Florida
Standards Assessments

Florida Standards Assessments

2014–2015

Volume 7
Special Studies

FLORIDA DEPARTMENT OF
EDUCATION
fldoe.org

Table of Contents

This volume consists of a series of independent special studies, conducted by AIR, FDOE, Alpine Testing Solutions, Inc., and Smarter Balanced Assessment Consortium. This volume will be updated each year to reflect studies relevant to the respective administration.

Calibrating Reading and Writing Scores

American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007-3835 | 202.403.5000 | TTY 877.334.3499 | www.air.org

Table of Contents

BACKGROUND

Students participating in the FSA English Language Arts (ELA) tests are administered two general types of items to produce an overall ELA score. Students are first administered a Writing prompt. This portion of the test contains a single prompt that is scored according to a rubric measuring three distinct dimensions. Students are then administered the Reading portion of the FSA. This portion of the test contains approximately 50 items, and the items on this portion measure the different dimensions of Reading found in the Florida standards.

When a student submits his or her final Reading and Writing tests, an overall ELA scaled score is produced that reflects a student's combined performance on the Writing and Reading portions of the test. Overall scores on both the ELA and Mathematics tests must have the highest degree of reliability, given their use for high stake decisions in the State of Florida. That is, because these overall scores have potential consequences for educators and students alike, it is important to establish overall scores meeting the highest degree of scientific reliability.

Because the Writing and Reading items are both used to form this overall ELA score, it is important to describe how this overall score is established. This document provides a brief description for how that score is created as well as the rationale used.

RELIABILITY AND VALIDITY

Generally speaking, scoring a test has two major objectives. The first objective is to find a way to produce the most reliable score. Reliability refers to the consistency in the test scores; a score is said to be *reliable* if the student would obtain approximately the same score when retaking the test. Reliability is a number that ranges from 0 to 1, where 0 denotes no reliability and 1 denotes perfect reliability. For any assessment, the higher the reliability, the better, and reliabilities for assessments that are used to inform promotion decisions should generally be at or above 0.80.

A second objective is validity. A score is said to be *valid* if it is measuring the targets it intends to measure. For example, suppose a test intends to measure basic Mathematics, but the test items require students to read large passages which contain mathematical issues that they must parse out in order to answer the Mathematics questions (e.g., a "story" problem). In this case, we may not know if a low score on a test like this is due to the low Reading ability or the low Mathematics ability of the student. It is important to note that a test cannot have high validity evidence if it is not reliable. Thus, we first require reliability evidence and only then can we consider validity.

RELIABILITY OF THE ELA SCORE

Because reliability and validity are key objectives, a study was performed to investigate various methods of how the ELA test could be scored in order to obtain the highest degree of reliability. The study was based on preliminary data created prior to the actual administration of the FSA in 2015. The purpose of the study was to compute an overall ELA score in various ways by combining the Reading and Writing scores and investigate which method produced the score with the highest level of reliability.

Two different types of scores were produced. The first score type used Writing items and Reading items collectively to form the overall ELA score. In this way, we do not calculate separate Reading and Writing scaled scores and then combine them; we only form a single, overall ELA score using all Reading and

Writing items simultaneously. Under the second score type, two separate scaled scores were produced and then combined. That is, we form a Reading scaled score and then a Writing scaled score, and then weighted the reading and writing scores with different methods to create a *composite* score. For example, one of the approaches was to weight the Reading score at 90% and the Writing score at 10%. Five different weight combinations were used.

The results showed that the highest reliability is obtained when a single overall ELA score is produced. An important trend we observed is that the reliability of the overall ELA score decreases when Writing has a larger weight in the overall score. This occurs because when a Writing score is computed, it alone cannot portray much information about a student's ability because it is a single Writing prompt. The Reading portion of the test, on the other hand, has higher reliability than the Writing prompt given that it contains upwards of 50 test items. When a score with lower reliability (i.e., the Writing score) has larger weight, it decreases the reliability of the entire overall score.

TECHNICAL ADVISORY COMMITTEE

The complete study was presented to the Florida Technical Advisory Committee (TAC), which consists of a group of nationally recognized experts in testing and measurement. Based on these results, as well as their experiences with similar scenarios in other states, they also recommended that the State of Florida create an overall ELA score by combining the Reading and Writing tests collectively instead of creating a composite score by weighting the two different tests differentially.

FDOE Final Decision

Using the study by AIR and the recommendation by TAC, FDOE decided that a single overall ELA score would be produced by combining the Reading and Writing tests. Raw scores are reported at the dimension level so that parents and educators can view how a student performed on different aspects of the test. However, these dimension scores always have lower reliability than the overall test score because they are based on a smaller number of items, and fewer items always provide less information regarding student performance.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Automated Scoring Engine

American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007-3835 | 202.403.5000 | TTY 877.334.3499 | www.air.org

Table of Contents

List of Tables

BACKGROUND

Students were administered a Writing prompt in grades 4-10 as one component of the FSA ELA test. Students in grades 4-7 were administered paper-based Writing prompts, while students in grades 8-10 were administered Writing prompts online. To maximize rater reliability, 100% of the papers were double-scored, and a full adjudication process was implemented to resolve any discrepant scores. For paper tests and for grade 10, the first and second rater (and resolution rater as needed) were always human raters. For grades 8 and 9, the second rater was an online scoring engine referred to as Autoscore.

This document details how Autoscore, AIR's automated scoring engine, was implemented to provide scores for online prompts and additionally demonstrates how it meets industry-standard rates for rater reliability. The document begins by revisiting the methods used in the independent field test (IFT), as the data from this IFT serve as the basis for training the engine on newly developed prompts, and subsequently describes how Autoscore uses maximally valid papers to train the engine and implement the operational score analysis.

REVIEW OF INDEPENDENT FIELD TEST METHODS

The Florida Department of Education (FDOE) administered an online Writing independent field test (IFT) to a sample of Florida students from December 2014 to early February 2015. The objectives for the IFT were:

- to obtain item statistics on the newly developed Writing prompts for grades 4 to 10; and
- to review the item statistics and choose Writing prompts that would be used as operational items beginning in the Spring 2016 school year.

A simulation study was previously provided to FDOE describing various statistical issues related to sample sizes and pool size. Upon the consideration of this simulation study, FDOE communicated to AIR that the following principles were to be used for the Writing IFT:

- Each student would respond to two Writing prompts.
- A stratified random sample would be drawn to represent the state.
- A total of 22,500 students within each grade would participate in the Writing IFT.

A scientific sample was used to identify and select the 22,500 students for the IFT. A stratified random sample of intact schools was used, thus requiring all students within a school to participate in the IFT. The generalized selection methods were as follows:

Let $k_{(j)g}$ denote the number of students in grade g in the jth school $j = \{1, 2, \dots N_g\}$ and $K_g = \sum_{J=1}^{N_g} k_{(j)g}$. N_g is the total number of eligible schools in grade g. FDOE required a nominal sample size of 22,500 students. Hence, assuming a typical sample size of

$$\bar{k}_g = \frac{K_g}{N_g},$$

we obtained the total number of schools required for sampling to be

$$M_g = \frac{22{,}500}{\bar{k}_g}.$$

Rather than making an arbitrary assumption regarding the value of $\bar{k}_g$, AIR derived the value for each grade from the data provided in the State Student Results (SSR) files. Intact schools were sampled, and then all relevant grades within the school were sampled for the IFT.

Stratified Sampling

In order to use a proportionate stratification method, we first identified the proportion of schools across the state within stratum l with the number of students $l_{n,g}$ as

$$P_{l,g} = \frac{l_{n,g}}{N_g},$$

and then within each stratum, $m_{l,g} = P_{l,g} M_g$ schools were sampled. The sampling method used both explicit and implicit strata.

For hierarchic serpentine sorting, we sorted the first variable in ascending order within a stratum. Then after sorting the first variable, within the first level of the first variable, we sorted the second variable in ascending order, and within the second level of the first variable, we sorted the second variable in descending order. We applied this process to all levels and all variables, thus making the sorting equivalent to alternative ascending and descending by each variable. The implicit strata are further described in the next section.

Strata

In order to yield a representative sample of students from the testing population, it was first necessary to identify the sampling strata in order to capture the important characteristics of the state population. For this reason, both explicit and implicit strata were used.

The state was first divided into various geographic regions, which served as the explicit stratum in this sampling design, with the intent to capture the differences in student populations across the state. For consistency with prior sampling methods used in Florida, only the North, Central, and South regions were used as stratification variables. The number of regions was collapsed from five to three, so that regions 01 and 02 formed the northern region, regions 03 and 04 formed the central region, and region 05 remained as the southern region. Any student with a region variable of 06 or blank was removed from sampling. Within this explicit stratum, schools were sorted in a serpentine order (alternating ascending and descending order) by the implicit strata, binned as quintiles. $m_{l,g}$ schools were sampled systematically by the implicit strata listed below.

1. Percent Proficient within a given school on the prior year's Reading test
    - This variable was intended to capture the ability of students across the population.
2. School Size

- This variable was intended to ensure schools of various sizes were represented in the sample.

3. Curriculum Group
   - Standard, LEP, ESE
4. Gender
5. Percent Ethnicity
   - The demographic variables White, African American, and Hispanic were used.

Probability Weights

If $m_{l,g}$ was the number of schools selected from stratum *l*, then the probability of the *j*th school within stratum *l* being selected was

$$p_{lj} = \frac{m_{j,l}}{m_{l,g}},$$

where $m_{j,l}$ is the *j*th school in the *l*th stratum. All classes were selected within the selected school, so the selection probability of class *k* in school *j* from stratum *l* iwas

$$p_{ljk} = 1.$$

Finally, all students were selected from each selected class so that the selection probability of student *i* is

$$p_{iljk} = 1$$

The overall selection probability for a student *i* being selected for the sample was calculated as

$$p_i = p_{ij}p_{ijk}p_{ijlk} = p_{lj}.$$

The overall weight was calculated as $w_i = \frac{1}{p_i}$, with the weight then normalized within each stratum to the total number of sampled students. More specifically, suppose that there were a total of n_l sampled students in stratum *l,* and the weighted sample size for stratum *l* was calculated as $W_l = \sum_{i \in l} w_i$. Then, the normalized weight for sampled student *i* in stratum *l* would be

$$\widetilde{w}_i = \frac{n_l}{W_l} w_i.$$

Field Test Algorithm

The algorithm employed by AIR's field test engine ensured that items from the available pool were drawn according to a simple random sample.

Assuming the pool has *R* total items, we computed the expected number of responses in the sample, $\mathrm{N_p}$, to each of the Writing prompts as follows:

$$\mathrm{E(N_p)} = 22{,}500\left(\frac{2}{\mathrm{R}}\right)$$

This represented the nominal response rate to each item in the total population. For this design, the schools served as the clusters. However, Writing prompts were randomly assigned to all individuals; hence, the cluster size was the number of students within a school that were assigned the same Writing prompt. Again, using the binomial, we computed the expected number of students that were administered the same Writing prompt within each school as follows:

$$\mathrm{E(N_c)} = \left(\frac{2}{\mathrm{R}}\right)\bar{k}_g$$

Based on the expected cluster size and assuming the intra-class correlation, ρ, was fixed at 0.15, a value derived from a variance decomposition of the 2013 grade 5 ELA data, and the design effect was then computed as

$$\mathrm{DEFF} = 1 + .15(\mathrm{E(N_c)} - 1).$$

These estimates were then used to find the expected effective sample size (ESS), which represented the effective number of students used to calibrate each Writing prompt:

$$\mathrm{E(ESS)} = \frac{\mathrm{E(N_p)}}{\mathrm{DEFF}}$$

Table 1 provides the estimated design effects and effective sample sizes based on the projected values of R and Mg.

Table 1: Expected Sample Sizes With Fixed Population Size

Grade	Pool Size (R)	Nominal Expected Sample Size (per item)	Expected Cluster Size	Design Effect	Expected Effective Sample Size (per item)	Overall N
4	14	3214	13	2.78	1157	22,500
5	15	3000	12	2.65	1132	22,500
6	16	2813	13	2.82	998	22,500
7	16	2813	14	2.96	950	22,500
8	17	2647	12	2.70	979	22,500
9	14	3214	11	2.46	1308	22,500
10	15	3000	11	2.50	1200	22,500

The random assignment of items to students generated the randomized sparse matrix design represented in Table 2.

Table 2: Sample Design 1 Linkage

Student	Item 1	Item 2	Item 3	…	Item R
S1	x	x		…	
S2	x			…	x
S3			x	…	x
S4		x	x	…	
. . .	x			…	x
N	x		x	…	

The benefit of this design was that every item in the pool served as a common item linking various students, thus allowing for a joint calibration of all items simultaneously. For example, the table shows that item 1 was shared by students S1 and S2. In addition, student S1 was linked to student S4 via item 2. This linkage allowed for all *R* items in the pool to serve as the set of common items.

SPRING 2015 OPERATIONAL PROMPT

During form construction for Spring 2015, a single Writing prompt per grade was chosen for operational use. This prompt did not come from the Writing IFT; instead the prompt was chosen from the Utah SAGE item bank. The same scientific methods for sample selection were used to identify papers in order to train the scoring engine, as described in Volume 1, Section 6.

There are essentially two phases that were implemented during the spring of 2015 in order to train and implement the Autoscore process. The first critical task was to establish a maximally valid set of papers that could be used to train the engine. This set of papers was taken from the scientific sample chosen as part of the early processing sample, and each paper in this subset was 100% double scored by two human raters and assigned a final, resolution score. All discrepant records in this set were sent for final resolution and then assigned a final resolution score only when the two raters disagreed on the dimensions scores. This set of papers was deemed maximally valid, as the final, resolved scores were provided entirely via human readers and underwent a complete adjudication process.

All other papers in grades 8 and 9 were scored only by a single human rater, and the second score was obtained from the Autoscore engine. While Autoscore provided the second score in lieu of a second human rater, the process for assigning a final score in grades 8 and 9 was consistent with all other grades. When the scores of the human rater and the Autoscoring engine were adjacent, the higher score of the two scores was assigned as the final score. When a discrepant record (i.e., not adjacent) existed, then the paper was sent for a resolution read by a human scorer, and the adjudication process and assigning of the final score was completed at Data Recognition Corporation (DRC).

Table 3 presents descriptive statistics based on the final scores sent to AIR after resolution. Students received scores ranging from 0-2 on the Conventions dimension and scores ranging from 1-4 on both the Evidence & Elaboration and Purpose, Focus, & Organization dimensions.

Table 3: Descriptive Statistics on Each Operational Writing Item

	Conventions		Evidence & Elaboration		Purpose, Focus, & Organization	
Grade	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
4	1.72	0.55	1.94	0.80	2.02	0.80
5	1.84	0.45	2.11	0.77	2.22	0.78
6	1.74	0.59	2.06	0.85	2.16	0.88
7	1.73	0.54	2.17	0.82	2.28	0.84
8	1.74	0.54	2.34	0.80	2.43	0.82
9	1.72	0.55	2.42	0.78	2.57	0.76
10	1.72	0.54	2.29	0.86	2.43	0.87

METHODS FOR ONLINE ESSAY SCORING

AIR's essay scoring engine, Autoscore, uses a statistical process to evaluate Writing prompts. Autoscore evaluates papers against the same rubric used by human raters, but a statistical process is used to analyze each paper and assign scores for each of the three dimensions. The engine uses the same process for scoring essays every time a new prompt is submitted, regardless of whether the data is obtained from an operational assessment or an IFT.

Statistical rubrics are, effectively, proxy measures. Although they can directly measure some aspects of Writing conventions (e.g., use of passive voice, misspellings, run-on sentences), they do not directly measure argument structure or content relevance. Hence, although statistical rubrics often prove useful for scoring essays and even for providing some diagnostic feedback in Writing, they do not develop a sufficiently specific model of the correct semantic structure to score many propositional items. Furthermore, they cannot provide the explanatory or diagnostic information available from an explicit rubric. For example, the frequency of incorrect spellings may predict whether a response to a factual item is correct— higher-performing students may also have better spelling skills. Spelling may prove useful in predicting the human score, but it is not the actual reason that the human scorer deducts points. Indeed, statistical rubrics are not about explanation or reason but rather about a prediction of how a human would score the response.

AIR's essay-scoring engine uses a statistical rubric with great success, as measured by the rater agreements observed relative to the human-to-human rater agreements. This technology is similar to all essay-scoring systems in the field. Although some systems replace the statistical process with a "neural network" algorithm, that algorithm functions like the statistical model. Not all descriptions of essay-scoring algorithms are as transparent as AIR's, but whenever a training set is used for the machine to "learn a rubric," the same technology is being used.

Training Set

The engine is designed to employ a "training set," a set of essays scored with maximally valid scores that are used to form the basis of the prediction model. The quality of the human-assigned scores is critical to the identification of a valid model and final performance of the scoring engine. Moreover, an ideal training sample over-represents higher-and lower-scoring papers and is selected according to a scientific sampling design with known probabilities of selection.

The training process of the scoring engine has two phases. The first phase requires over-sampled, high- and low-scoring papers, leaving an equally weighted representative sample for the second phase. The first phase is used to identify concepts that are proportionately represented in higher-scoring papers. Here, concepts are defined as words and their synonyms, as well as clusters of words used meaningfully in proximity.

The second phase takes a series of measures on each essay in the remaining training set. These measures include latent semantic analysis (LSA) measures based on the concepts identified in the first phase; other semantic measures indicate the coherence of concepts within and across paragraphs and a range of word-use and syntactic measures. The LSA is similar to a data reduction method identifying common concepts within the narrative and reducing the data to a configurable number of LSA dimensions.

For each trait in the rubric, the system estimates an appropriate statistical model where these LSA and other syntactic characteristics described above serve as the independent variables, and the final, resolved score serves as the dependent variable in an ordered probit regression. This model, along with its final parameter estimates, is used to generate a predicted or "proxy" score. The probability of scoring in the *p*th category is compared to a random draw form the uniform distribution, and a final score point of 1 through 4 is determined from this comparison.

Cross Validation Set Analysis

In addition to the training set, an independent random sample of responses is drawn for the cross-validation of the identified scoring rubric. As with the training set, student responses in the cross-validation study are hand scored, and the LSA and other syntactic characteristics of the papers are computed. Subsequently, a second machine score is generated by applying the model coefficients obtained from the ordered probit in the training set. This forms a predicted score for the papers in the cross-validation set for each dimension in the rubric, which can then be used to evaluate the agreement rates between the human and Autoscore engine.

When implementing the scoring engine, we expect that the computer-to-human agreement rates to be at least as high as the human-to-human agreement rates obtained from the double-scored process. If the engine yields scores with rater agreement rates that are at least as high as the human rater agreement rates, then the scoring engine can be deployed for operational scoring. If the computer-to-human agreement rates are not at least as high as the human-to-human rates, then adjustments to the scoring engine statistical model are necessary in order to find a scoring model that yields rater agreement rates that match the human-to-human rates.

SUMMARY OF FINAL SOLUTIONS

Table 4 and Table 5 provide the human-to-human rater agreement rates for the maximally valid set of papers used to train the engine in grades 8 and 9, obtained from DRC's scoring on each of the three dimensions. These values serve as a benchmark for later evaluating Autoscore, as the computer-to-human rater agreement rates should be similar to the human-to-human rates observed in these tables.

Table 4. Human-to-Human Rater Agreement Rates for Grade 8

Grade 8	% Exact	% Adjacent	% Not Adjacent
Purpose	67	32	1
Conventions	69	30	1
Elaboration	76	24	1

Table 5. Human-to-Human Rater Agreement Rates for Grade 9

Grade 9	% Exact	% Adjacent	% Not Adjacent
Purpose	65	34	1
Conventions	72	28	0
Elaboration	73	26	1

A subset of approximately 1,500 of the maximally valid papers were used for training the scoring engine, and approximately 600 papers were used in the cross-validation process.

In comparison, Table 6 and Table 7 provide the computer-to-human agreement rates from the scoring engine model that were deployed for operational scoring in Spring 2015. In all instances, the scoring engine yielded scores that were comparable to the human-to-human rates provided in the maximally informative training set.

Table 6: Computer-to-Human Rater Agreement Rates for Grade 8

Grade 8	% Exact	% Adjacent	% Not Adjacent
Purpose	68.69%	30.66%	0.66%
Conventions	80.82%	18.69%	0.49%
Elaboration	71.97%	27.87%	0.16%

Table 7: Computer-to-Human Rater Agreement Rates for Grade 9

Grade 9	% Exact	% Adjacent	% Not Adjacent
Purpose	75.57%	24.10%	0.33%
Conventions	76.72%	22.79%	0.49%
Elaboration	72.13%	27.54%	0.33%

CONCLUSION

Overall, Autoscore produces scores that are at least as reliable as the human-to-human rater agreement, which we take as a benchmark in order to deploy the scoring engine. On this basis, we can assume that the scoring engine yielded results that were comparable to what another human would have produced, had all papers been double-scored by two humans. As the engine is trained each year when a new Writing prompt is introduced, the model coefficients and results in this report are not generalizable over time. Each year, AIR will update the model coefficients and describe any nuances in the Autoscore that may be different from those used in the Spring 2015 administration.

AMERICAN INSTITUTES FOR RESEARCH®

Classification Accuracy

American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007-3835 | 202.403.5000 | TTY 877.334.3499 | www.air.org

Table of Contents

List of Tables

Background

When students complete the Florida Standards Assessments (FSA), they are placed into one of five achievement levels given their observed scaled score. Volume 3 of the FSA technical reports provides details on the FSA standard-setting process and the recommended cut scores for student classification into the different achievement levels.

During test construction, techniques are implemented to minimize misclassification of students, which can occur on any assessment. In particular, standard error of measurement (SEM) curves can be constructed to ensure that smaller SEMs are expected near important cut scores of the test.

Misclassification probabilities are computed for all achievement level standards, i.e. for the cuts between levels 1 and 2, levels 2 and 3, levels 3 and 4, and levels 4 and 5. The achievement level cut between level 2 and level 3 is of primary interest because students are classified as Satisfactory or Below Satisfactory using this cut. Students with observed scores far from the level 3 cut are expected to be classified more accurately as Satisfactory or Below Satisfactory than students with scores near this cut. This report estimates classification reliabilities based on observed abilities.

Classification Accuracy

Observed score approach to computing misclassification probabilities and is designed to explore the following research question: What is the classification accuracy rate index for each individual performance cut within the test?

Method

Data

We used students from the Spring 2015 FSA SSR files with the status of reported scores (score status flag of 1). Table 1 provides the sample size, mean, and standard deviation of the observed theta. The theta scores are MLE based estimations obtained from AIR's scoring engine.

Table 1: Descriptive Statistics from Data

ELA Grade	Sample Size	Average Theta	Standard Deviation of Theta	Mathematics Grade/EOC Subject	Sample Size	Average Theta	Standard Deviation of Theta
3	215317	0.00	1.07	3	215473	0.02	1.06
4	197681	0.03	1.00	4	199351	0.02	1.09
5	196812	0.00	1.01	5	199010	0.00	1.06
6	192614	-0.04	1.06	6	191091	-0.06	1.09
7	192024	0.00	1.06	7	179194	-0.01	1.11
8	198412	-0.01	1.05	8	123928	-0.01	1.10
9	201252	0.04	1.05	Algebra 1	203235	-0.10	1.21
10	191080	0.01	1.07	Algebra 2	158254	-0.20	1.32
blank	blank	blank	blank	Geometry	195113	-0.05	1.11

Implementation

The observed score approach (Rudner, 2001) implemented to assess classification accuracy is based on the probability that the true score, θ, for student i is within performance level $j = 1,2,\cdots,J$. This probability can be estimated from evaluating the following integral

$$p_{ij} = \Pr(\lambda_l \le \theta_i < \lambda_u | \hat{\theta}_i, \hat{\sigma}_i^2) = \int_{\lambda_l}^{\lambda_u} f\left(\theta_i \middle| \hat{\theta}_i, \hat{\sigma}_i^2\right) d\theta_{i,}$$

where λ_u and λ_l denote the score corresponding to the upper and lower limits of the performance level, respectively, $\hat{\theta}_i$ is the ability estimate of the ith student with standard error of measurement of $\hat{\sigma}_i$ and using the asymptotic property of normality of the maximum likelihood estimate, $\hat{\theta}_i$, we take $f(\cdot)$ as asymmetrically normal, so the above probability can be estimated by

$$p_{ij} = \Phi\left(\frac{\lambda_u - \hat{\theta}_i}{\hat{\sigma}_i}\right) - \Phi\left(\frac{\lambda_l - \hat{\theta}_i}{\hat{\sigma}_i}\right),$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (CDF).

The expected number of students at level j based on students from observed level k can be expressed as

$$E_{kj} = \sum_{pl_i \in k} p_{ij},$$

where pl_i is the ith student's performance level, the values of E_{kj} are the elements used to populate the matrix $\boldsymbol{E}$, a 5 × 5 matrix of conditionally expected numbers of students to score within each performance level bin based on their true scores.

The classification accuracy index for the individual cuts is estimated by forming square partitioned blocks of the matrix $\boldsymbol{E}$ and taking the summation over all elements within the block as follows:

$$\text{CAIC} = \left(\sum_{k=1}^{p}\sum_{j=1}^{p} E_{kj} + \sum_{k=p+1}^{5}\sum_{j=p+1}^{5} E_{kj}\right)\Big/ N,$$

where $N = \sum_{k=1}^{5} N_k$, N_k is the observed number of students scoring in performance level k, p is the elements of one of the cuts of interest.

Results

Table 2 and Table 3 provide the classification accuracy index for the individual cuts (CAIC) for the English Language Arts (ELA), Mathematics and EOC tests, respectively, based on the observed score approach. The overall cut accuracy rates denote that the degree to which we can reliably differentiate students between adjacent performance levels is typically above or close to 0.9 for ELA, Mathematics, and EOC assessments.

Table 2: Classification Accuracy Index (ELA)

Grade	Cut Accuracy			
	Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4	Cut 4 and Cut 5
3	0.926	0.912	0.931	0.965
4	0.933	0.907	0.919	0.958
5	0.932	0.912	0.923	0.959
6	0.939	0.924	0.932	0.964
7	0.932	0.921	0.927	0.953
8	0.943	0.924	0.926	0.951
9	0.942	0.922	0.926	0.954
10	0.939	0.912	0.922	0.959

Table 3: Classification Accuracy Index (Mathematics and EOC)

Grade/Subject	Cut Accuracy			
	Cut 1 and Cut 2	Cut 2 and Cut 3	Cut 3 and Cut 4	Cut 4 and Cut 5
3	0.948	0.932	0.927	0.953
4	0.942	0.934	0.941	0.958
5	0.946	0.934	0.939	0.958
6	0.933	0.927	0.942	0.969
7	0.932	0.929	0.949	0.971
8	0.906	0.901	0.939	0.972
Algebra 1	0.885	0.884	0.936	0.967
Algebra 2	0.883	0.915	0.959	0.972
Geometry	0.924	0.924	0.957	0.974

Summary

These results are based on the Spring 2015 test administration, when performance classifications were not yet assigned. FSA performance cuts were established on January 6, 2016, and will be applied to students beginning in Spring 2016.

These results demonstrate that classification reliabilities are generally high. The classification cut accuracy rates in Mathematics and EOC range from 0.883 in Algebra 2 to 0.974 in Geometry. Similarly, the classification cut accuracy rates in ELA range from 0.907 in grade 4 to 0.965 in grade 3.

References

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation, 7*(14).

Vertical Linking

American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007-3835 | 202.403.5000 | TTY 877.334.3499 | www.air.org

Table of Contents

List of Appendices

List of Tables

List of Figures

BACKGROUND

The first operational administration of the Florida Standards Assessments (FSA) occurred during spring of 2015. Administered test forms contained operational items and embedded field test (EFT) items in pre-determined slots across each form. Operational items were items used to calculate student scores. The EFT items were non-scored items and were used either to populate the FSA test bank for future operational use or to establish linkages to other test forms.

This document is concerned with the latter of these non-scoring items, and it describes the development of a new vertical scale in Mathematics and English Language Arts (ELA) based on a common-item, non-equivalent groups design (Kolen & Brennan, 2004). In Mathematics, grades 3 to 8 were linked on a vertical scale, and each of the End-of-Course tests remained on their separate scales. In ELA, all tests in grades 3 to 10 were placed on the same vertical scale.

Prior to the Spring 2015 administration, AIR proposed developing a series of variants (A-G) of the vertical scale during the operational season, from which the Department could choose. Each of these variants, described in Initial Variants on page 3 below, were vetted by the Technical Advisory Committee (TAC). However, during the operational season, additional variants were developed between AIR and FDOE, and ultimately, a new variant H was chosen to develop the final vertical scale. The creation of the final vertical scale is described in Section 6.4 of Volume 1 of the 2015 technical reports.

This standalone chapter contains details about the entire process used to create the new vertical scale. The remainder of this document is organized to describe vertical linking terminology, common item design, the methods for analysis and linking, and the specific steps used to create the FSA vertical scale.

GENERAL DESCRIPTION OF VERTICAL SCALE DEVELOPMENT

In this section we describe the terminology used throughout the vertical linking report.

Linking Direction

The term *linking direction* is used to mean the grade from which the common items used in the vertical scale are measuring the on-grade content. This section describes three common approaches and details the benefits and risks of the various approaches.

Backwards linking occurs when items intending to measure on-grade content from grade $g+1$ are placed onto the test forms in grade g. An example of backwards linking is when items measuring grade 4 content are placed onto test forms as linking items in grade 3.

Forwards linking occurs when items intending to measure on-grade content from grade g are placed onto the test forms in grade $g+1$. An example of forwards linking is when items measuring grade 3 content are placed onto test forms as linking items in grade 4.

Mixed linking occurs when both forwards and backwards linking methods are combined to create a vertical scale. Items measuring content from grades $g – 1$ as well as $g + 1$ are on test forms in grade g. The only exception is for tests at the extreme grades. For example, there is no grade 2 test; hence grade 3 forms contain only items from grade 4.

AIR presented all three linking methods to the TAC in November 2014. After consideration, and based on experience, the TAC recommended that FDOE and AIR use a mixed linking design. Other methods were implemented by AIR to evaluate the complete set of options.

Stocking-Lord Method

In order to compare item characteristics from different calibrations, the calibrations must be placed on the same scale. Stocking-Lord (1983) is a method commonly used alongside the 3-parameter logistic model and Generalized Partial Credit Model and establishes the linking constants, *A* and *B*, that minimize the squared distance between two test characteristic curves. *A* is often referred to as the slope and *B* is often referred to as the intercept. The approach evaluates the following integral, where the indices *i* denote a common item and *a* and *b* denote separate forms:

$$SL = \int \sum_{i=1}^{I} \left(ICC_{ai}(\theta) - ICC_{bi}(\theta)\right)^2 f(\theta; \mu, \sigma^2) d\theta,$$

where for dichotomous items,

$$ICC_i(\theta) = c_i + (1 - c_i) \frac{exp\left(Da_i(\theta - b_i)\right)}{1 + exp\left(Da_i(\theta - b_i)\right)},$$

and for polytomous items,

$$ICC_i(\theta) = \sum_{l=1}^{K_i} \frac{lExp\left(Da_i \sum_{k=1}^{l}(\theta - b_{i,k})\right)}{1 + \sum_{m=1}^{K_i} Exp\left(Da_i \sum_{k=1}^{m}(\theta - b_{i,k})\right)},$$

where K_i is the maximum item score point, $ICC_{ai}(\theta)$ is the probability of a correct response on item i on form a, given an ability of θ and $ICC_{bi}(\theta)$ is the probability of a correct response on item i on form *b* given an ability of θ, and the marginal density is $f(\theta) \sim \mathrm{N}(0,1)$.

The linking constants from the vertical scale represent the first and second moments of the ability distributions in the grades in which they are applied. For this reason, we expect the linking constant *B* to increase from grade 3 to the highest grade (grade 8 in Mathematics and grade 10 in ELA) representing an increase in the means over grades. The linking constant *A* often tends to become larger in higher grades, indicating larger variances between students over grades.

COMMON ITEM LINKING DESIGN

Operational test forms for the Spring 2015 FSA administration were developed during form construction meetings in the summer of 2014. During that time, common items used to develop the vertical scale were selected and placed onto test forms.

During form construction, items from both the upper grade as well as the lower grade were placed onto the on-grade forms, thus forming the mixed linking design. This common item linking design was established in advance knowing that this design would allow for all three options (backwards, forwards, and mixed linking) to be considered for the linking design.

The primary goal when developing the common item linking design was to administer a linking set that represented the content of the tests from which the items were derived. For example, the grade 4 items placed onto the grade 3 form were intended to represent the grade 4 test blueprint. This design supported the inference that the scaled score from the vertical scale represented both the on-grade performance as well as the location of a student's performance on the upper grade test.

Appendix A provides a set of tables describing the common item linking design. The tables show the representativeness of the linking set vis-à-vis the test blueprint from which the items were derived, the total number of test forms, the number of vertical linking forms, and sample sizes by grade and vertical linking form.

INITIAL VARIANTS

As proposed, multiple variants of the vertical scale were implemented to explore the effects of various methods and provide options for FDOE. Versions A–G were initially vetted by TAC and are outlined below. While these were not ultimately chosen as the final linking versions, they were all calculated, as initially proposed, and were useful in the development of the final variant, Version H.

1) **Version A:** Backwards linking: This method used items that were on-grade in grade *g* and vertical linking in grade *g-1*.

    a. For example, items measuring on-grade content in grade 4 were placed onto the grade 3 test forms. These were the backward linked items. The two sets of parameters from the calibrations were used to find the linking constants between the tests.

2) **Version B:** Forwards linking: This approach used items that were on-grade in grade *g* and vertical linking in grade *g+1*.

    a. Items measuring on-grade content in grade 3 were placed onto the grade 4 test forms. These were the forward linked items. The two sets of parameters from the calibrations were used to find the linking constants between the tests.

3) **Version C:** Mixed linking: This method used items that were on-grade in grade *g* and vertical linking in grade *g-1* and additionally used items that were on-grade in grade *g-1* and vertical linking in grade *g*.

    a. As an example, grade 4 contained on-grade items placed onto the grade 3 forms and also included items from grade 3 placed onto the grade 4 form. Both sets of these items were used to find the linking constants between grades 3 and 4.

4) **Version D:** Mixed Linking: This type of linking used the calibrated item parameters from Version C but dropped items with poor model fit. The Q1 statistic was used as the measure of model fit.

5) **Version E:** Mixed Linking: This version used the calibrated item parameters from Version C but removed items if p-values were reversed over grades.

6) **Version F:** Mixed Linking: This version used the calibrated item parameters from Version C, but removed items if poor model fit or if p-values were reversed.

7) **Version G:** Mixed Linking: This version used the calibrated item parameters from Version C but removed items with poor classical and/or D^2statistics (see flagging criteria below).

FLAGGING CRITERIA

As indicated in the outline of the versions above, items were dropped in variants D-F for poor item fit, based on the Q1 fit statistic or reversal of p-values across grades. In addition, items were flagged for inspection if there were poor classical item statistics or large position shifts between grades. If any of the flagged items required removal, they were dropped according to Variant G outlined above.

Table 1 outlines the flagging criteria that were used. The flagging rules outlined in Table 1 did not necessarily require that items be removed, but they were used as guidelines for further review before a final decision was made on whether to keep or remove a particular item.

Classical and IRT item statistics can be found in Appendix B. For each of the vertical linking items administered, their role (e.g., on-grade), IRT item parameters, Q1 fit flag, p-value, and point biserial are listed for each grade. In addition a table lists the number of items flagged by each of the criteria.

Table 1: Flagging Criteria for Vertical Linking

Rule	Flagging Criteria	Rationale
p-value	For multiple choice items, flag if p < 0.25 or p > 0.95	Items are too difficult and p-value is less than expected from random chance or item is too easy for population
Relative mean	For polytomous items, flag if relative mean is < 0.15 or > 0.95	Item difficulty is too difficult or too easy
Point Biserial/polyserial	Flag if < 0.15	Non-discriminating item
Distractor p-value	Flag if p-value for distractor is larger than p-value for key	Potentially problematic item
Distractor biserial	Flag if biserial for any distractor is larger than biserial for key	Distractor is more discriminating than the keyed response
Item Position shift	Flag if item shifts more than 5 positions	Item position may affect item performance
D^2 and ICCs	Flag if D^2 greater than 3 standard deviations	Difference between grades is too large
Convergence Issues	Flag IRT statistics if IRTPRO does not converge	The number of iterations and convergence should be noted in a table.

Calculating the Q1 Fit Statistic

To evaluate model fit, the Q1 statistic was calculated for all operational items. Q1 is a fit statistic that compares observed and expected item performance. MAP estimates from IRTPRO were used for student ability estimates in the calculations. Q1 is calculated as

$$\mathrm{Q}_{1\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \frac{\mathrm{N}_{\mathrm{ij}}(\mathrm{O}_{\mathrm{ij}} - \mathrm{E}_{\mathrm{ij}})^2}{\mathrm{E}_{\mathrm{ij}}\left(1 - \mathrm{E}_{\mathrm{ij}}\right)}$$

where N_{ij} is the number of examinees in cell *j* for item *i*, $\mathrm{O}_{\mathrm{ij}}$and E_{ij} are the observed and predicted proportions of examinees in cell *j* for item *i*. The expected or predicted proportion is calculated as

$$\mathrm{E}_{\mathrm{ij}} = \frac{1}{\mathrm{N}_{\mathrm{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N}_{\mathrm{ij}}} \mathrm{P}_{\mathrm{i}}\left(\hat{\theta}_{\mathrm{a}}\right)$$

where $\mathrm{P}_{\mathrm{i}}\left(\hat{\theta}_{\mathrm{a}}\right)$ is the item characteristic function for item i and examinee a. The summation is taken over examinees in cell *j*. The generalization of Q1, or Generalized Q1, for items with multiple response categories is

$$gen\ \mathrm{Q}_{1\mathrm{i}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{k}=1}^{\mathrm{m}_{\mathrm{i}}} \frac{\mathrm{N}_{\mathrm{ij}}(\mathrm{O}_{\mathrm{ikj}} - \mathrm{E}_{\mathrm{ikj}})^2}{\mathrm{E}_{\mathrm{ikj}}}$$

with

$$\mathrm{E}_{\mathrm{ikj}} = \frac{1}{\mathrm{N}_{\mathrm{ij}}} \sum_{\mathrm{aej}}^{\mathrm{N}_{\mathrm{ij}}} \mathrm{P}_{\mathrm{ik}}\left(\hat{\theta}_{\mathrm{a}}\right).$$

Both the Q1 and Generalized Q1 results are transformed into the statistic ZQ1, and are compared to a criterion, $\mathrm{ZQ}_{\mathrm{crit}}$, to determine acceptable fit. These are defined as

$$\mathrm{ZQ}_1 = \frac{\mathrm{Q}_1 - \mathrm{df}}{\sqrt{2\mathrm{df}}}$$

and

$$\mathrm{ZQ}_{\mathrm{crit}} = \frac{\mathrm{N}}{1500} * 4,$$

where Q is either Q1 or Generalized Q1, and df is the degrees of freedom for the statistic. The degrees of freedom is calculated as 10 – number of parameters estimated. For example, multiple choice items have df = 7 Poor fit is indicated where ZQ1 is greater than $\mathrm{ZQ}_{\mathrm{crit}}$.

Calculating the D^2 Statistic

After performing the Stocking-Lord, the equated parameters were compared by rescaling items to be on the same scale. D^2 the sum of the squared differences between ICCs, was calculated.

The D^2, or the MSD, is computed by integrating out θ as follows:

$$D^2 = \int \left(ICC_{ai}(\theta) - ICC_{bi}(\theta)\right)^2 f(\theta; \mu, \sigma^2) d\theta.$$

The integral does not have a closed form solution, and so its approximation is based on the weighted summation over j={1, 2, …, 30} quadrature points, all taken from equally spaced points interior to the normal density, *w*, between -4 and 4 of the marginal distribution.

$$\mathrm{D}^2 = \sum_{j=1}^{30} w_j \left(ICC_{ai}(\theta_j) - ICC_{bi}(\theta_j)\right)^2$$

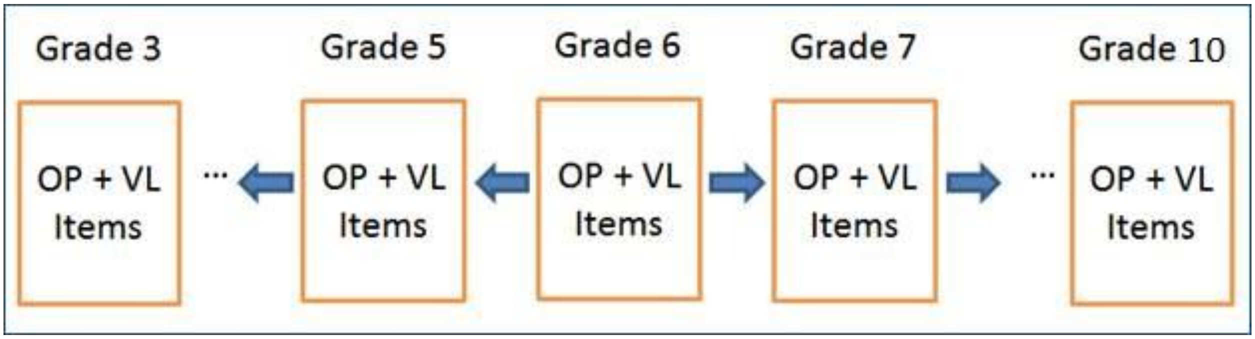
D^2 was calculated and ICCs plotted. Items with D^2 values more than 3 standard deviations were flagged for review.

Implementation Steps

To complete the vertical linking, four IRT calibrations were performed per grade: (1) the operational items only, (2) operational items with backward vertical linking items, (3) operational items with forward vertical linking items, and (4) operational items with on-grade vertical linking items. For example, in grade 5, the on-grade calibration included the vertical linking items that came from grade 5, the forward calibration included the vertical linking items that came from grade 4, and the backward calibration included the vertical linking items that came from grade 6. Note that the grades at the end of the vertical scale required one less calibration. In each calibration, the operational items and the vertical linking items were freely calibrated. After each calibration, the number of iterations and any convergence issues were discussed.

After IRT calibrations were complete, AIR implemented each initial vertical scale version A-G previously described. Items that met the flagging criteria were discussed and removed according to decisions by FDOE. Adjacent grades were chain linked using the Stocking-Lord method to obtain linking constants.

For initial variants A-G, grade 6 was used as the base (or anchor) grade in the vertical linking, as depicted in the figure below.



Implementation Steps to Find Linking Constants – Initial Variants

This section lists the step-by-step procedures that were completed for constructing the linkages. The steps below were generally applied to both subjects with one notable exception. In Mathematics, linking occurred for grades 3 to 8 only. In ELA, linking occurred for grades 3 to 10. Initially, grade 6 was used as the base grade for the development of the variants.

1) Data files were prepared for grade *g*.
   a. Data included only the operational items and the vertical linking items.
2) A separate calibration was conducted of the operational items administered to each grade using IRTPRO.
3) A second calibration was performed, including the operational items and the vertical linking items.
4) For each version, chain linking was implemented via the Stocking-Lord procedure, removing flagged items if necessary, according to the following plan:
   a. Link grade 6 to grade 5 and identify linking constants A_{56} and B_{56}
   b. Link grade 5 to grade 4 to find linking constants A_{45} and B_{45}
   c. Link grade 4 to grade 3 to find linking constants A_{34} and B_{34}
   d. Link grade 6 to grade 7 and identify linking constants A_{67} and B_{67}
   e. Link grade 7 to grade 8 to find linking constants A_{78} and B_{78}
   f. Link grade 8 to grade 9 to find linking constants A_{89} and B_{89}
   g. Link grade 9 to grade 10 to find linking constants $A_{9,10}$ and $B_{9,10}$
5) Linking constants were updated via the following transformations:
   a. $A'_{56} = A_{56}$ and $B'_{56} = B_{56}$
   b. $A'_{45} = A'_{56}\, A_{45}$ and $B'_{45} = B'_{56} + A'_{56}B_{45}$
   c. $A'_{34} = A'_{45}\, A_{34}$ and $B'_{34} = B'_{45} + A'_{45}B_{34}$
   d. $A'_{67} = A_{67}$ and $B'_{67} = B_{67}$
   e. $A'_{78} = A'_{67}\, A_{78}$ and $B'_{78} = B'_{67} + A'_{67}B_{78}$
   f. $A'_{89} = A'_{78}\, A_{89}$ and $B'_{89} = B'_{78} + A'_{78}B_{89}$
   g. $A'_{9,10} = A'_{89}\, A_{9,10}$ and $B'_{9,10} = B'_{89} + A'_{89}B_{9,10}$

EVALUATING THE INITIAL VARIANTS

Steps 1–5 above were completed for each variant of the vertical scale. The results based on these steps were presented to FDOE and FDOE's TAC. In addition, TDC reviewed the vertical linking items, their statistics, and content coverage.

For these initial variants, the forward linking method (Version B) identified a unique pattern that was not observed, to the same degree, as the other approaches. In both subjects the following characteristics were observed:

- Backwards linking yielded the largest growth.
- Forwards linking yielded the smallest growth.
- Mixed linking yielded growth that was intermediate compared to the backwards and forwards linking methods.
- Effect sizes in Mathematics were larger than those observed in ELA

Initial Variant results for Mathematics

The number of items included in each variant and the number of dropped items are shown in Table 2 and Table 3. The Stocking-Lord slopes are shown in Table 4, and the intercepts are shown in Table 5. Recall that the slopes or *A* values can be thought of as the standard deviation and that in general they increase over grades. The slopes or *B* values represent the means and in general they also increase over grades. The raw growth and effect sizes are given in Table 6 and Table 7 respectively.

Figure 1 is a graphical representation of the growth over grades for each of the initial variants for Mathematics, and Figure 2 shows the slope over grades.

In Mathematics, either low or negative growth patterns (as shown by the slope, *B*) were observed between grades 5 and 6 and grades 7 and 8 in all but one method. Students often transition to middle school between grades 5 and 6, and this could account for the low growth in these grades. Additionally, in grade 8, about half the population was administered the Algebra 1 test instead of the grade 8 Mathematics assessment, which is a likely cause of the low or negative growth pattern. Note that ELA did not exhibit negative growth for grades 7 and 8.

Table 2: Initial Variants, Number of Linking Items - Mathematics

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	30	30	60	58	57	55	48
G45	30	30	60	56	50	46	45
G56	30	30	60	55	36	34	39
G67	30	30	60	51	48	40	41
G78	30	30	60	54	29	27	48

Table 3: Initial Variants, Number of Dropped Items - Mathematics

Grade	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	2	3	5	12
G45	4	10	14	15
G56	5	24	26	21

Grade	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G67	9	12	20	19
G78	6	31	33	12

Table 4: Initial Variants, Linking Constants (M1/A) - Mathematics

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
3	0.73	1.10	0.77	0.78	0.77	0.76	0.84
4	0.85	1.13	0.87	0.87	0.88	0.87	0.90
5	0.97	1.08	0.97	0.97	1.00	0.99	0.92
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	1.02	0.93	0.99	0.97	0.99	0.98	0.95
8	0.88	0.72	0.81	0.78	0.92	0.89	0.82

Table 5: Initial Variants, Linking Constants (M2/B) - Mathematics

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
3	-1.76	-0.62	-1.20	-1.19	-1.52	-1.51	-1.11
4	-0.94	-0.10	-0.56	-0.55	-0.83	-0.83	-0.49
5	-0.31	0.08	-0.13	-0.11	-0.29	-0.29	-0.17
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.40	0.15	0.28	0.26	0.36	0.35	0.20
8	0.86	-0.05	0.38	0.37	0.87	0.84	0.29

Table 6: Initial Variants, Raw Growth - Mathematics

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	0.82	0.53	0.64	0.64	0.70	0.68	0.63
G45	0.63	0.18	0.43	0.43	0.53	0.54	0.32
G56	0.31	-0.08	0.13	0.11	0.29	0.29	0.17
G67	0.40	0.15	0.28	0.26	0.36	0.35	0.20
G78	0.46	-0.20	0.10	0.11	0.51	0.48	0.09

Table 7: Initial Variants, Effect Size - Mathematics

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	1.13	0.48	0.84	0.82	0.91	0.90	0.75
G45	0.74	0.16	0.50	0.50	0.61	0.62	0.36
G56	0.32	-0.07	0.13	0.12	0.29	0.29	0.18
G67	0.40	0.15	0.28	0.26	0.36	0.35	0.20
G78	0.46	-0.22	0.10	0.11	0.52	0.50	0.10

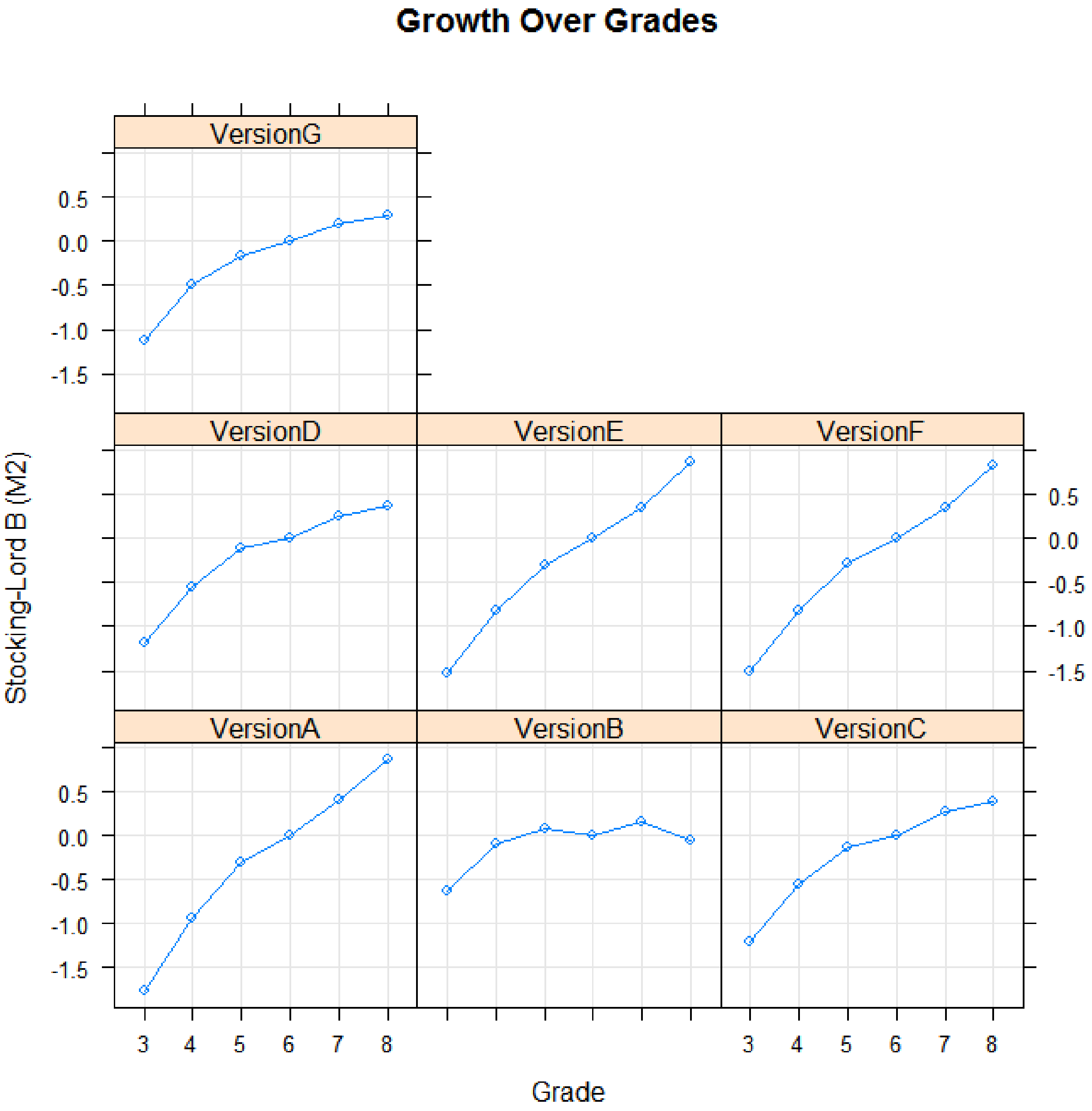


Figure 1: Growth Over Grades by Initial Variant - Mathematics

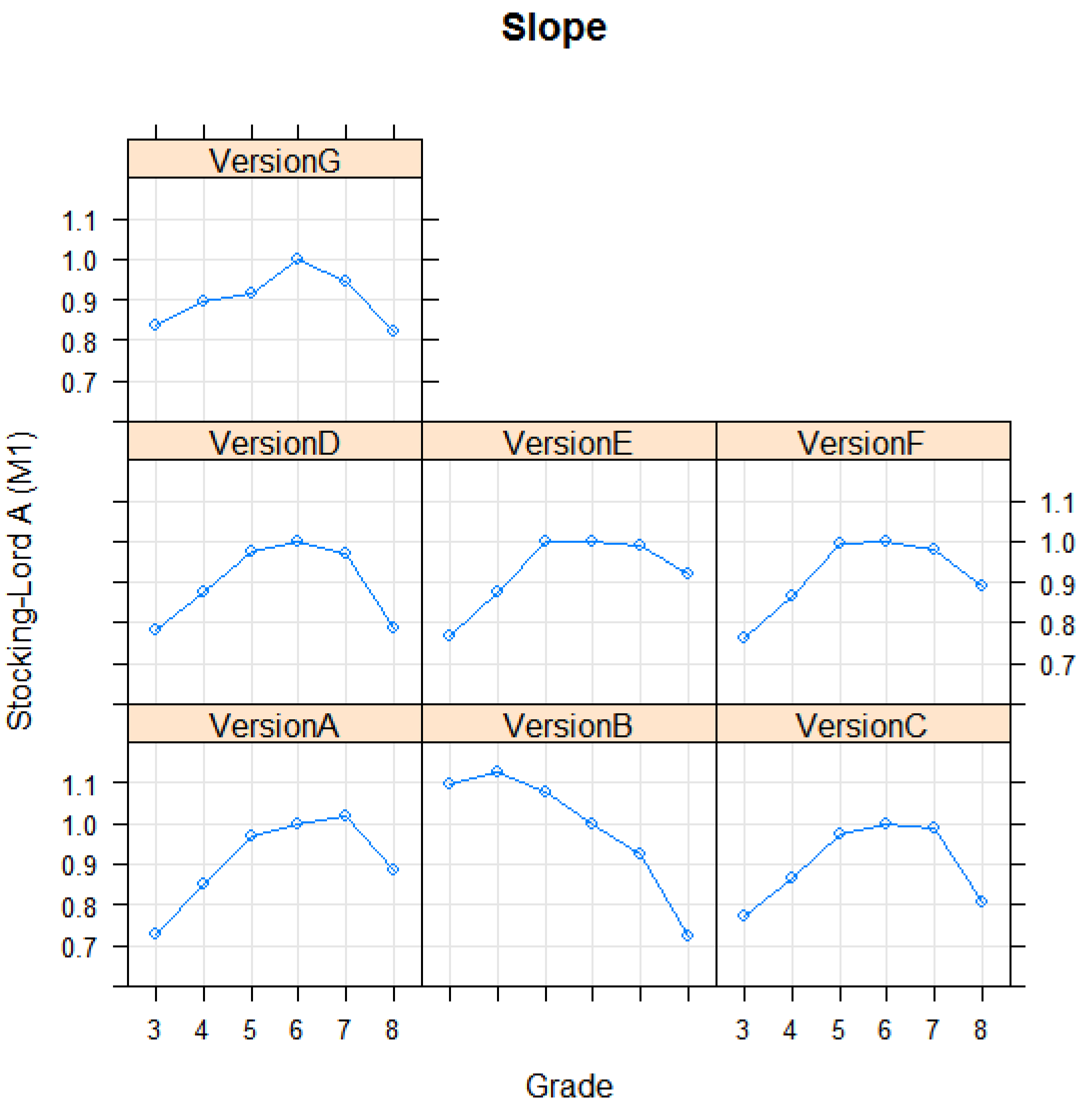


Figure 2: Slope Over Grades by Initial Variant – Mathematics

Initial Variant Results for ELA

For the initial ELA variants, the number of items included in each variant and the number of dropped items are shown in Table 8 and Table 9. The Stocking-Lord slopes are shown in Table 10, and the intercepts are shown in Table 11. In general the slope, which can be thought of as the standard deviation, increases over grades. The mean across grades, as given by the intercept, increases over grades. The raw growth and effect sizes are given in Table 12 and Table 13 respectively.

Furthermore, Figure 3 is a graphical representation of the growth over grades for each of the initial variants for ELA and Figure 4 shows the slope over grades.

Table 8: Initial Variants, Number of Linking Items - ELA

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	28	29	57	56	57	56	54
G45	29	28	57	57	54	54	53
G56	25	29	54	52	46	44	50
G67	26	25	51	42	46	40	46
G78	29	26	55	45	48	38	48
G89	28	29	57	50	49	42	52
G910	25	28	53	49	51	48	49

Table 9: Initial Variants, Number of Dropped Items - ELA

Grade	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	1	0	1	3
G45	0	3	3	4
G56	2	8	10	4
G67	9	5	11	5
G78	10	7	17	7
G89	7	8	15	5
G910	4	2	5	4

Table 10: Initial Variants, Linking Constants (M1/A) - ELA

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
3	0.91	0.89	0.88	0.88	0.88	0.87	0.89
4	0.90	0.88	0.87	0.87	0.87	0.86	0.88
5	0.91	0.92	0.90	0.90	0.89	0.89	0.90
6	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7	0.99	0.98	0.98	0.99	0.97	0.97	0.99
8	0.98	0.97	0.98	0.97	0.96	0.95	1.00
9	1.00	1.00	1.00	1.00	0.98	0.98	1.02
10	0.99	0.99	0.99	1.00	0.98	0.98	1.02

Table 11: Initial Variants, Linking Constants (M2/B) - ELA

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
3	-1.24	-1.01	-1.10	-1.10	-1.14	-1.14	-1.04
4	-0.69	-0.50	-0.59	-0.59	-0.63	-0.63	-0.52
5	-0.22	-0.10	-0.16	-0.16	-0.19	-0.19	-0.13
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.36	0.30	0.33	0.34	0.35	0.36	0.33
8	0.64	0.56	0.60	0.62	0.62	0.64	0.60
9	0.83	0.74	0.79	0.79	0.81	0.82	0.77
10	1.08	1.02	1.06	1.06	1.09	1.09	1.05

Table 12: Initial Variants, Raw Growth - ELA

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	0.55	0.51	0.52	0.52	0.52	0.51	0.52
G45	0.47	0.39	0.43	0.42	0.44	0.44	0.39
G56	0.22	0.10	0.16	0.16	0.19	0.19	0.13
G67	0.36	0.30	0.33	0.34	0.35	0.36	0.33
G78	0.29	0.26	0.27	0.28	0.27	0.28	0.27
G89	0.18	0.18	0.18	0.17	0.19	0.18	0.17
G910	0.26	0.28	0.27	0.27	0.28	0.27	0.27

Table 13: Initial Variants, Effect Size - ELA

Grade	Backward (A)	Forward (B)	Mixed (C)	Mixed-Q1 (D)	Mixed - p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)
G34	0.60	0.57	0.59	0.59	0.59	0.59	0.58
G45	0.52	0.45	0.49	0.49	0.51	0.51	0.44
G56	0.25	0.11	0.18	0.18	0.21	0.21	0.15
G67	0.36	0.30	0.33	0.34	0.35	0.36	0.33
G78	0.29	0.26	0.28	0.28	0.28	0.28	0.27
G89	0.19	0.19	0.19	0.18	0.20	0.19	0.17
G910	0.26	0.28	0.27	0.27	0.28	0.28	0.27

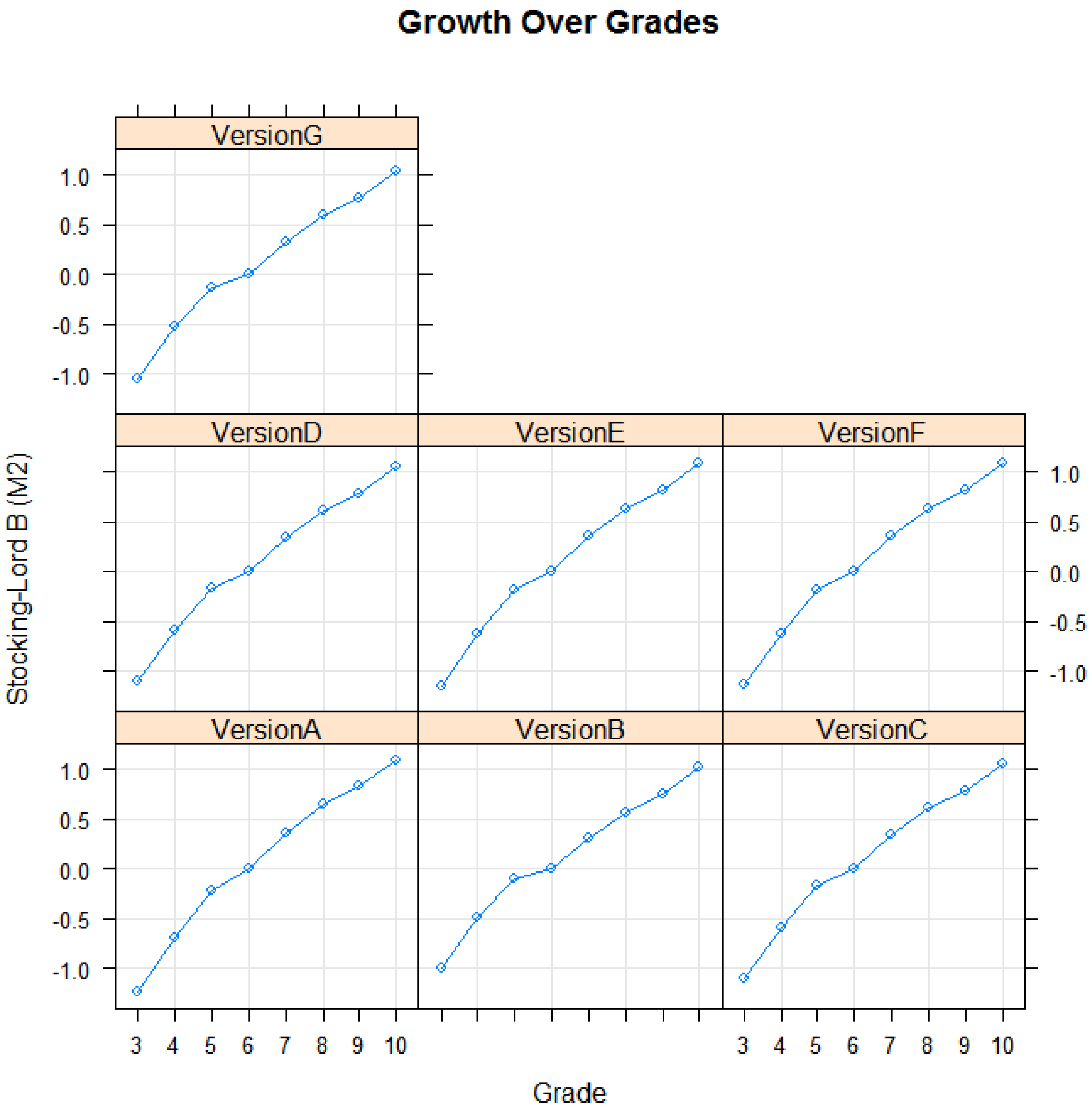


Figure 3: Growth Over Grades by Initial Variant - ELA

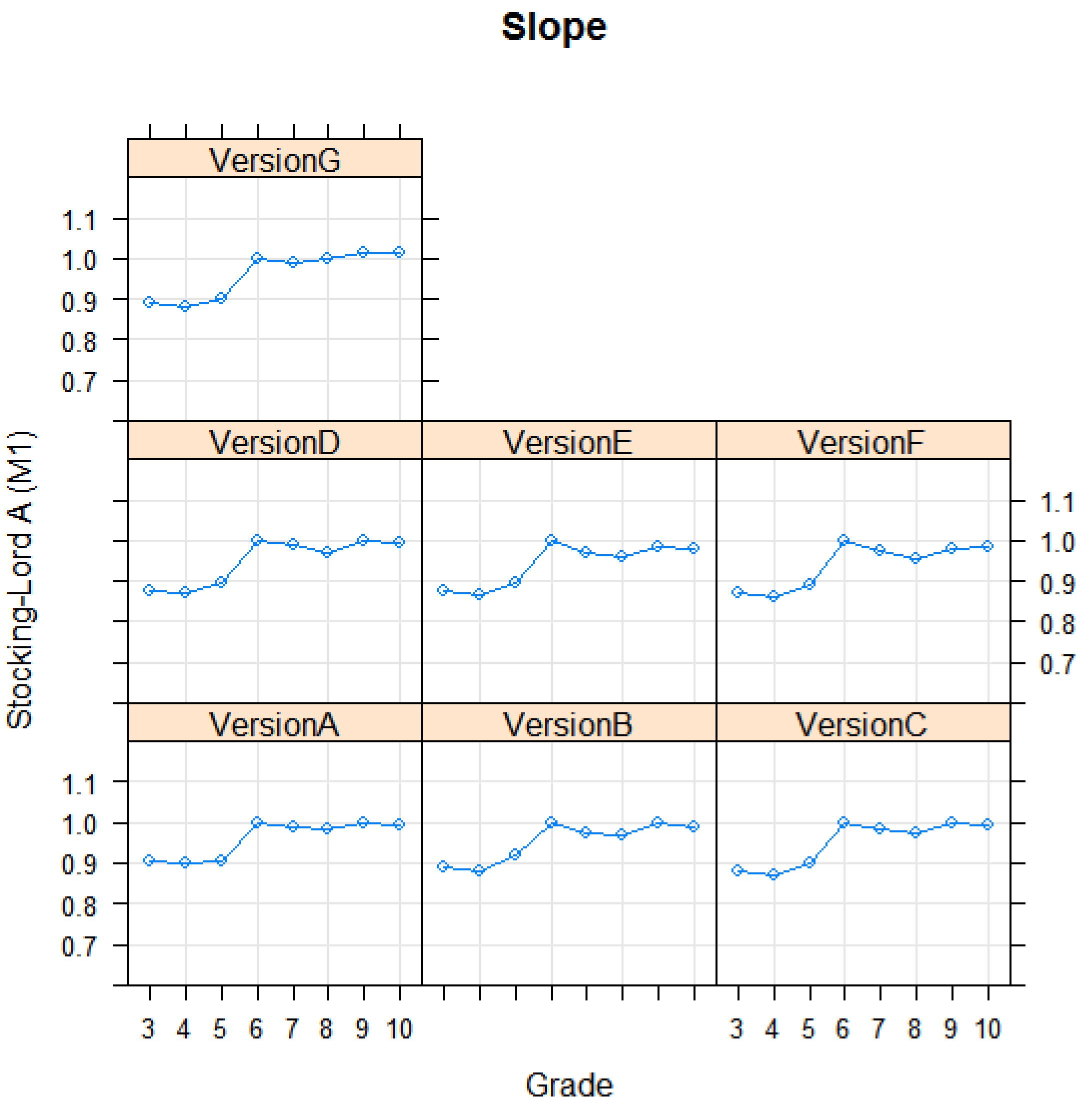


Figure 4: Slope Over Grades by Initial Variant – ELA

ADDITIONAL VARIANTS

Upon examination of the initial variants A-G and using the input from TDC and the TAC members, FDOE requested that AIR rerun versions C through G using grade 3 as the base grade, rather than grade 6 as previously described. Additionally, three new versions were requested. Furthermore, IRT calibration that contained operational-plus-vertical-linking items were linked to the operational-item-only calibration before any linking was complete.

Versions G1, G2, and H were similar to version G, except that they differed in the number of items dropped from the linking set. Version G1 implemented mixed linking, and items were dropped from the linking set based on Q1 statistics, D^2 statistics, convergence issues, point-biserial flags, and p-value flags. Version G2 implemented mixed linking as well, but items were dropped only according to Q1 statistics, D^2 statistics, and convergence issues.

After evaluating the performance of other variants, FDOE proposed version H, by listing a set of items to be excluded from the linking set with additional input from TDC on balancing the content specifications of the linking set. Version H is labeled as Final throughout the remainder of this document, as this was selected by FDOE to produce the FSA vertical scales.

Implementation Steps to Find Linking Constants – Additional Variants

This section lists the updated step-by-step procedures used to construct the linkages for the additional variants. The main differences between these steps and those listed for the initial variants occur in steps 4-6. Note that steps 1-3 are the same as the initial variants and were not repeated. For the additional variants, the operational-plus-vertical-linking item calibration was linked to the operational item only calibration in step 4, prior to any linking. Also, in steps 5 and 6, grade 3 was used as the base grade.

1) The data file was prepared for grade *g*. Data included only the operational items and the vertical linking items.
2) We conducted a separate calibration of the operational items administered in each grade using IRTPRO.
3) A second calibration was performed including the operational and vertical linking items.
4) We linked operational-plus-vertical-linking item calibrations to the operational-item-only calibration using the Stocking-Lord procedure to put vertical linking items on the scale of a given grade level.
5) For each version, chain linking was implemented via the Stocking-Lord procedure, removing flagged items if necessary, according to the following plan:
   i. Link grade 3 to grade 4 to find linking constants $\mathbf{A_{34}}$ and $\mathbf{B_{34}}$
   ii. Link grade 4 to grade 5 to find linking constants $\mathbf{A_{45}}$ and $\mathbf{B_{45}}$
   iii. Link grade 5 to grade 6 and identify linking constants $\mathbf{A_{56}}$ and $\mathbf{B_{56}}$
   iv. Link grade 6 to grade 7 and identify linking constants $\mathbf{A_{67}}$ and $\mathbf{B_{67}}$
   v. Link grade 7 to grade 8 to find linking constants $\mathbf{A_{78}}$ and $\mathbf{B_{78}}$
   vi. Link grade 8 to grade 9 to find linking constants $\mathbf{A_{89}}$ and $\mathbf{B_{89}}$
   vii. Link grade 9 to grade 10 to find linking constants $\mathbf{A_{9,10}}$ and $\mathbf{B_{9,10}}$
6) Linking constants were updated via the following transformations:
   i. $\boldsymbol{A'_{34} = A_{34}}$ and $\boldsymbol{B'_{34} = B_{34}}$
   ii. $\boldsymbol{A'_{45} = A'_{34}\, A_{45}}$ and $\boldsymbol{B'_{45} = B'_{34} + A'_{34}B_{45}}$
   iii. $\boldsymbol{A'_{56} = A'_{45}\, A_{56}}$ and $\boldsymbol{B'_{56} = B'_{45} + A'_{45}B_{56}}$
   iv. $\boldsymbol{A'_{67} = A'_{56}\, A_{67}}$ and $\boldsymbol{B'_{67} = B'_{56} + A'_{56}B_{67}}$

v. $\boldsymbol{A'_{78}} = \boldsymbol{A'_{67}}\ \boldsymbol{A_{78}}$ and $\boldsymbol{B'_{78}} = \boldsymbol{B'_{67}} + \boldsymbol{A'_{67}}\boldsymbol{B_{78}}$

vi. $\boldsymbol{A'_{89}} = \boldsymbol{A'_{78}}\ \boldsymbol{A_{89}}$ and $\boldsymbol{B'_{89}} = \boldsymbol{B'_{78}} + \boldsymbol{A'_{78}}\boldsymbol{B_{89}}$

vii. $\boldsymbol{A'_{9,10}} = \boldsymbol{A'_{89}}\ \boldsymbol{A_{9,10}}$ and $\boldsymbol{B'_{9,10}} = \boldsymbol{B'_{89}} + \boldsymbol{A'_{89}}\boldsymbol{B_{9,10}}$

EVALUATING THE ADDITIONAL VARIANTS

Additional Variant Results for Mathematics

Summary tables for the additional variants in Mathematics are found in Table 14 - Table 18. Figure 5 shows the new growth over grade by variant, and Figure 6 shows the slope over grade by variant. Historical effect sizes in Florida for 2011 and 2007, in addition to national effect sizes from 2007 (Dadey and Briggs, 2012), have been added to the last three columns of Table 18.

While there were no negative growth patterns, as shown by the intercept in Table 16, the same low growth patterns were observed. In general the slopes increase over grades, with the exception of grades 7 to 8.

Table 14: Additional Variants, Number of Linking Items - Mathematics

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
G34	60	58	57	55	48	54	55	54
G45	60	56	50	46	45	50	51	52
G56	60	55	36	34	39	52	54	53
G67	60	51	48	40	41	48	48	50
G78	60	54	29	27	48	48	52	38

Table 15: Additional Variants, Linking Constants (M1/A) - Mathematics

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.12	1.11	1.13	1.13	1.07	1.07	1.09	1.04
5	1.27	1.26	1.31	1.31	1.10	1.10	1.14	1.10
6	1.27	1.26	1.28	1.29	1.17	1.17	1.14	1.08
7	1.23	1.20	1.24	1.24	1.09	1.09	1.08	1.02
8	1.04	1.00	1.19	1.16	0.98	0.98	0.94	1.00

Table 16: Additional Variants, Linking Constants (M2/B) - Mathematics

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.82	0.80	0.89	0.88	0.73	0.78	0.78	0.68
5	1.39	1.38	1.59	1.60	1.13	1.21	1.21	1.09
6	1.61	1.57	2.03	2.03	1.38	1.35	1.36	1.26
7	1.93	1.87	2.45	2.45	1.58	1.61	1.61	1.51
8	2.04	1.99	3.10	3.07	1.67	1.66	1.64	1.65

Table 17: Additional Variants, Raw Growth - Mathematics

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
G34	0.82	0.80	0.89	0.88	0.73	0.78	0.78	0.68
G45	0.58	0.57	0.71	0.73	0.40	0.43	0.43	0.41
G56	0.22	0.20	0.43	0.43	0.25	0.14	0.15	0.17
G67	0.32	0.30	0.43	0.42	0.20	0.26	0.26	0.24
G78	0.12	0.12	0.64	0.61	0.09	0.05	0.03	0.14

Table 18: Additional Variants, Effect Size - Mathematics

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)	FL Hist. (2011)	FL Hist. (2007)	National (2007)
G34	0.82	0.80	0.89	0.88	0.73	0.78	0.78	0.68	0.60	0.38	0.53
G45	0.52	0.51	0.63	0.64	0.37	0.40	0.39	0.39	0.36	0.42	0.48
G56	0.17	0.16	0.33	0.32	0.22	0.13	0.13	0.16	0.26	0.06	0.39
G67	0.25	0.24	0.33	0.33	0.17	0.22	0.22	0.22	0.42	0.56	0.33
G78	0.09	0.10	0.52	0.50	0.08	0.05	0.03	0.14	0.34	0.43	0.33

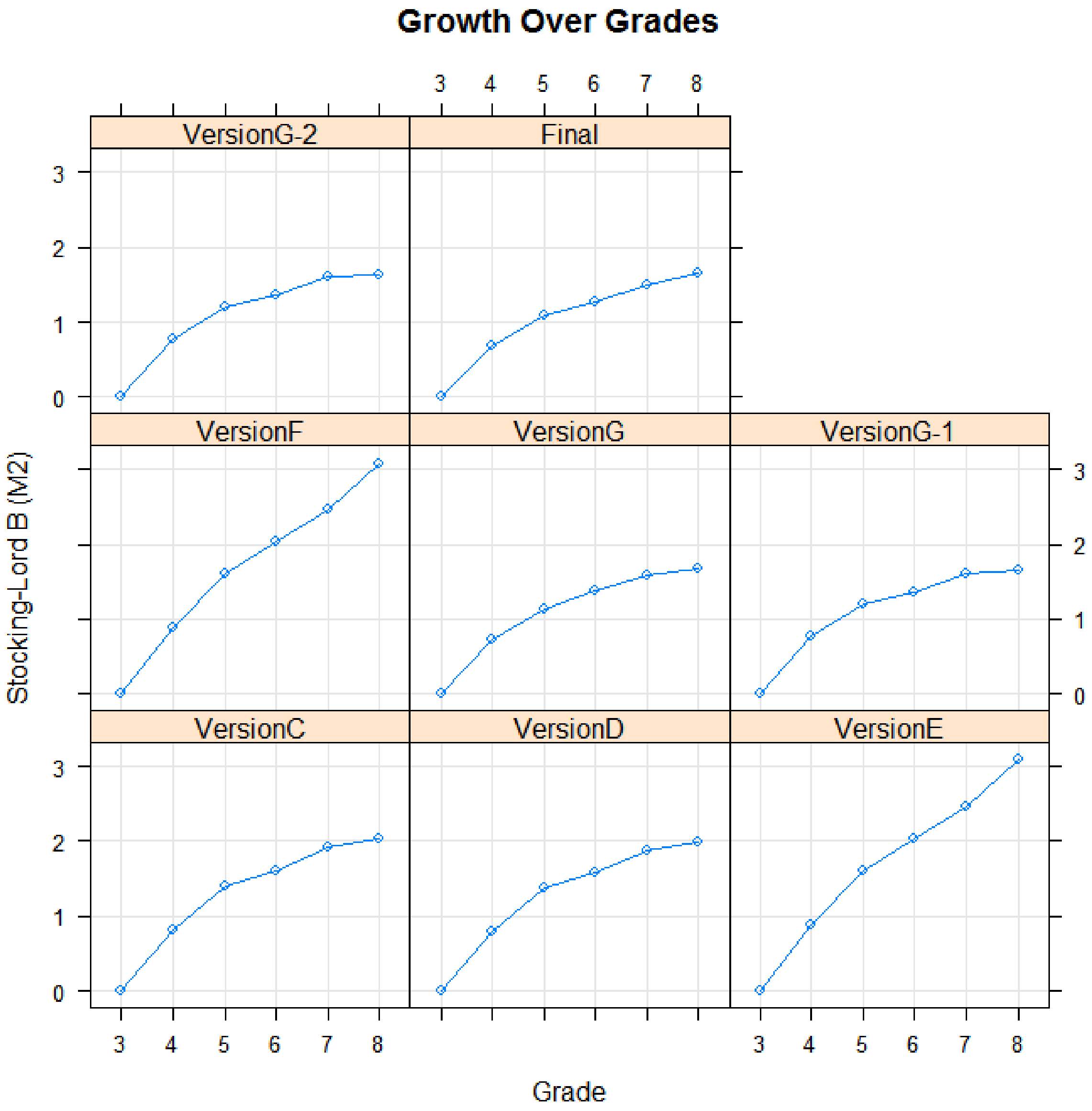


Figure 5: Growth Over Grades by Additional Variant – Mathematics

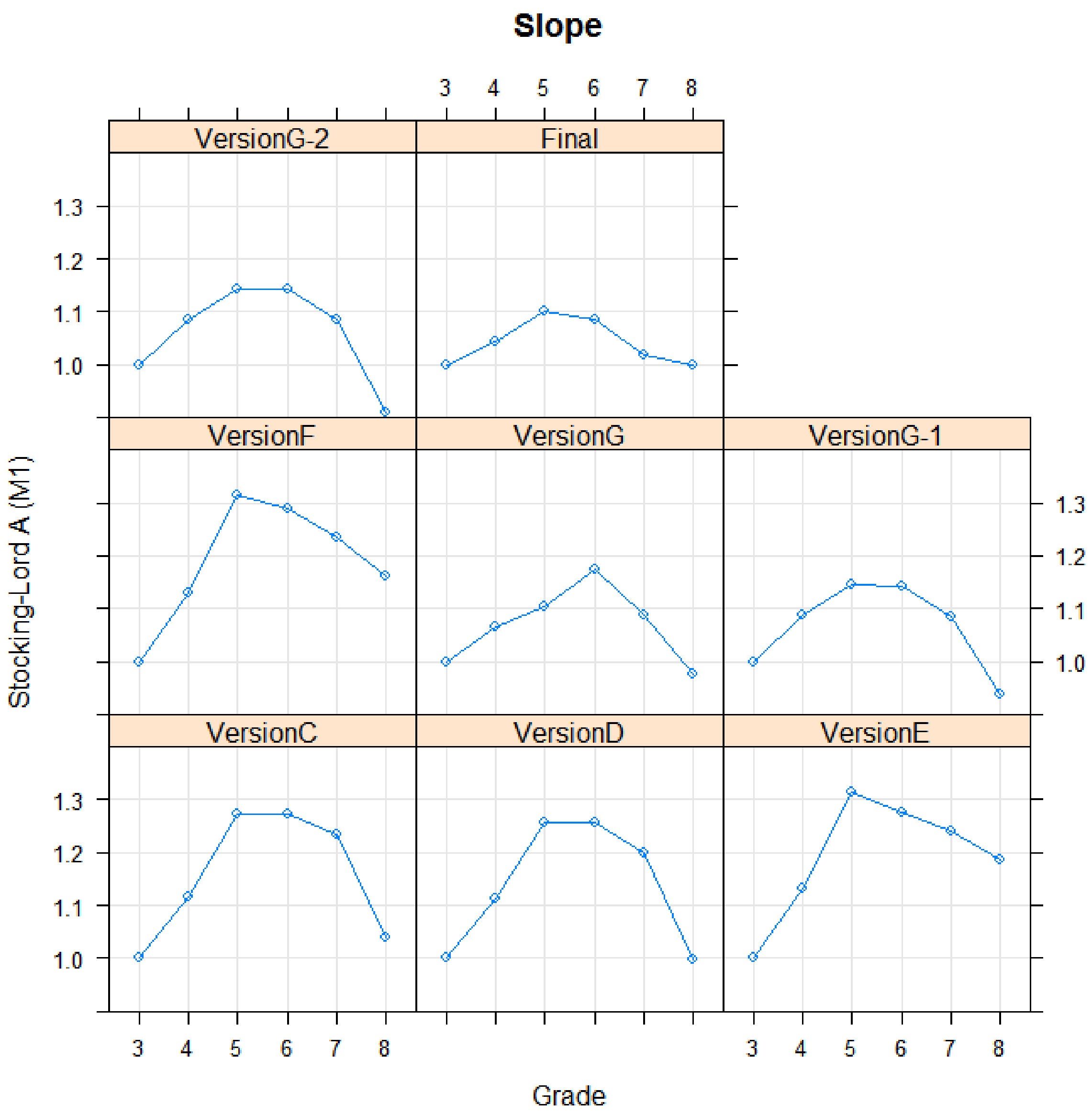


Figure 6: Slope Over Grades by Additional Variant – Mathematics

Additional Variant Results for ELA

Summary tables for the additional variants in ELA are found in Table 19 - Table 23. Figure 7 shows the growth over grade by variant, and Figure 8 shows the slope over grade by variant. In the last three columns of Table 23, the historical effect sizes in Florida for 2011 and 2007, in addition to national effect sizes from 2007 (Dadey and Briggs, 2012), have been added.

In general the slope remains the same or increases over grades. There are no negative growth patterns observed, as given by the intercept in Table 21.

Table 19: Additional Variants, Number of Linking Items - ELA

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
G34	57	52	57	52	53	50	51	52
G45	57	57	52	52	53	55	55	52
G56	54	51	45	42	49	46	48	45
G67	51	40	46	38	46	37	38	40
G78	53	42	47	36	48	39	40	38
G89	54	47	47	40	49	44	46	43
G910	52	48	49	46	47	45	46	48

Table 20: Additional Variants, Linking Constants (M1/A) - ELA

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.02	1.01	1.02	1.01	1.01	1.00	1.00	1.01
5	1.06	1.05	1.05	1.05	1.05	1.04	1.04	1.06
6	1.12	1.11	1.12	1.12	1.11	1.10	1.10	1.09
7	1.10	1.10	1.08	1.08	1.09	1.08	1.08	1.08
8	1.08	1.08	1.07	1.06	1.09	1.07	1.06	1.08
9	1.11	1.10	1.09	1.09	1.11	1.10	1.08	1.09
10	1.08	1.08	1.06	1.06	1.09	1.08	1.06	1.06

Table 21: Additional Variants, Linking Constants (M2/B) - ELA

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.57	0.57	0.57	0.57	0.57	0.56	0.56	0.57
5	1.06	1.06	1.10	1.10	1.01	1.01	1.01	1.05
6	1.30	1.29	1.37	1.37	1.21	1.20	1.20	1.25
7	1.64	1.65	1.73	1.73	1.56	1.55	1.55	1.61
8	1.97	1.98	2.06	2.06	1.88	1.87	1.86	1.92
9	2.16	2.15	2.25	2.25	2.04	2.02	2.01	2.09
10	2.49	2.48	2.60	2.60	2.38	2.35	2.34	2.42

Table 22: Additional Variants, Raw Growth - ELA

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)
G34	0.57	0.57	0.57	0.57	0.57	0.56	0.56	0.57
G45	0.49	0.49	0.53	0.53	0.44	0.45	0.45	0.48
G56	0.23	0.23	0.26	0.27	0.20	0.19	0.19	0.21
G67	0.35	0.36	0.37	0.38	0.35	0.35	0.35	0.35
G78	0.33	0.33	0.33	0.33	0.32	0.31	0.31	0.32
G89	0.19	0.17	0.19	0.19	0.17	0.15	0.16	0.17
G910	0.34	0.33	0.34	0.34	0.33	0.33	0.33	0.33

Table 23: Additional Variants, Effect Size - ELA

Grade	Mixed (C)	Mixed-Q1 (D)	Mixed-p-val (E)	Mixed-Q1, p-val (F)	Mixed (G)	Mixed (G1)	Mixed (G2)	Final (H)	FL Hist. 2011	FL Hist. 2007	Nat'l 2007
G34	0.57	0.57	0.57	0.57	0.57	0.56	0.56	0.57	0.53	0.57	0.55
G45	0.48	0.48	0.52	0.52	0.44	0.45	0.45	0.47	0.36	0.15	0.34
G56	0.22	0.22	0.25	0.25	0.19	0.18	0.18	0.19	0.29	0.25	0.28
G67	0.31	0.32	0.33	0.34	0.31	0.32	0.32	0.32	0.26	0.30	0.32
G78	0.30	0.30	0.30	0.31	0.29	0.29	0.28	0.29	0.23	0.57	0.29
G89	0.17	0.16	0.18	0.17	0.15	0.14	0.15	0.15	0.16	-	-
G910	0.30	0.30	0.32	0.31	0.30	0.30	0.30	0.30	0.24	-	-

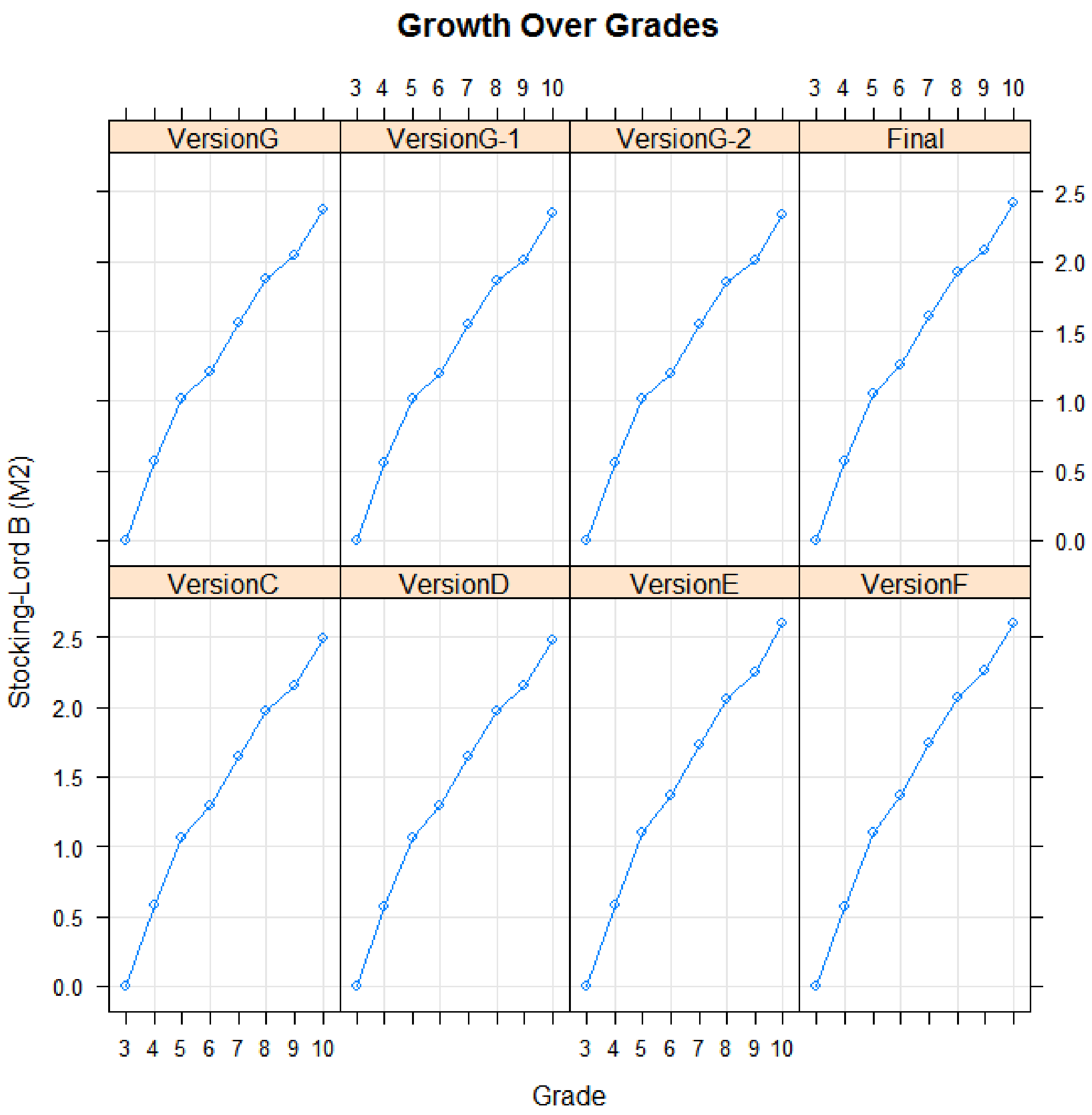


Figure 7: Growth Over Grades by Additional Variant *– ELA*

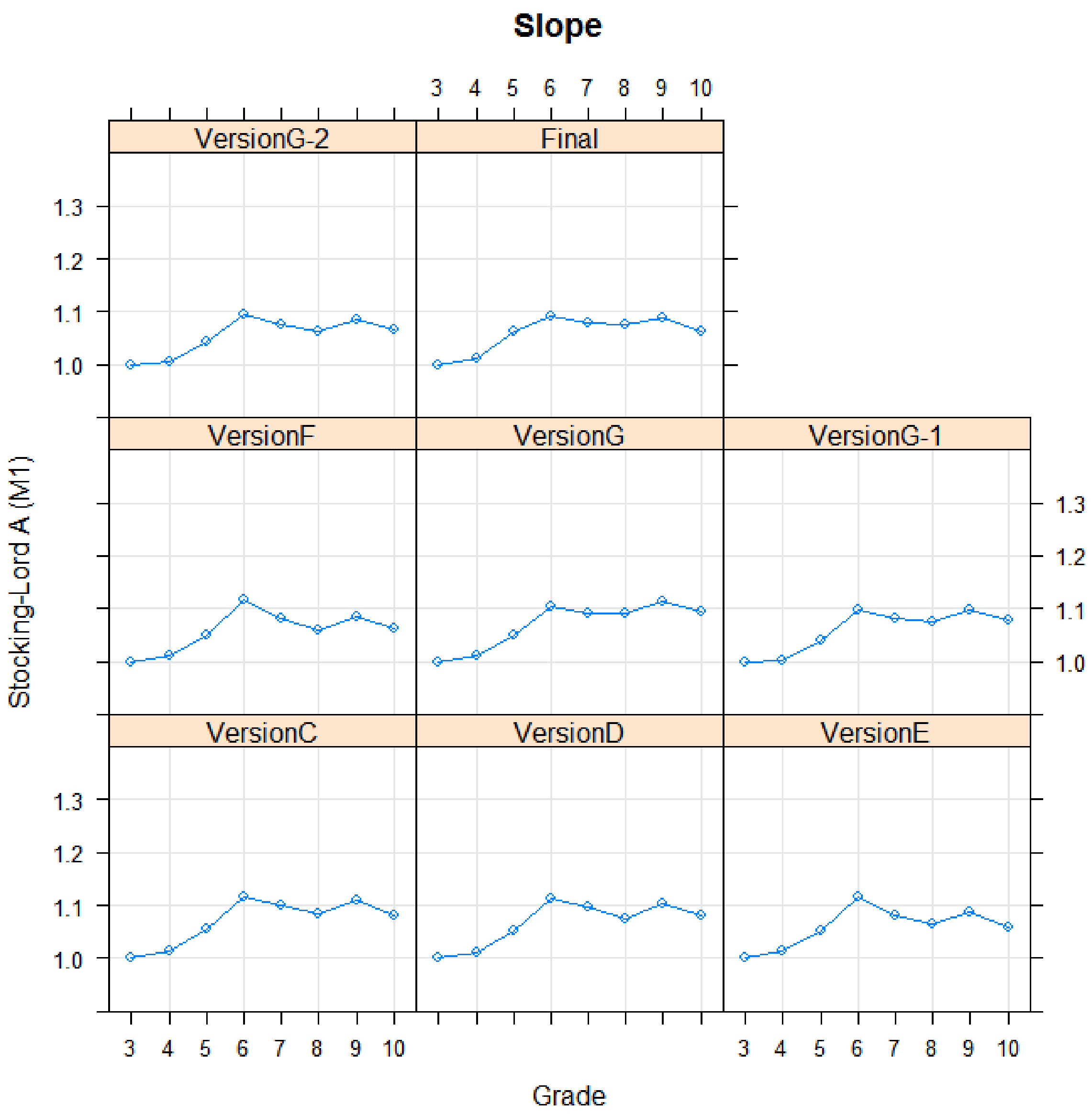


Figure 8: Slope Over Grades by Additional Variant – *ELA*

Final Vertical Scale: Variant H

Once versions C–G, G1, and G2 were updated and the additional version H was complete, AIR presented the results to FDOE. Based on the feedback from TAC and TDC's review of the vertical linking sets, FDOE selected final version H that created a smooth transition from one grade level to the next with an increasing intercept, and the vertical scale was established. Final vertical scaling constants are given in Table 24 and Table 25.

Item characteristic curves (ICCs) for the final version H can be found in Appendix C. For a given item the graph shows the ICCs for adjacent grades. The WRMSD listed is the D^2 value discussed under flagging criteria.

Appendix D shows the cumulative frequency plots for each of the additional variants. Appendix E shows the growth by subgroup for the final version H.

Table 24: Vertical Scaling Constants for FSA Mathematics

Grade	Slope (a)	Intercept (b)
3	1.000000	0.000000
4	1.044966	0.680890
5	1.102538	1.090128
6	1.084225	1.264961
7	1.018981	1.507877
8	0.997639	1.647321

Table 25: Vertical Scaling Constants for FSA ELA

Grade	Slope (a)	Intercept (b)
3	1.000000	0.000000
4	1.011871	0.570848
5	1.061502	1.048071
6	1.093056	1.253075
7	1.079095	1.606216
8	1.076568	1.921636
9	1.087592	2.087487
10	1.064215	2.416427

On-grade MLE estimates are converted to a vertically scaled theta as follows:

$$\theta_{VS} = a * \theta_G + b$$

where θ_{VS} is the vertical scale theta value, θ_G is the on-grade MLE estimate of theta, and a and b are the vertical scaling constants given in Table 24 and Table 25.

For a given grade and subject in ELA and Mathematics, the on-grade theta to on-grade scale score transformation equation is

$$SS_G = A * \theta_G + B$$

where $A = 20$ and $B = 300$ for all grades. Replacing the on-grade theta with the vertically scaled theta will yield the vertical scale score. The vertical theta can be replaced using the equation above to find the final on-grade theta to vertical scale score transformation equation.

$$SS_{VS} = A * \theta_{VS} + B$$

$$SS_{VS} = A * (a * \ \theta_G + b) + B$$

$$SS_{VS} = A * a * \ \theta_G + (A * b + B)$$

$$SS_{VS} = 20 * a * \ \theta_G + (20 * b + 300)$$

$$SS_{VS} = a' * \ \theta_G + b'$$

Applying the vertical scaling constants, the final intercept and slope are provided in Table 26 and Table 27.

Table 26: Intercept and Slope Values for FSA Mathematics

Grade	**Slope ($\boldsymbol{a}'$)**	**Intercept ($\boldsymbol{b}'$)**
3	20.000000	300.000000
4	20.899320	313.617800
5	22.050760	321.802560
6	21.684500	325.299220
7	20.379620	330.157540
8	19.952780	332.946420

Table 27: Intercept and Slope Values for FSA ELA

Grade	**Slope ($\boldsymbol{a}'$)**	**Intercept ($\boldsymbol{b}'$)**
3	20.000000	300.000000
4	20.237420	311.416960
5	21.230040	320.961420
6	21.861120	325.061500
7	21.581900	332.124320
8	21.531360	338.432720
9	21.751840	341.749740
10	21.284300	348.328540

REFERENCES

Dadey, N. & Briggs, D. C. (2012). A Meta-Analysis of Growth Trends from Vertically Scaled Assessments. *Practical Assessment, Research & Evaluation*, 17 (14).

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. (2nd ed.) New York, NY: Springer.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201 - 210.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

NATIONAL BENCHMARKS

For State Achievement Standards

February 22, 2016

Gary W. Phillips
Vice President and Institute Fellow
American Institutes for Research

Contents

List of Tables

Executive Summary

State achievement standards represent how much the state expects their students to learn in order to reach various levels of academic proficiency. In this study, the academic subjects are English language arts (ELA) and mathematics. In the past, these achievement standards were used by each state to report adequate yearly progress (AYP) under No Child Left Behind (NCLB) federal legislation, and they are currently being used for federal reporting under the Every Student Succeeds Act (ESSA) of 2015. These standards are also used by the state to monitor progress from year to year and to report on the success of each classroom, school, and district to parents and the public.

This report uses national benchmarking as a common metric to examine state achievement standards and compare how high these standards are compared to the National Assessment of Educational Progress (NAEP) achievement levels. It also compares how much students are expected to learn in some states with how much they are expected to learn in other states. The study uses NAEP grades 4 and 8 reading and mathematics as benchmarks for individual state achievement standards. The study also benchmarks the achievement standards of Smarter Balanced Assessment Consortium (referred to in this study as *Smarter Balanced*), Partnership for Assessment of Readiness for College and Careers (PARCC), and ACT Aspire. Benchmarking Smarter Balanced, PARCC, and ACT Aspire provides a common metric (i.e., the NAEP scale) that can be used to compare the stringency of their achievement standards. The most important findings in the study relate to achievement standards that represent college readiness. Each of these consortium tests in grades 4 and 8 has achievement standards that indicate the student is on track to be college ready. The college-ready standards are Level 3 (Met) for Smarter Balanced, Level 4 (Met) for PARCC, and Level 3 (Ready) for ACT Aspire.

The overall findings in the study are:

1. Smarter Balanced college-ready standards (Level 3) are comparable in difficulty to the NAEP Basic levels.
2. Smarter Balanced college-ready standards (Level 3) are significantly below PARCC college-ready standards (Level 4) by about one-quarter of a standard deviation. In the statistical literature, a standard deviation unit is referred to as an *effect size*. The effect sizes are for ELA grades 4 and 8, and mathematics grades 4 and 8 are −.26, −.28, −.26 and −.36, respectively.
3. Smarter Balanced college-ready grade 8 standards are comparable to ACT Aspire college-ready grade 8 standards. However, for grade 4, the Smarter Balanced college-ready standard is significantly below the ACT Aspire college-ready standard for Reading (effect size = −.26) but significantly above the ACT Aspire college-ready standard for mathematics (effect size = +.29).
4. PARCC college-ready standards (Level 4) are comparable in difficulty to the NAEP Basic level for ELA and comparable to the NAEP Proficient level for mathematics.
5. PARCC college-ready standards (Level 4) are comparable in difficulty to the ACT Aspire college-ready standard for Reading grade 4. However, PARCC standards are significantly

above ACT Aspire college-ready standards for ELA grade 8 (effect size = +.28), mathematics grade 4 (effect size = +.55), and mathematics grade 8 (effect size = +.48).

6. ACT Aspire college-ready standards (Ready) are comparable in difficulty to the NAEP Basic levels.
7. Individual states that have college-readiness standards that map to the NAEP Proficient level are:

    a. ELA grade 4—Florida and New York;

    b. ELA grade 8—Florida, Kansas, and New York;

    c. Mathematics grade 4—Florida and Kansas; and

    d. Mathematics grade 8—Alaska, Florida, Kansas, New York, and Pennsylvania.

Note that Iowa, Nebraska, and Texas have three achievement levels, instead of the usual four levels or five levels in other states. At the time of this report, the author was unable to determine which levels in these states represented college readiness.

Brief History of Common Core–Related Activities

Role of NCLB: Probably the biggest contributor to the development of the Common Core State Standards (CCSS) was the passage of the No Child Left Behind Act of 2001. A fundament problem with NCLB demonstrated the need for the CCSS. NCLB required each state to have challenging content standards and performance standards but left it up to the state to define what "challenging" meant. Some states used low standards in order to report higher levels of proficiency. States with low standards were living in a kind of Lake Wobegon world where more and more students were being reported as proficient but fewer and fewer students were prepared for college. This led the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) to see if there was a way to make state standards more competitive and consistent.

Role of NGA and CCSSO: In 2006–2007, Arizona Governor Janet Napolitano chaired the NGA. In order to find a way to make America's educational system internationally competitive, she created a task force of state and national education policy leaders that released a report titled "Benchmarking for Success: Ensuring U.S. Students Receive a World-Class Education" (2008). The state leaders responsible for the report were the NGA and the CCSSO as well as the nonprofit group Achieve. The concepts in this report caught on, and in 2009 state leaders launched CCSS. These three groups obtained the support of other organizations that were critical in the development of the CCSS. These organizations included the American Federation of Teachers, the National Education Association, the National Council of Teachers of Mathematics, the National Council of Teachers of English, and the International Reading Association.

Role of U.S. Federal Government: The CCSS was a state-led effort and was not initiated by the federal government. The NGA and the CCSSO received no financial support from the federal government to develop the CCSS. However, once CCSS was developed, the federal government used the bully pulpit to encourage many states to implement internationally competitive common standards. For example, in 2009 President Obama, in a speech to the U.S. Hispanic Chamber of Commerce, recognized the need for high and consistent standards. He stated:

> *Let's challenge our states to adopt world-class standards that will bring our curriculums into the 21st century. Today's system of 50 different sets of benchmarks for academic success means fourth-grade readers in Mississippi are scoring nearly 70 points lower than students in Wyoming—and getting the same grade.*

The federal government also provided seed money to help states implement common standards. The funding was provided in the 4.35 billion dollar Race to the Top grant as part of the American Recovery and Reinvestment Act of 2009, which was part of the federal economic stimulus package.

Role of Smarter Balanced and PARCC: Part of the Race to the Top grant was awarded to PARCC and Smarter Balanced to develop tests that measure the CCSS. Over several years of development, some states dropped out of the initiative. By spring 2015, 18 states had given the first operational administration of the Smarter Balanced assessment, and 11 states plus the District of Columbia gave the first operational administration of the PARCC assessment. These are the jurisdictions on which the current consortium results are based. The Virgin Islands were also administered the Smarter Balanced assessment, but they were excluded in this mapping study because they did not participate in the 2015 NAEP assessment.

ACT Aspire: In 2015, ACT Aspire was administered in two states—Alabama and South Carolina—which represents a group of states taking the same assessment. Recognizing that a large portion of students were graduating high school unprepared for college, ACT developed an assessment that was built around college readiness beginning in elementary school. The ACT Aspire replaced the ACT Explore (grades 8 and 9) and ACT Plan (grade 10) and was administered in grades 3–10.

Benchmarking State Achievement Standards

Benchmarking is a way to calibrate the difficulty level of state achievement standards so they can be compared to each other and to national standards. This type of benchmarking is similar to benchmarking in business and industry. For example, the fuel efficiency and quality of American-built cars are often benchmarked against those of cars built in Japan and South Korea. Such benchmarking is important in education if we are to expect our students to compete in a global economy. In this study, we use the NAEP as a national benchmark.

Some terminology clarification is needed in order to navigate through the results of this study. This report is about benchmarking (or comparing) state achievement standards (cut-scores on the state accountability test used to report results to the federal government under ESSA) to the NAEP achievement levels. In some testing programs, achievement standards are referred to as *performance standards*. The comparisons are obtained through equipercentile linking (described in the Appendix). An achievement standard is a specific number, or cut-score, on the scale such as those in Tables 1–3. What this study does is determine the NAEP equivalent of the state achievement standard (or cut-score) and report the NAEP achievement level in which the NAEP equivalent falls. For example, the Smarter Balanced ELA grade 4 cut-score for Level 3 is 2473 (see One caveat in the study is that for Smarter Balanced and PARCC we are mapping *ELA* standards, which include writing, to NAEP *Reading* standards, which do not include writing. This should not make much difference because, generally, the dis-attenuated correlations between reading and writing are very high.

Table 1). The linking analysis shows this is equivalent in difficulty to a NAEP score of 222 (see Table 6). The NAEP equivalent of 222 falls within the range of the NAEP Basic level (208-237; see Table 4).

Grades 4 and 8 Achievement Standards for Smarter Balanced, PARCC, ACT Aspire, and NAEP

Each of the assessments used by groups of states in 2015 has its own achievement standards. In each case, the standards were set through a consortium or national consensus process and represent how much we expect students to know and be able to do at different levels of achievement. Possibly the most important achievement standard is the one that indicates the student is on track to be college ready by the end of high school. For Smarter Balanced this is Level 3, for PARCC this is Level 4, and for ACT Aspire this is Level 3. The achievement standards for each assessment—Smarter Balanced, PARCC, ACT Aspire, and NAEP—are indicated in Tables 1–4.

One caveat in the study is that for Smarter Balanced and PARCC we are mapping *ELA* standards, which include writing, to NAEP *Reading* standards, which do not include writing. This should not make much difference because, generally, the dis-attenuated correlations between reading and writing are very high.

Table 1: Smarter Balanced Achievement Standards

Subject	Grade	Level 2 Nearly Met	Level 3 Met	Level 4 Exceeded
ELA	4	2416	2473	2533
ELA	8	2487	2567	2668
Mathematics	4	2411	2485	2549
Mathematics	8	2504	2586	2653

Table 2: PARCC Performance Standards

Subject	Grade	Level 2 Partially Met	Level 3 Approached	Level 4 Met	Level 5 Exceeded
ELA	4	700	725	750	790
ELA	8	700	725	750	794
Mathematics	4	700	725	750	796
Mathematics	8	700	725	750	801

Table 3: ACT Aspire Achievement Standards

Subject	Grade	Level 2 Close	Level 3 Ready	Level 4 Exceeding
Reading	4	412	417	422
Reading	8	418	424	430
Mathematics	4	411	416	421
Mathematics	8	419	425	431

Table 4: NAEP Achievement Standards

Subject	Grade	Basic	Proficient	Advanced
Reading	4	208	238	268
Reading	8	243	281	323
Mathematics	4	214	249	282
Mathematics	8	262	299	333

Using NAEP as a National Benchmark

NAEP represents probably the best assessment against which to benchmark state achievement standards. First, the NAEP content standards and achievement standards were developed through an elaborate national process that has been exhaustively evaluated. NAEP standards have been demonstrated to be internationally competitive and are often referred to as the gold standard against which other standards can be compared. Second, NAEP provides biennial state representative assessments that can be treated as randomly equivalent to the local state testing population. This facilitates comparisons between local state testing results and state NAEP testing results. Third, because NAEP is administered in each state, the NAEP scale can be used as an anchor test to provide a common metric to compare local state-by-state testing results. This was the strategy used in this study.

National NAEP Benchmarks for Smarter Balanced

In 2015, 18 states and the Virgin Islands administered the Smarter Balanced assessment. Because they all used the same test, a weighted average of the percentage at and above each achievement level for the 18 states was used for the analysis. The weights were based on the student population size in each state. The 18 jurisdictions were California, Connecticut, Delaware, Hawaii, Idaho, Maine, Michigan, Missouri, Montana, Nevada, New Hampshire, North Dakota, Oregon, South Dakota, Vermont, Washington, West Virginia, and Wisconsin. The Virgin Islands were excluded because they did not participate in NAEP in 2015. For ELA, Wisconsin and Missouri were excluded from the weighted average because their administration deviated from the Smarter Balanced blueprint. North Dakota was excluded for both ELA and mathematics because the author was unable to find their results on their state web site. Aggregate NAEP estimates for the Smarter Balanced states were obtained from the NAEP Data Explorer (NDE).

The national NAEP benchmarks for Smarter Balanced are contained in Tables 5–7. The most important Smarter Balanced level to benchmark is Level 3, considered to represent being on track to be college ready. We see in Table 6 that each of the Smarter Balanced Level 3 cut-scores maps to the NAEP Basic achievement level.

Table 5: NAEP Equivalents of Smarter Balanced Achievement Standards for Level 2

Subject	Grade	Smarter Cut-Score for Level 2 Nearly Met	Percent at and Above Smarter Level 2 Nearly Met	NAEP Scaled Score Equivalent of Level 2 Nearly Met	Standard Error of NAEP Equivalent of Level 2 Nearly Met	NAEP Achievement Level Equivalent of Level 2 Nearly Met
ELA	4	2416	66	201	2.0	Below Basic
ELA	8	2487	77	236	1.0	Below Basic
Math	4	2411	74	216	1.0	Basic
Math	8	2504	61	269	1.0	Basic

Table 6: NAEP Equivalents of Smarter Balanced Achievement Standards for Level 3

Subject	Grade	Smarter Cut-Score for Level 3 Met	Percent at and Above Smarter Level 3 Met	NAEP Scaled Score Equivalent of Level 3 Met	Standard Error of NAEP Equivalent of Level 3 Met	NAEP Achievement Level Equivalent of Level 3 Met
ELA	4	2473	44	222	2.0	Basic
ELA	8	2567	48	264	1.0	Basic
Math	4	2485	40	244	1.0	Basic
Math	8	2586	35	294	1.0	Basic

Table 7: NAEP Equivalents of Smarter Balanced Achievement Standards for Level 4

Subject	Grade	Smarter Cut-Score for Level 4 Exceeded	Percent at and Above Smarter Level 4 Exceeded	NAEP Scaled Score Equivalent of Level 4 Exceeded	Standard Error of NAEP Equivalent of Level 4 Exceeded	NAEP Achievement Level Equivalent of Level 4 Exceeded
ELA	4	2533	22	247	2.0	Proficient
ELA	8	2668	13	302	1.0	Proficient
Math	4	2549	15	268	1.0	Proficient
Math	8	2653	17	315	1.0	Proficient

National NAEP Benchmarks for PARCC

In 2015, 11 states and the District of Columbia administered the PARCC assessment. Because they all used the same test, a weighted average of the percentage at and above each achievement level for the 12 jurisdictions was used for the analysis. The weights were based on the student population size in each state. The 12 jurisdictions were Arkansas, Colorado, District of Columbia, Illinois, Louisiana, Maryland, Massachusetts, Mississippi, New Jersey, New Mexico, Ohio, and Rhode Island. The aggregate NAEP estimate for the PARCC jurisdictions was obtained from the NDE.

The national NAEP benchmarks for PARCC are contained in Tables 8–11. The most important PARCC level to benchmark is Level 4, considered to represent being on track to be college ready. We see in Table 10 that each of the PARCC Level 4 cut-scores maps to the NAEP Basic achievement level for ELA and the NAEP Proficient achievement level for mathematics.

Table 8: NAEP Equivalents of PARCC Performance Standards for Level 2

Subject	Grade	PARCC Cut-Score for Level 2 Partially Met	Percent at and Above PARCC Level 2 Partially Met	NAEP Scaled Score Equivalent of Level 2 Partially Met	Standard Error of NAEP Equivalent of Level 2 Partially Met	NAEP Achievement Level Equivalent of Level 2 Partially Met
ELA	4	700	89	179	1.0	Below Basic
ELA	8	700	86	229	1.0	Below Basic
Math	4	700	88	200	1.0	Below Basic
Math	8	700	78	255	1.0	Below Basic

Table 9: NAEP Equivalents of PARCC Performance Standards for Level 3

Subject	Grade	PARCC Cut-Score for Level 3 Approached	Percent at and Above PARCC Level 3 Approached	NAEP Scaled Score Equivalent of Level 3 Approached	Standard Error of NAEP Equivalent of Level 3 Approached	NAEP Achievement Level Equivalent of Level 3 Approached
ELA	4	725	70	205	1.0	Below Basic
ELA	8	725	67	250	1.0	Basic
Math	4	725	62	228	1.0	Basic
Math	8	725	52	282	1.0	Basic

Table 10: NAEP Equivalents of PARCC Performance Standards for Level 4

Subject	Grade	PARCC Cut-Score for Level 4 Met	Percent at and Above PARCC Level 4 Met	NAEP Scaled Score Equivalent of Level 4 Met	Standard Error of NAEP Equivalent of Level 4 Met	NAEP Achievement Level Equivalent of Level 4 Met
ELA	4	750	41	232	1.0	Basic
ELA	8	750	42	273	1.0	Basic
Math	4	750	32	252	1.0	Proficient
Math	8	750	27	307	1.0	Proficient

Table 11: NAEP Equivalents of PARCC Performance Standards for Level 5

Subject	Grade	PARCC Cut-Score for Level 5 Exceeded	Percent at and Above PARCC Level 5 Exceeded	NAEP Scaled Score Equivalent of Level 5 Exceeded	Standard Error of NAEP Equivalent of Level 5 Exceeded	NAEP Achievement Level Equivalent of Level 5 Exceeded
ELA	4	790	7	277	1.0	Advanced
ELA	8	794	7	318	1.0	Proficient
Math	4	796	3	297	2.0	Advanced
Math	8	801	3	358	1.0	Advanced

National NAEP Benchmarks for ACT Aspire

In 2015, two states administered the ACT Aspire test. They were Alabama and South Carolina. Because both states used the same test, a weighted average of the percentage at and above each achievement level was used for the analysis. The weights were based on the student population size in each state. The aggregate NAEP estimate for the ACT Aspire jurisdictions was obtained from the NDE. The national NAEP benchmarks for ACT Aspire are contained in Tables 12–14. The most important ACT Aspire level to benchmark is Level 3, considered to represent being on track to be college ready. We see in Table 13 that each of the ACT Aspire college-ready cut-scores map to the NAEP Basic achievement level.

Table 12: NAEP Equivalents of ACT Aspire Achievement Standards Level 2

Subject	Grade	ACT Aspire Cut-Score for Level 2 Close	Percent at and Above ACT Aspire Level 2 Close	NAEP Scaled Score Equivalent of Level 2 Close	Standard Error of NAEP Equivalent of Level 2 Close	NAEP Achievement Level Equivalent of Level 2 Close
Reading	4	412	67	202	2.0	Below Basic
Reading	8	418	72	240	1.0	Below Basic
Math	4	411	91	195	1.0	Below Basic
Math	8	419	59	263	2.0	Basic

Table 13: NAEP Equivalents of ACT Aspire Achievement Standards Level 3

Subject	Grade	ACT Aspire Cut-Score for Level 3 Ready	Percent at and Above ACT Aspire Level 3 Ready	NAEP Scaled Score Equivalent of Level 3 Ready	Standard Error of NAEP Equivalent of Level 3 Ready	NAEP Achievement Level Equivalent of Level 3 Ready
Reading	4	417	35	232	2.0	Basic
Reading	8	424	45	264	1.0	Basic
Math	4	416	49	235	1.0	Basic
Math	8	425	30	290	2.0	Basic

Table 14: NAEP Equivalents of ACT Aspire Achievement Standards Level 4

Subject	Grade	ACT Aspire Cut-Score for Level 4 Exceeding	Percent at and Above ACT Aspire Level 4 Exceeding	NAEP Scaled Score Equivalent of Level 4 Exceeding	Standard Error of NAEP Equivalent of Level 4 Exceeding	NAEP Achievement Level Equivalent of Level 4 Exceeding
Reading	4	422	13	260	2.0	Proficient
Reading	8	430	13	298	2.0	Proficient
Math	4	421	14	266	1.0	Proficient
Math	8	431	14	309	2.0	Proficient

National NAEP Benchmarks for Nonconsortium States

Across most of the nonconsortium states with four achievement levels, Level 3 is considered on track to be college ready. For many states with five achievement levels, Level 4 is considered on track to be college ready. However, this is not universally true. For Indiana, the author was not able to obtain the 2015 state results at the present time.

The results of NAEP benchmarks for ELA grade 4 individual states are reported in

Table 15. The reading grade 4 NAEP achievement level cut-scores are Basic = 208, Proficient = 238, and Advanced = 268. The only state with four achievement levels for which Level 3 maps to the NAEP Proficient level is New York. The only state with five achievement levels for which Level 4 maps to the NAEP Proficient level is Florida.

Table 15: ELA Grade 4 NAEP Benchmarks for Nonconsortium States

State	ELA Grade 4											
	Level 2			Level 3			Level 4			Level 5		
	Percent at and Above Level 2	NAEP Equivalent for Level 2	NAEP Achievement Level Comparable to Level 2	Percent at and Above Level 3	NAEP Equivalent for Level 3	NAEP Achievement Level Comparable to Level 3	Percent at and Above Level 4	NAEP Equivalent for Level 4	NAEP Achievement Level Comparable to Level 4	Percent at and Above Level 5	NAEP Equivalent for Level 5	NAEP Achievement Level Comparable to Level 5
Alaska	59	203	Below Basic	40	224	Basic	9	271	Advanced			
Arizona	59	206	Below Basic	42	223	Basic	6	277	Advanced			
DoDEA	93	193	Below Basic	72	217	Basic	37	243	Proficient			
Florida	79	202	Below Basic	54	224	Basic	27	246	Proficient	8	271	Advanced
Georgia	71	204	Below Basic	37	233	Basic	9	267	Proficient			
Iowa	76	198	Below Basic	29	244	Proficient						
Kansas	88	176	Below Basic	55	217	Basic	11	269	Advanced			
Kentucky	81	199	Below Basic	52	226	Basic	14	263	Proficient			
Minnesota	79	192	Below Basic	58	216	Basic	18	259	Proficient			
Nebraska	81	196	Below Basic	38	237	Basic						
New York	68	206	Below Basic	32	240	Proficient	11	267	Proficient			
North Carolina	77	201	Below Basic	59	218	Basic	47	228	Basic	7	275	Advanced
Oklahoma	85	188	Below Basic	70	205	Below Basic	4	279	Advanced			
Pennsylvania	87	185	Below Basic	59	219	Basic	22	255	Proficient			
Tennessee	88	172	Below Basic	45	224	Basic	14	261	Proficient			
Texas	74	194	Below Basic	21	247	Proficient						
Utah	69	208	Basic	42	233	Basic	13	267	Proficient			
Virginia	97	160	Below Basic	77	202	Below Basic	20	260	Proficient			
Wyoming	85	195	Below Basic	61	219	Basic	18	259	Proficient			

The results of NAEP benchmarks for ELA grade 8 individual states are reported in Table 16. The reading grade 8 NAEP achievement level cut-scores are Basic = 243, Proficient = 281, and Advanced = 323. The states with four achievement levels for which Level 3 maps to the NAEP Proficient level are Kansas and New York. The only state with five achievement levels for which Level 4 maps to the NAEP Proficient level is Florida.

Table 16: ELA Grade 8 NAEP Benchmarks for Nonconsortium States

State	ELA Grade 8											
	Level 2			Level 3			Level 4			Level 5		
	Percent at and Above Level 2	NAEP Equivalent for Level 2	NAEP Achievement Level Comparable to Level 2	Percent at and Above Level 3	NAEP Equivalent for Level 3	NAEP Achievement Level Comparable to Level 3	Percent at and Above Level 4	NAEP Equivalent for Level 4	NAEP Achievement Level Comparable to Level 4	Percent at and Above Level 5	NAEP Equivalent for Level 5	NAEP Achievement Level Comparable to Level 5
Alaska	80	228	Below Basic	31	279	Basic	2	337	Advanced			
Arizona	61	253	Basic	35	276	Basic	8	311	Proficient			
DoDEA	96	230	Below Basic	79	256	Basic	41	283	Proficient			
Florida	78	239	Below Basic	55	259	Basic	29	281	Proficient	11	303	Proficient
Georgia	76	238	Below Basic	39	272	Basic	8	312	Proficient			
Iowa	75	246	Basic	24	291	Proficient						
Kansas	78	241	Below Basic	29	285	Proficient	2	333	Advanced			
Kentucky	79	241	Below Basic	54	264	Basic	18	299	Proficient			
Minnesota	75	248	Basic	56	265	Basic	20	299	Proficient			
Nebraska	79	244	Basic	36	281	Basic						
New York	60	254	Basic	22	291	Proficient	7	317	Proficient			
North Carolina	79	231	Below Basic	53	257	Basic	42	269	Basic	10	309	Proficient
Oklahoma	87	226	Below Basic	75	241	Below Basic	16	295	Proficient			
Pennsylvania	89	225	Below Basic	58	262	Basic	15	306	Proficient			
Tennessee	91	221	Below Basic	50	265	Basic	11	306	Proficient			
Texas	78	234	Below Basic	23	286	Proficient						
Utah	66	256	Basic	42	276	Basic	15	304	Proficient			
Virginia	96	207	Below Basic	75	244	Basic	11	309	Proficient			
Wyoming	79	244	Basic	52	268	Basic	12	305	Proficient			

The results of NAEP benchmarks for mathematics grade 4 individual states are reported in Table 17. The mathematics grade 4 NAEP achievement level cut-scores are Basic = 214, Proficient = 249, and Advanced = 282. The only state with four achievement levels for which Level 3 maps to the NAEP Proficient level is Kansas. The only state with five achievement levels for which Level 4 maps to the NAEP Proficient level is Florida.

Table 17: Mathematics Grade 4 NAEP Benchmarks for Nonconsortium States

State	Mathematics Grade 4											
	Level 2			Level 3			Level 4			Level 5		
	Percent at and Above Level 2	NAEP Equivalent for Level 2	NAEP Achievement Level Comparable to Level 2	Percent at and Above Level 3	NAEP Equivalent for Level 3	NAEP Achievement Level Comparable to Level 3	Percent at and Above Level 4	NAEP Equivalent for Level 4	NAEP Achievement Level Comparable to Level 4	Percent at and Above Level 5	NAEP Equivalent for Level 5	NAEP Achievement Level Comparable to Level 5
Alaska	86	202	Below Basic	39	245	Basic	8	279	Proficient			
Arizona	72	220	Basic	42	244	Basic	10	276	Proficient			
DoDEA	88	218	Basic	66	237	Basic	39	255	Proficient			
Florida	77	222	Basic	59	236	Basic	31	256	Proficient	12	275	Proficient
Georgia	80	212	Below Basic	40	244	Basic	9	275	Proficient			
Iowa	79	219	Basic	29	260	Proficient						
Kansas	85	211	Below Basic	35	252	Proficient	8	282	Advanced			
Kentucky	80	219	Basic	49	243	Basic	16	270	Proficient			
Minnesota	85	217	Basic	70	233	Basic	36	261	Proficient			
Nebraska	77	223	Basic	24	263	Proficient						
New York	73	219	Basic	43	242	Basic	19	262	Proficient			
North Carolina	79	221	Basic	56	239	Basic	49	245	Basic	18	270	Proficient
Oklahoma	90	206	Below Basic	72	224	Basic	27	256	Proficient			
Pennsylvania	75	222	Basic	44	248	Basic	17	273	Proficient			
Tennessee	85	211	Below Basic	50	240	Basic	21	264	Proficient			
Texas	73	227	Basic	17	271	Proficient						
Utah	71	226	Basic	51	242	Basic	26	261	Proficient			
Virginia	97	193	Below Basic	84	218	Basic	28	263	Proficient			
Wyoming	88	214	Basic	51	246	Basic	13	278	Proficient			

The results of NAEP benchmarks for mathematics grade 8 individual states are reported in

Table 18. The mathematics grade 8 NAEP achievement level cut-scores are Basic = 262, Proficient = 299, and Advanced = 333. The only states with four achievement levels for which Level 3 maps to the NAEP Proficient level are Alaska, Kansas, New York, and Pennsylvania. The only state with five achievement levels for which Level 4 maps to the NAEP Proficient level is Florida.

In some states, some of the grade 8 students took the Algebra 1 test. In this benchmarking study, this factor could have had the effect of making the grade 8 mathematics standards appear higher.

Table 18: Mathematics Grade 8 NAEP Benchmarks for Nonconsortium States

State	Mathematics Grade 8											
	Level 2			Level 3			Level 4			Level 5		
	Percent at and Above Level 2	NAEP Equivalent for Level 2	NAEP Achievement Level Comparable to Level 2	Percent at and Above Level 3	NAEP Equivalent for Level 3	NAEP Achievement Level Comparable to Level 3	Percent at and Above Level 4	NAEP Equivalent for Level 4	NAEP Achievement Level Comparable to Level 4	Percent at and Above Level 5	NAEP Equivalent for Level 5	NAEP Achievement Level Comparable to Level 5
Alaska	89	236	Below Basic	26	304	Proficient	1	361	Advanced			
Arizona	59	275	Basic	34	298	Basic	14	323	Proficient			
DoDEA	96	236	Below Basic	78	267	Basic	45	295	Basic			
Florida	71	256	Below Basic	45	280	Basic	18	308	Proficient	7	328	Proficient
Georgia	75	254	Below Basic	37	291	Basic	12	321	Proficient			
Iowa	75	262	Below Basic	24	312	Proficient						
Kansas	62	274	Basic	22	309	Proficient	4	344	Advanced			
Kentucky	85	242	Below Basic	44	283	Basic	11	320	Proficient			
Minnesota	80	264	Basic	58	287	Basic	27	317	Proficient			
Nebraska	68	270	Basic	22	313	Proficient						
New York	60	271	Basic	22	308	Proficient	7	334	Advanced			
North Carolina	70	262	Below Basic	43	288	Basic	36	295	Basic	11	328	Proficient
Oklahoma	79	248	Below Basic	53	272	Basic	11	314	Proficient			
Pennsylvania	62	271	Basic	30	304	Proficient	8	338	Advanced			
Tennessee	81	246	Below Basic	54	274	Basic	29	299	Basic			
Texas	75	261	Below Basic	6	336	Advanced						
Utah	70	267	Basic	41	294	Basic	14	325	Proficient			
Virginia	93	235	Below Basic	74	265	Basic	9	336	Advanced			
Wyoming	84	255	Below Basic	47	289	Basic	10	327	Proficient			

Comparing Achievement Standards for Smarter Balanced, PARCC, and ACT Aspire

One of the advantages of mapping state achievement standards to NAEP is that the NAEP scale can serve as a common metric with which to compare the achievement standards of Smarter Balanced, PARCC, and ACT Aspire. The strategy is to obtain the NAEP equivalent of each consortium achievement standard and then compare their NAEP equivalents. The procedure used in this report is to compare their NAEP equivalents by using a two-tailed *Z* test with $p < .05$. The standard error used in the *Z* test is described in the Appendix.

The most important comparisons are between the college-ready standards of the group assessments. Comparing Smarter Balanced versus PARCC in Table 19 we find that

1. Smarter Balanced college-ready standards (Level 3) are comparable in difficulty to the NAEP Basic levels, and
2. Smarter Balanced college-ready standards (Level 3) are significantly below PARCC college-ready standards (Level 4) by about one-quarter of a standard deviation.

Table 19: Smarter Balanced Versus PARCC

		Smarter Balanced			PARCC			Difference	
Subject	Grade	NAEP Equivalent Level 3 Met	Standard Error NAEP Equivalent	NAEP Achievement Level	NAEP Equivalent Level 4 Met	Standard Error NAEP Equivalent	NAEP Achievement Level	Significant Difference ($p < .05$)	Effect Size Difference Smarter Minus PARCC
ELA	4	222	2	Basic	232	1	Basic	YES	−.26
ELA	8	264	1	Basic	273	1	Basic	YES	−.28
Math	4	244	1	Basic	252	1	Proficient	YES	−.26
Math	8	294	1	Basic	307	1	Proficient	YES	−.36

We can also compare the achievement standards of Smarter Balanced to those of ACT Aspire. When we compare the college-ready standards in Table 20, we find that

1. both Smarter Balanced and ACT Aspire college-ready standards (Ready) are comparable in difficulty to the NAEP Basic level; and
2. Smarter Balanced college-ready grade 8 standards are statistically comparable to ACT Aspire college-ready grade 8 standards. However, for grade 4, the Smarter Balanced college-ready standard is significantly below the ACT Aspire college-ready standard for ELA and reading (effect size = −.26) but significantly above the ACT Aspire college-ready standard for mathematics (effect size = +.29).

Table 20: Smarter Balanced Versus ACT Aspire

		Smarter Balanced			ACT Aspire			Difference	
Subject	Grade	NAEP Equivalent Level 3 Met	Standard Error NAEP Equivalent	NAEP Achievement Level	NAEP Equivalent Level 4 Met	Standard Error NAEP Equivalent	NAEP Achievement Level	Significant Difference ($p < .05$)	Effect Size Difference Smarter Minus ACT Aspire
ELA/ Reading	4	222	2	Basic	232	2	Basic	YES	−.26
ELA/ Reading	8	264	1	Basic	264	1	Basic	NO	
Math	4	244	1	Basic	235	1	Basic	YES	.29
Math	8	294	1	Basic	290	2	Basic	NO	

Similarly, we can compare PARCC and ACT Aspire college-ready standards. From Table 21, PARCC college-ready standards (Level 4) are statistically comparable in difficulty to the ACT Aspire college-ready standard for ELA and reading grade 4. However, PARCC standards are significantly above ACT Aspire college-ready standards for ELA and reading grade 8 (effect size = +.28), mathematics grade 4 (effect size = +.55), and mathematics grade 8 (effect size = +.48).

Table 21: PARCC Versus ACT Aspire

		PARCC			ACT Aspire			Difference	
Subject	Grade	NAEP Equivalent Level 4 Met	Standard Error NAEP Equivalent	NAEP Achievement Level	NAEP Equivalent of Ready Level	Standard Error NAEP Equivalent	NAEP Achievement Level	Significant Difference ($p < .05$)	Effect Size Difference PARCC Minus ACT Aspire
ELA/ Reading	4	232	1	Basic	232	2	Basic	NO	
ELA/ Reading	8	273	1	Basic	264	1	Basic	YES	.28
Math	4	252	1	Proficient	235	1	Basic	YES	.55
Math	8	307	1	Proficient	290	2	Basic	YES	.48

Conclusion

There are essentially three overall findings in this study.

1. A handful of nonconsortium states have college-ready standards that are at least as stringent as the NAEP Proficient level. These are

    a. ELA grade 4—Florida and New York;

    b. ELA grade 8—Florida, Kansas, and New York;

    c. Mathematics grade 4—Florida and Kansas; and

    d. Mathematics grade 8—Alaska, Florida, Kansas, New York, and Pennsylvania.

2. For the group-based assessments, only PARCC mathematics, grades 4 and 8, have college-ready standards comparable in difficulty to the NAEP Proficient level.
3. The Smarter Balanced achievement standards are about one-quarter of a standard deviation lower than the PARCC performance standards.

The benchmarking study reported here should give policy makers insight into what states are expecting from their students. Some states expect more, and some expect less. The study is intended to provide a way to benchmark and compare state achievement standards and benchmark and compare the achievement standards of Smarter Balanced, PARCC, and ACT Aspire. The study does not intend to evaluate state achievement standards or make policy recommendations.

Caveats

There are several caveats that are important to note in this study. First, the results in this report do not provide final and complete information about each state. The author was unable to obtain the results for several states, and some states have reported their results as preliminary. In the future, the National Center for Education Statistics (NCES) will conduct their biennial state mapping study. By that time, the NCES should be able to provide a more definitive and comprehensive mapping study.

Second, in some states, some of the grade 8 mathematics students took an end-of-course test, such as Algebra 1. In this benchmarking study, this factor could have had the effect of making the state grade 8 mathematics standards appear higher.

Third, this study maps state achievement standards to NAEP achievement levels and highlights those state standards that reach the NAEP Proficient level. This should not be interpreted to mean that NAEP's Proficient levels in grades 4 and 8 are the gold standards for deciding whether our students are on track to be ready for college. No evidence has been presented by NAEP that the proficient standard in grades 4 and 8 predicts college success. It is the case that NAEP used 12th grade college-ready cut-scores (2013) to report that about 38% of students have the reading skills, and 39% have the math skills that make them ready for college. The cut-scores were 302 for reading and 163 for mathematics. The reading college-ready cut-score was equal to the reading proficient standard, and the mathematics cut-score was just below the mathematics proficient standard.

Fourth, there are some interpretive nuances related to the methodology used in this study. This report uses statistical linking to map state achievement standards onto the NAEP scale. Holland (2007) has outlined three broad categories of linking. These are equating, scale alignment, and prediction. A fundamental difference among the three methods is related to the degree to which they assume the two tests measure the same content and have the same administrative procedures.

- In equating, both tests must be constructed to measure the same identical content, be equally reliable, and both tests must use the same administrative procedures.
- In scale alignment, both tests measure similar but not identical content, may not be equally reliable, and there can be variation in administrative procedures. Scale alignment can provide a good ballpark estimate of how scores line up, but is less precise than equating.
- In prediction, there are no assumptions at all about content, reliability or administrative procedures.

This report uses the second type: scale alignment. The scales we are aligning will not measure identical constructs[1], will not be equally reliable, and will not use identical administrative

[1] A recent study for mathematics by the NAEP Validity Study (NVS) panel found that 79% of NAEP items were matched to content in the CCSS in the 4th grade and 87% in the 8th grade (Daro, Hughes and Stancavage, 2015).

procedures. The method of alignment is equiprecentile linking based on the aggregate reporting of NAEP and the state assessments. It is the scales of the total aggregate distributions that are aligned, so the linking should not be used for disaggregated reporting of individual students or demographic subgroups (such as race/ethnicity or gender) or subpopulations (such as schools). Also, the reader should be aware that the concordance between NAEP and the state assessments established in this report for 2015 may not be applicable in subsequent years.

Fifth, this report does not, in any way, address or evaluate the quality of the CCSS. The CCSS are *content* standards, while this report deals only with *achievement* standards. Content standards represent the curriculum that teachers should teach, and the scope and sequence of what students should learn in school. Achievement standards are cut-scores on the state test that represent performance expectations. For example, what level of performance on the test do we think represents being on track to be college ready.

References

Daro, P., Hughes, G. B., and Stancavage, F. (2015). Study of the Alignment of the 2015 NAEP Mathematics Items at Grade 4 and 8 to the Common Core State Standards (CCSS) for Mathematics. (see http://www.air.org/project/naep-validity-studies-nvs-panel)

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*, New York, NY: Springer.

National Governors Association, Council of Chief State School Officers, Achieve. (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education*. Washington, DC: National Governors Association.

NAEP as an Indicator of Students' Academic Preparedness for College. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2013 Mathematics and Reading Assessments.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Obama, B. (2009, March 10). President Obama's remarks to the Hispanic Chamber of Commerce. *New York Times*.

Appendix: Methodology

This study uses equipercentile linking to benchmark state achievement standards against NAEP achievement levels. The derivations described below make two assumptions. First, we assume the state test scores and the NAEP test scores are normal distributions. Second, we assume the NAEP examinee sample is randomly equivalent to the population of examinees who took the state test.

NAEP scores are assumed to have a normal distribution $N(\hat{\mu}_N, \hat{\sigma}_N^2)$, where the standard error of $\hat{\mu}_N$ is estimated by $\hat{\sigma}_{\hat{\mu}_N}$, the standard error of $\hat{\sigma}_N$ is $\hat{\sigma}_{\hat{\sigma}_N}$, and the covariance between $\hat{\mu}_N$ and $\hat{\sigma}_N$ is $\hat{\sigma}_{\hat{\mu}_N,\hat{\sigma}_N}$, usually 0 if from a normal sample.

If the state-level proportion at and above the cut *c* is $\hat{p}_c$ with standard error of $\hat{\sigma}_{\hat{p}_c}$, the corresponding NAEP equivalent score, $\boldsymbol{\hat{s}_N}$ assuming random equivalent group tests, can be estimated by solving the equation

$$1 - \hat{p}_c = \int_{-\infty}^{\hat{s}_N} \frac{Exp\left(-\frac{(x-\hat{\mu}_N)^2}{2\hat{\sigma}_N^2}\right)}{\sqrt{2\pi}\hat{\sigma}_N} dx.$$

Let $y = \frac{x-\hat{\mu}_N}{\hat{\sigma}_N}$,and making the change of variable, we obtain

$$1 - \hat{p}_c = \int_{-\infty}^{\frac{\hat{s}_N-\hat{\mu}_N}{\hat{\sigma}_N}} \frac{Exp\left(-\frac{y^2}{2}\right)}{\sqrt{2\pi}} dy,$$

or

$$\frac{\hat{s}_N - \hat{\mu}_N}{\hat{\sigma}_N} = \Phi^{-1}(1 - \hat{p}_c).$$

So

$$\hat{s}_N = \hat{\mu}_N + \hat{\sigma}_N \Phi^{-1}(1 - \hat{p}_c).$$

Using delta method, the variance of the NAEP equivalent score $\hat{s}_N$ can be estimated by

$$\begin{aligned} Var(\hat{s}_N) = {} & Var(\hat{\mu}_N) + 2\Phi^{-1}(1-\hat{p}_c)Cov(\hat{\mu}_N, \hat{\sigma}_N) + Var(\hat{\sigma}_N)\left(\Phi^{-1}(1-\hat{p}_c)\right)^2 \\ & + \hat{\sigma}_N^2 Var\left(\Phi^{-1}(1-\hat{p}_c)\right), \end{aligned}$$

or

$$Var(\hat{s}_N) = \hat{\sigma}_{\hat{\mu}_N}^2 + 2\Phi^{-1}(1-\hat{p}_c)\hat{\sigma}_{\hat{\mu}_N,\hat{\sigma}_N} + \left(\Phi^{-1}(1-\hat{p}_c)\right)^2 \hat{\sigma}_{\hat{\sigma}_N}^2 + \hat{\sigma}_N^2 \left(\varphi\left(\Phi^{-1}(1-\hat{p}_c)\right)\right)^{-2} \hat{\sigma}_{\hat{p}_c}^2.$$

The standard error of the NAEP equivalent score $\hat{s}_N$ is then estimated by

$$\hat{\sigma}_{\hat{s}_N} = \sqrt{\hat{\sigma}^2_{\hat{\mu}_N} + 2\Phi^{-1}(1-\hat{p}_c)\hat{\sigma}_{\hat{\mu}_N,\hat{\sigma}_N} + \left(\Phi^{-1}(1-\hat{p}_c)\right)^2\hat{\sigma}^2_{\hat{\sigma}_N} + \hat{\sigma}^2_N\left(\varphi\left(\Phi^{-1}(1-\hat{p}_c)\right)\right)^{-2}\hat{\sigma}^2_{\hat{p}_c}}$$

where Φ is the cumulative distribution function of the standard normal distribution and φ is the probability density function of the standard normal distribution. Assuming $Cov(\hat{\mu}_N, \hat{\sigma}_N) = 0$, this is simplified to

$$\hat{\sigma}_{\hat{s}_N} = \sqrt{\hat{\sigma}^2_{\hat{\mu}_N} + \left(\Phi^{-1}(1-\hat{p}_c)\right)^2\hat{\sigma}^2_{\hat{\sigma}_N} + \hat{\sigma}^2_N\left(\varphi\left(\Phi^{-1}(1-\hat{p}_c)\right)\right)^{-2}\hat{\sigma}^2_{\hat{p}_c}}.$$

The values of $\hat{\sigma}_{\hat{s}_N}$were rounded up to the nearest NAEP scaled score unit.

For Smarter Balanced, PARCC, and ACT Aspire, the aggregate state-level proportion at and above the cut *c* is $\hat{p}_c$ with standard error $\hat{\sigma}_{\hat{p}_c}$ and was based on the weighted average of the states and jurisdictions within the consortium. The weights were the population sizes within each state. For Smarter Balanced, PARCC, and ACT Aspire, the state NAEP aggregate scores $\hat{s}_N$ were estimated with the NCES NDE (http://nces.ed.gov/nationsreportcard/naepdata/).

- Aggregate Smarter Balanced results are based on the weighted average of 18 states: California, Connecticut, Delaware, Hawaii, Idaho, Maine, Michigan, Missouri, Montana, Nevada, New Hampshire, North Dakota, Oregon, South Dakota, Vermont, Washington, West Virginia, and Wisconsin. For ELA, Missouri and Wisconsin were excluded because they did not follow the Smarter Balanced blueprint.
- Aggregate PARCC results are based on the weighted average of 12 jurisdictions: Arkansas, Colorado, District of Columbia, Illinois, Louisiana, Maryland, Massachusetts, Mississippi, New Jersey, New Mexico, Ohio, and Rhode Island. In grade 8 mathematics, in some PARCC states, some students took the Algebra 1 test. In the mapping study, this factor could have had the effect of making the grade 8 mathematics PARCC standards appear higher.
- Aggregate ACT Aspire results are based on the weighted average of two jurisdictions: Alabama and South Carolina.

AIR®

AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY: 877.334.3499

www.air.org

Making Research Relevant

2016 Regular Session

The Florida Senate

COMMITTEE MEETING EXPANDED AGENDA

EDUCATION PRE-K - 12
Senator Legg, Chair
Senator Detert, Vice Chair

MEETING DATE:	Thursday, September 17, 2015
TIME:	1:30—3:00 p.m.
PLACE:	*Pat Thomas Committee Room,* 412 Knott Building
MEMBERS:	Senator Legg, Chair; Senator Detert, Vice Chair; Senators Benacquisto, Brandes, Bullard, Clemens, Gaetz, Galvano, Garcia, Montford, and Sobel

TAB	BILL NO. and INTRODUCER	BILL DESCRIPTION and SENATE COMMITTEE ACTIONS	COMMITTEE ACTION
1	Status Update on the Implementation of Education Accountability Legislative Policy Requirements: CS/HB 7069 (2015) - Education Accountability		
	Other Related Meeting Documents		

Independent Verification of the Psychometric Validity for the Florida Standards Assessments

Andrew Wiley, Ph.D. Alpine Testing Solutions, Inc.

Ellen Forte, Ph.D. edCount LLC

Presentation prepared for the Committee on Education Pre-K – 12, The Florida Senate

September 17, 2015

Florida House Bill 7069 (April 2015)

- Mandated an "independent verification of the psychometric statewide, standardized assessments" be completed
- Created a 3-person panel responsible for selecting the organization
  - One appointed by the Governor of Florida,
  - One appointed by the President of the Florida Senate,
  - One appointed by the Speaker of the Florida House of Representatives
- Conduct a review of the development, production, administration, scoring and reporting of the grades 3-10 ELA, grades 3-8 Math, and Algebra 1, Algebra 2, and Geometry EOC assessments

Validity – Key Concepts

- Evaluation of uses, not the test or scores
- Based on available evidence
- Continuum (matter of degree) rather than a dichotomy (yes/no)

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.
>
> *Test Standards*, 2014, p.11

FSA Test Score Uses (Table 2, Page 27)

		Individual Student			Teacher	School			District	State
Content Area	Grade	Grade Promotion	Graduation Eligibility	Course Grade	Teacher Evaluation	School Grade	School Improvement Rating	Opportunity Scholarship	District Grade	State Accountability
English/ Language Arts	3	x			X	x	x	x	x	x
	4				X	x	x	x	x	x
	5				X	x	x	x	x	x
	6				X	x	x	x	x	x
	7				X	x	x	x	x	x
	8				X	x	x	x	x	x
	9				X	x	x	x	x	x
	10		X		X	x	x	x	x	x
Mathematics	3				x	x	x	x	x	x
	4				x	x	x	x	x	x
	5				x	x	x	x	x	x
	6				x	x	x	x	x	x
	7				x	x	x	x	x	x
	8				x	x	x	x	x	x
Algebra 1			X	X	x	x	x	x	x	x
Geometry				X	x	x	x	x	x	x
Algebra 2				X	x	x	x	x	x	x

Overview of Independent Verification

- Study 1 – Test Items
- Study 2 – Field Testing
- Study 3 – Test Blueprints & Construction
- Study 4 – Test Administration
- Study 5 – Scaling, Equating, and Scoring
- Study 6 – Psychometric Validity
- Final conclusions were reached using the data and information across all six studies

Evaluation Design

- FSA procedures compared to *Test Standards*
- Sources of Data
  - 700+ documents from FLDOE and vendors
  - In-person interviews with FLDOE and vendors
  - Item review with FL stakeholders
  - Survey of district assessment coordinators
  - Focus groups with district representatives
- For each study
  - Review of key activities
  - Findings
  - Commendations
  - Recommendations

Study 1 – Test Items

» Key Activities

- Review of More than 380 FSA items by Florida stakeholders for alignment to FL standards, fairness, depth of knowledge, etc.

» Findings

- Non-traditional item review and selection (completed by FLDOE rather than Florida educators) is an acceptable process under the development timeline
- Independent review indicated some differences in standards match and that all but 2 reviewed items matched one or more Florida Standards
- Independent review indicates need for closer attention to cognitive complexity in future item development

Study 2 – Field Testing

Key Activities

- Review of:
  - Utah student sample
  - Decision rules for accepting items for use in Florida
  - Post-administration item review procedures

Findings

- Field testing followed a somewhat unusual process (completed in Utah rather than Florida)
- Procedures adhered to industry practice

Study 3 – Test Blueprints & Construction

» Key Activities

- Review of:
  - Test Construction Specifications
  - Test Blueprints
  - Draft Score Reports

» Findings

- Procedures adhered to industry practice
- Test blueprints were generally consistent with the Florida standards but should reflect cognitive complexity at the item level
- Ongoing development of score reports and supplemental materials meant these could not be reviewed for this study

Study 4 – Test Administration

» Key Activities

- Collection of feedback from district representatives (nearly 70% of Florida districts participated)
- Discussions with FLDOE and test vendors about administration events
- Review of quantitative data provided by AIR

» Findings

- Computer-based testing faced challenges and issues
- There were differences in estimated impact from districts and test vendor
- Standardization of the test administration could be impacted
- Review of 2015 FSA test scores demonstrated consistency in performance across impacted and non impacted students

Study 5 – Scaling, Equating, & Scoring

» Key Activities

- Review of:
  - Scoring procedures
  - Calibration activities and analyses
  - Item statistics review process
  - Scaling specifications

» Findings

- Procedures adhered to industry practice
- Work related to these activities was ongoing at the time of this study

Study 6 – Psychometric Validity

Key Activities

- Review of:
  - FSA items (see Study 1)
  - FSA item statistics
  - Linking of FSA to FCAT 2.0 for Grade 10 ELA and Algebra 1

Findings

- Non-traditional interim standards (linking rather than standard setting)
- Procedures adhered to industry practice

Conclusions – Use of Scores

» Student-level

- "… test scores should not be used as a sole determinant in decisions such as the prevention of advancement to the next grade, graduation eligibility, or placement into a remedial course." p. 120

» Group-level

- "… the evidence appears to support the use of these data in the aggregate." p. 120
- "… cases may exist where a notably high percentage of students in a given classroom or school were impacted…" p. 121

Recommendations

(complete text for all recommendations is available in the final report)

Study 1 – Test Items

- **Recommendation 1.1:** FLDOE should phase out the Utah items as quickly as possible and use items on FSA assessments written to target the Florida standards.
- **Recommendation 1.2:** FLDOE should conduct an external alignment study on the entire pool of items appearing on the future FSA assessment with the majority of items targeting Florida standards to ensure documentation and range of complexity as intended for the FSA items across grades and content areas.
- **Recommendation 1.3:** FLDOE should conduct cognitive laboratories involving the capture and analysis of data about how students engage with test items and the content within each of the items during administration.

Study 2 – Field Testing

- **Recommendation 2.1:** FLDOE should provide further documentation and dissemination of the review and acceptance of Utah state items.

Recommendations

(complete text for all recommendations is available in the final report)

Study 3 Test Blueprints and Construction

- » **Recommendation 3.1** FLDOE should finalize and publish documentation related to test blueprint construction.
- » **Recommendation 3.2** FLDOE should include standard specific cognitive complexity expectations (DOK) in each grade-level content area blueprint.
- » **Recommendation 3.3** FLDOE should document the process through which the score reports and online reporting system for various stakeholders was developed, reviewed, and incorporated usability reviews, when appropriate.
- » **Recommendation 3.4** FLDOE should develop interpretation guides to accompany the score reports provided to stakeholders.

Recommendations

(complete text for all recommendations is available in the final report)

Study 4 Test Administration

- **Recommendation 4.1:** FLDOE and its vendors should be more proactive in the event of test administration issues.
- **Recommendation 4.2:** FLDOE and its FSA partners should engage with school districts in a communication and training program throughout the 2015-16 year.
- **Recommendation 4.3:** FLDOE should review and revise the policies and procedures developed for the FSA administration to allow for more efficient test delivery.

Study 5 Scaling, Equating, and Scoring

- **Recommendation 5.1:** Documentation of the computer-based scoring procedures should be provided in an accessible manner to stakeholders and test users.

Study 6 Psychometric Validity

- **Recommendation 6.1:** FLDOE should more clearly outline the limitations of the interim passing scores for the grade 10 ELA and Algebra 1 tests for stakeholders.

Additional Resources

House Bill 7069

https://www.flsenate.gov/Session/Bill/2015/7069

Florida Standards Assessment Review Selection Panel

(including complete copies of the final report and executive summary)

http://www.flgov.com/fl-standards-assessment-review-selection-panel/

Independent Verification of the Psychometric Validity for the Florida Standards Assessment

Executive Summary

August 31, 2015

Submitted to:
Vince Verges
Florida Department of Education
325 W. Gaines St.
Tallahassee FL 32399

Prepared by:
Andrew Wiley
Tracey R. Hembry
Chad W. Buckendahl
Alpine Testing Solutions, Inc.

and

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Table of Contents

Acknowledgments

This final report of the evaluation of the Florida Standards Assessment (FSA) benefited from the contributions of many people outside and within the Florida Department of Education (FLDOE). The evaluation team extends its appreciation to these individuals and acknowledges those whose assistance made this final report possible. More specifically, from FLDOE, we thank Vince Verges, Victoria Ash, Salih Binici, Susan Lee, Qian Liu, and Steve Ash. In addition, we thank Charlene Sozio from Volusia County Schools, Sally Shay and Oria O. Mcauliff from Miami-Dade County Public Schools, Cynthia G. Landers and Maxwell A. Barrington Jr. from Orange County Public Schools, and Gillian Gregory from Leon County Schools, who provided invaluable services as we conducted our survey and focus groups with Florida representatives.

We also thank the organizations and individuals that serve as vendors for the components of FSA that were included in the evaluation. These organizations were American Institutes for Research (AIR), Data Recognition Corporation (DRC), and Human Resources Research Organization (HumRRO).

Finally, because the foundation for this report is based on multiple studies and data collection efforts that comprised the evaluation, a number of people played key roles in the project. We appreciate the efforts of these individuals in contributing to the success of the evaluation. Specifically, we want to thank: Erica Brown, Brett Foley, and Jennifer Paine of Alpine Testing Solutions; and Alycia Hardy of edCount, LLC.

Although we have received feedback from FLDOE and vendors during the investigation, the judgments expressed in this report are those of the authors. This fulfills the spirit and intent that this evaluation be independent.

Andrew Wiley
Tracey R. Hembry
Chad W. Buckendahl
Alpine Testing Solutions, Inc.

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Executive Summary

Alpine Testing Solutions (Alpine) and edCount, LLC (edCount) were contracted to conduct an Independent Verification of the Psychometric Validity of the Florida Standards Assessments (FSA). Collectively, this evaluation team's charge was to conduct a review and analysis of the development, production, administration, scoring and reporting of the grades 3 through 10 English Language Arts (ELA), grades 3 through 8 Mathematics, and Algebra 1, Algebra 2, and Geometry End-of-Course assessments developed and administered in 2014-2015 by American Institutes for Research (AIR). To conduct the work, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; *Test Standards*), along with other seminal sources from the testing industry including *Educational Measurement*, 4th ed. (Brennan, 2006) and the *Handbook for Test Development* (Downing & Haladyna, 2006) were the guidelines to which all work was compared and served as the foundation of the evaluation.

As articulated in the Request for Offers, this investigation was organized into six separate studies; each study contributed to the overall evaluation of the FSA. These studies focused on evaluating several areas of evidence: 1) test items, 2) field testing, 3) test blueprint and construction, 4) test administration, 5) scaling, equating and scoring, and 6) specific questions of psychometric validity. For each of the six studies, the evaluation used a combination of document and data review, data collection with Florida educators, and discussions with staff from the Florida Department of Education (FLDOE) and its testing vendors. Although organized into separate studies, the synthesis of the results formed the basis for our findings, commendations, recommendations, and conclusions that emerged in this report.

This Executive Summary provides a high-level summary of the evaluation work including results of each of the six studies along with the overall findings and recommendations. In the body of the report, further detail for each of the six studies is provided, including the data and evidence collected, the interpretation of the evidence relative to the *Test Standards* and industry practice, findings, commendations, and recommendations. Following the discussion of the studies individually, we provide a synthesis of recommendations along with conclusions from the evaluation regarding the psychometric validity of the FSA scores for their intended uses.

Summary of the Evaluation Work

> "Evidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test" (Test Standards, 2014, p. 11).

The process of validation refers not to a test or scores but rather to the uses of test scores. By reviewing a collection of evidence gathered throughout the development and implementation of a testing program, an evaluation can provide an indication of the degree to which the available evidence supports each intended use of test scores. As such, the evaluation of the FSA program began with the identification of the uses and purposes of the tests. Per legislation and as outlined within FLDOE's *Assessment Investigation* (2015) document, FSA scores will contribute to decisions made regarding students, teachers, schools, districts, and the state. These uses across multiple levels of aggregation incorporate FSA data taken from a single year as well as measures of student growth from multiple years of data.

To consider the validity of each of these uses, the evaluation team worked with FLDOE and AIR to collect available documentation and information regarding each of the FSA program activities within the six studies. These materials were supplemented by regular communication via email and phone as well as interviews with relevant staff. Together, the evaluation team, FLDOE, and AIR worked together to identify key data points relevant to the evaluation. In addition, the evaluation team collected data related to the FSA items and the FSA administrations through meetings with Florida educators and a survey of district assessment coordinators.

This evidence was then compared to industry standards of best practice using sources like the *Test Standards* as well as other key psychometric texts. For each of the six studies, this comparison of evidence to standards provided the basis for the findings, recommendations, and commendations. These results were then evaluated together to reach overall conclusions regarding the validity evidence related to the use of FSA scores for decision-making at the levels of student, teacher, school, district, and state.

Evaluation of Test Items

This evaluation study is directly connected to the question of whether FSA follows procedures that are consistent with the *Test Standards* in the development of test items. This study included a review of test materials and included analyses of the specifications and fidelity of the development processes.

Findings

The review of FSA's practices allowed the evaluation team to explore many aspects of the FSA program. Except for the few noted areas of concern below, the methods and procedures used for the development and review of test items for the FSA were found to be in compliance with the *Test Standards* and with commonly accepted standards of practice.

Commendations

- Processes used to create and review test items are consistent with common approaches to assessment development.
- Methods for developing and reviewing the FSA items for content and bias were consistent with the *Test Standards* and followed sound measurement practices.

Recommendations:

Recommendation 1.1 Phase out items from the spring 2015 administration and use items written to specifically target Florida standards.

Every item that appears on the FSA was reviewed by Florida content and psychometric experts to determine content alignment with the Florida standards; however, the items were originally written to measure the Utah standards rather than the Florida standards. While alignment to Florida standards was confirmed for the majority of items reviewed via the item review study, many were not confirmed, usually because these items focused on slightly different content within the same anchor standards. It would be more appropriate to phase-out the items originally developed for use in Utah and replace them with items written to specifically target the Florida standards.

Recommendation 1.2 Conduct an independent alignment study

FLDOE should consider conducting an external alignment study on the entire pool of items appearing on future FSA assessments to ensure that items match standards. Additionally such a review could consider the complexity of individual items as well as the range of complexity across items and compare this information to the intended complexity levels by item as well as grade and content area. Further, the specifications for item writing relating to cognitive complexity should be revisited and items should be checked independently for depth of knowledge (DOK) prior to placement in the FSA item pool.

Recommendation 1.3 The FLDOE should conduct a series of cognitive labs

FLDOE should consider conducting cognitive laboratories, cognitive interviews, interaction studies involving the capture and analysis of data about how students engage with test items during administration, or other ways to gather response process evidence during the item development work over the next year.

Evaluation of Field Testing

Appropriate field testing of test content is a critical step for many testing programs to help ensure the overall quality of the assessment items and test forms. For this evaluation, the item development was started as part of the Utah Student Assessment of Student Growth and Excellence (SAGE) assessment program. Therefore, this study began with a review of the field testing practices that were followed for SAGE. The evaluation team also completed a review of the procedures that were followed once the SAGE assessments were licensed and the steps followed to identify items for the FSA.

Findings

For this study, the policies and procedures used in the field testing of test forms and items were evaluated and compared to the expectations of the *Test Standards* and industry best practices. While the FSA field testing was completed through a nontraditional method, the data collected and the review procedures that were implemented were consistent with industry-wide practices. The rationale and procedures used in the field testing provided appropriate data and information to support the development of the FSA test, including all components of the test construction, scoring, and reporting.

Commendations

- The field test statistics in Utah were collected from an operational test administration, thus avoiding questions about the motivation of test takers.
- During the Utah field testing process, the statistical performance of all items was reviewed to determine if the items were appropriate for use operationally.
- Prior to use of the FSA, all items were reviewed by educators knowledgeable of Florida students and the Florida Standards to evaluate whether the items were appropriate for use within the FSA program.
- After the FSA administration, all items went through the industry-expected statistical and content reviews to ensure accurate and appropriate items were delivered as part of the FSA.

Recommendations

Recommendation 2.1 Further documentation and dissemination on the review and acceptance of Utah state items.

The FLDOE should finalize and publish documentation that provides evidence that the FSA followed testing policies, procedures, and results that are consistent with industry expectations. While some of this documentation could be delayed due to operational program constraints that are still in process, other components could be documented earlier. Providing this information would be appropriate so that Florida constituents can be more fully informed about the status of the FSA.

Evaluation of Test Blueprints and Construction

This study evaluated evidence of test content and testing consequences related to the evaluation of the test blueprint and construction. This study focused on the following areas of review:

a) Review of the process for the test construction,
b) Review of the test blueprints to evaluate if the blueprints are sufficient for the intended purposes of the test,
c) Review of the utility of score reports for stakeholders by considering:
   i. Design of score reports for stakeholder groups
   ii. Explanatory text for appropriateness to the intended population
d) Information to support improvement of instruction

Findings

Given that the 2015 FSA was an adaptation of another state's assessments, much of the documentation about test development came from that other state. This documentation reflects an item development process that meets industry standards, although the documentation does not appear to be well represented in the body of technical documentation AIR offers. Likewise, the documentation of the original blueprint development process appears to have been adequate, but that information had to be pieced together with some diligence. The documentation about the process FLDOE undertook to adapt the blueprints and to select from the pool of available items reflects what would have been expected during a fast adaptation process.

The findings from the blueprint evaluation, when considered in combination with the item review results from Study 1, indicate that the blueprints that were evaluated (grades 3, 6, and 10 for English Language Arts, grades 4 and 7 for Math, and Algebra 1) do conform to the blueprint in terms of overall content match to the expected Florida standards. However, the lack of any cognitive complexity expectations in the blueprints mean that test forms could potentially include items that do not reflect the cognitive complexity in the standards and could vary in cognitive complexity across forms, thus allowing for variation across students, sites, and time.

In regards to test consequences and the corresponding review of score reporting materials, insufficient evidence was provided. The individual score reports must include scale scores and indicate performance in relation to performance standards. The performance level descriptors must be included in the report as must some means for communicating error. Currently, due to the timing of this study, this information is not included within the drafted FSA score reports.

Given the timing of this review, FLDOE and AIR have yet to develop interpretation guides for the score reports. These guides typically explicate a deeper understanding of score interpretation such as what content is assessed, what the scores represent, score precision, and intended uses of the scores.

Commendations

- FLDOE clearly worked intensely to establish an operational assessment in a very short timeline and worked on both content and psychometric concerns.

Recommendations

Recommendation 3.1 FLDOE should finalize and publish documentation related to test blueprint construction. Much of the current process documentation is fragmented among multiple data sources. Articulating a clear process linked to the intended uses of the FSA test scores provides information to support the validity of the intended uses of the scores.

Finalizing and publishing documentation related to test blueprint construction is highly recommended.

Recommendation 3.2 FLDOE should include standard specific cognitive complexity expectations (DOK) in each grade-level content area blueprint. While FLDOE provides percentage of points by depth of knowledge (DOK) level in the mathematics and ELA test design summary documents, this is insufficient to guide item writing and ensure a match between item DOK and expected DOK distributions.

Recommendation 3.3 FLDOE should document the process through which the score reports and online reporting system for various stakeholders was developed, reviewed, and incorporated usability reviews, when appropriate. Given the timing of this evaluation, the technical documentation outlining this development evidence for the FSA score reports was incomplete.

Recommendation 3.4 FLDOE should develop interpretation guides to accompany the score reports provided to stakeholders. The guides should include information that supports the appropriate interpretation of the scores for the intended uses, especially as it relates to the impact on instruction.

Evaluation of Test Administration

Prior to beginning the FSA evaluation, a number of issues related to the spring 2015 FSA administration were identified. These issues ranged from DDoS attacks, student login issues, and difficulty with the test administration process. The evaluation team gathered further information about all of these possible issues through reviews of internal documents from the FLDOE and AIR, data generated by the FLDOE and AIR, and focus groups and surveys with Florida district representatives.

Findings

The spring 2015 FSA administration was problematic. Problems were encountered on just about every aspect of the administration, from the initial training and preparation to the delivery of the tests themselves. Information from district administrators indicate serious systematic issues impacting a significant number of students, while statewide data estimates the impact to be closer to 1 to 5% for each test. The precise magnitude of the problems is difficult to gauge with 100% accuracy, but the evaluation team can reasonably state that the spring 2015 administration of the FSA did not meet the normal rigor and standardization expected with a high-stakes assessment program like the FSA.

Commendations

- Throughout all of the work of the evaluation team, one of the consistent themes amongst people the team spoke with and the surveys was the high praise for the FLDOE staff members who handled the day-to-day activities of the FSA. Many individuals took the time to praise their work and to point out that these FLDOE staff members went above and beyond their normal expectations to assist them in any way possible.

Recommendations

Recommendation 4.1 FLDOE and its vendors should be more proactive in the event of test administration issues.

Standard 6.3 from the *Test Standards* emphasizes the need for comprehensive documentation and reporting anytime there is a deviation from standard administration procedures. It would be appropriate for the FLDOE and its vendors to create contingency plans that more quickly react to any administration-related issues with steps designed to help ensure the reliability, validity, and fairness of the FSAs.

Recommendation 4.2 FLDOE and its FSA partners should engage with school districts in a communication and training program throughout the entire 2015-16 academic year.

The problematic spring 2015 FSA administration has made many individuals involved with the administration of the FSA to be extremely skeptical of its value. Given this problem, the FLDOE and its partners should engage in an extensive communication and training program throughout the entire academic year to inform its constituents of the changes that have been made to help ensure a less troublesome administration in 2016.

Recommendation 4.3 The policies and procedures developed for the FSA administration should be reviewed and revised to allow the test administrators to more efficiently deliver the test, and when required, more efficiently resolve any test administration issues.

Test administration for all FSAs should be reviewed to determine ways to better communicate policies to all test users. The process for handling any test administration issues during the live test administration must also be improved. Improved Help desk support should be one essential component.

Evaluation of Scaling, Equating, and Scoring

This study evaluated the processes for scaling, calibrating, equating, and scoring the FSA. The evaluation team reviewed the rationale and selection of psychometric methods and procedures that are used to analyze data from the FSA. It also included a review of the proposed methodology for the creation of the FSA vertical scale.

Findings

Based on the documentation and results available, acceptable procedures were followed and sufficient critical review of results was implemented. In addition, FLDOE and AIR solicited input from industry experts on various technical aspects of the FSA program through meetings with the FLDOE's Technical Advisory Committee (TAC).

Commendations

- Although AIR committed to the development of the FSA program within a relatively short timeframe, the planning, analyses, and data review related to the scoring and calibrations of the FSA (i.e., the work that has been completed to date) did not appear to be negatively impacted by the time limitations. The procedures outlined for these activities followed industry standards and were not reduced to fit within compressed schedules.

Recommendation

Recommendation 5.1 - Documentation of the computer-based scoring procedures, like those used for some of the FSA technology-enhanced items as well as that used for the essays, should be provided in an accessible manner to stakeholders and test users.

AIR uses computer-based scoring technology (i.e., like that used for the FSA technology-enhanced items and essays). Therefore, for other programs in other states, the documentation around these scoring procedures should already exist and be available for review (e.g., scoring algorithms for FSA technology-enhanced items was embedded within patent documents).

Specific Psychometric Validity Questions

This study evaluated specific components of psychometric validity that in some instances aligned with other studies in the broader evaluation. The evaluation team considered multiple sources of evidence, including judgmental and empirical characteristics of the test and test items, along with the psychometric models used. This study also included a review of the methodology compiled for linking the FSA tests to the FCAT 2.0.

Findings

During the scoring process, the statistical performance of all FSA items were evaluated to determine how well each item fit the scoring model chosen for the FSA and that the items fit within acceptable statistical performance. In regards to the linking of scores for grade 10 ELA and Algebra 1, FLDOE and AIR implemented a solution that served the purpose and requirement determined by the state. While some concerns about the requirements for linking the FSA to the FCAT were raised, the methodology used was appropriate given the parameters of the work required.

Commendations

- Given an imperfect psychometric situation regarding the original source of items and the reporting requirements, AIR and FLDOE appear to have carefully found a balance that delivered acceptable solutions based on the FSA program constraints.

Recommendation

Recommendation 6.1 The limitations of the interim passing scores for the grade 10 ELA and Algebra 1 tests should be more clearly outlined for stakeholders.

Unlike the passing scores used on FCAT 2.0 and those that will be used for subsequent FSA administrations, the interim passing scores were not established through a formal standard setting process and therefore do not represent a criterion-based measure of student knowledge and skills. The limitations regarding the meaning of these interim passing scores should be communicated to stakeholders.

Conclusions

As the evaluation team has gathered information and data about the Florida Standards Assessments (FSA), we note a number of commendations and recommendations that have been provided within the description of each of the six studies. The commendations note areas of strength while recommendations represent opportunities for improvement and are primarily focused on process improvements, rather than conclusions related to the test score validation question that was the primary motivation for this project.

As was described earlier in the report, the concept of validity is explicitly connected to the intended use and interpretation of the test scores. As a result, it is not feasible to arrive at a simple Yes/No decision when it comes to the question "Is the test score valid?" Instead, the multiple uses of the FSA must be considered, and the question of validity must be considered separately for each. Another important consideration in the evaluation of validity is that the concept is viewed most appropriately as a matter of degree rather than as a dichotomy. As evidence supporting the intended use accumulates, the degree of confidence in the validity of a given test score use can increase or decrease. For purposes of this evaluation, we provide specific conclusions for each study based on the requested evaluative judgments and then frame our overarching conclusions based on the intended uses of scores from the FSA.

Study-Specific Conclusions

The following provide conclusions from each of the six studies that make up this evaluation.

Conclusion #1 – Evaluation of Test Items

When looking at the item development and review processes that were followed with the FSA, **the policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the test items were determined to be error free, unbiased, and were written to support research-based instructional methodology, use student- and grade-appropriate language as well as content standards-based vocabulary, and assess the applicable content standard.

Conclusion #2 – Evaluation of Field Testing

Following a review of the field testing rationale, procedure, and results for the FSA, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the field testing design, process, procedures, and results support an assertion that the sample size was sufficient and that the item-level data were adequate to support test construction, scoring, and reporting for the purposes of these assessments.

Conclusion #3 – Evaluation of Test Blueprint and Construction

When looking at the process for the development of test blueprints, and the construction of FSA test forms, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards*.** The initial documentation of the item development reflects a process that meets industry standards, though the documentation could be enhanced and placed into a more coherent framework. Findings also observed that the blueprints that were evaluated do reflect the Florida Standards in terms of overall content match, evaluation of intended complexity as compared to existing complexity was not possible due to a lack of specific complexity information in the blueprint. Information for testing consequences, score reporting, and interpretive guides were not included in this study as the score reports with scale scores and achievement level descriptors along with the accompanying interpretive guides were not available at this time.

Conclusion #4 – Evaluation of Test Administration

Following a review of the test administration policies, procedures, instructions, implementation, and results for the FSA, **with some notable exceptions, the intended policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, some aspects of the test administration, such as the test delivery engine, and the instructions provided to administrators and students, were consistent with other comparable programs. However, for a variety of reasons, the spring 2015 FSA test administration was problematic, with issues encountered on multiple aspects of the computer-based test (CBT) administration. These issues led to significant challenges in the administration of the FSA for some students, and as a result, these students were not presented with an opportunity to adequately represent their knowledge and skills on a given test.

Conclusion #5 – Evaluation of Scaling, Equating, and Scoring

Following a review of the scaling, equating, and scoring procedures and methods for the FSA, and **based on the evidence available at the time of this evaluation, the policies, procedures, and methods are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the measurement model used or planned to be used, as well as the rationale for the models was considered to be appropriate, as are the equating and scaling activities associated with the FSA. Note that evidence related to content validity is included in the first and third conclusions above and not repeated here. There are some notable exceptions to the breadth of our conclusion for this study. Specifically, evidence was not available at the time of this study to be able to evaluate evidence of criterion, construct, and consequential validity. These are areas where more comprehensive studies have yet to be completed. Classification accuracy and consistency were not available as part of this review because achievement standards have not yet been set for the FSA.

Conclusion #6 – Evaluation of Specific Psychometric Validity Questions

Following a review of evidence for specific psychometric validity questions for the FSA, **the policies, methods, procedures, and results that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry with notable exceptions**. Evidence related to a review of the FSA items and their content are noted in the first conclusion above and not repeated here. The difficulty levels and discrimination levels of items were appropriate and analyses were conducted to investigate potential sources of bias. The review also found that the psychometric procedures for linking the FSA Algebra 1 and Grade 10 ELA with the associated FCAT 2.0 tests were acceptable given the constraints on the program.

Cross-Study Conclusions

Because validity is evaluated in the context of the intended uses and interpretations of scores, the results of any individual study are insufficient to support overall conclusions. The following conclusions are based on the evidence compiled and reviewed across studies in reference to the intended uses of the FSAs both for individual students and for aggregate-level information.

Conclusion #7 – Use of FSA Scores for Student-Level Decisions

With respect to student level decisions, **the evidence for the paper and pencil delivered exams support the use of the FSA at the student level. For the CBT FSA, the FSA scores for some students will be suspect. Although the percentage of students in the aggregate may appear small, it still represents a significant number of students for whom critical decisions need to be made. Therefore, test scores should not be used as a sole determinant in decisions such as the prevention of advancement to the next grade, graduation eligibility, or placement into a remedial course**. However, under a "hold harmless" philosophy, if students were able to complete their tests(s) and demonstrate performance that is considered appropriate for an outcome that is beneficial to the student (i.e., grade promotion, graduation eligibility), it would appear to be appropriate that these test scores could be used in combination with other sources of evidence about the student's ability. This conclusion is primarily based on observations of the difficulties involved with the administration of the FSA.

Conclusion #8 – Use of Florida Standards Assessments Scores for Group-Level Decisions

In reviewing the collection of validity evidence from across these six studies in the context of group level decisions (i.e., teacher, school, district or state) that are intended uses of FSA scores, **the evidence appears to support the use of these data in the aggregate**. **This conclusion is appropriate for both the PP and the CBT examinations.** While the use of FSA scores for individual student decisions should only be interpreted in ways that would result in student outcomes such as promotion, graduation, and placement, the use of FSA test scores at an aggregate level does appear to still be warranted. Given that the percentage of students

with documented administration difficulties remained low when combining data across students, schools and districts, it is likely that aggregate level use would be appropriate.

The primary reason that aggregate level scores are likely appropriate for use is the large number of student records involved. As sample sizes increase and approach a census level, and we consider the use of FSA at the district or state level, the impact of a small number of students whose scores were influenced by administration issues should not cause the mean score to increase or decrease significantly. However, cases may exist where a notably high percentage of students in a given classroom or school were impacted by any of these test administration issues. It would be advisable for any user of aggregated test scores strongly consider this possibility, continue to evaluate the validity of the level of impact, and implement appropriate policies to consider this potential differential impact across different levels of aggregation.

Independent Verification of the Psychometric Validity for the Florida Standards Assessment

Final Report

August 31, 2015

Submitted to:
Vince Verges
Florida Department of Education
325 W. Gaines St.
Tallahassee FL 32399

Prepared by:
Andrew Wiley
Tracey R. Hembry
Chad W. Buckendahl
Alpine Testing Solutions, Inc.

and

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Table of Contents

Acknowledgments

This final report of the evaluation of the Florida Standards Assessment (FSA) benefited from the contributions of many people outside and within the Florida Department of Education (FLDOE). The evaluation team extends its appreciation to these individuals and acknowledges those whose assistance made this final report possible. More specifically, from FLDOE, we thank Vince Verges, Victoria Ash, Salih Binici, Susan Lee, Qian Liu, and Steve Ash. In addition, we thank Charlene Sozio from Volusia County Schools, Sally Shay and Oria O. Mcauliff from Miami-Dade County Public Schools, Cynthia G. Landers and Maxwell A. Barrington Jr. from Orange County Public Schools, and Gillian Gregory from Leon County Schools, who provided invaluable services as we conducted our survey and focus groups with Florida representatives.

We also thank the organizations and individuals that serve as vendors for the components of FSA that were included in the evaluation. These organizations were American Institutes for Research (AIR), Data Recognition Corporation (DRC), and Human Resources Research Organization (HumRRO).

Finally, because the foundation for this report is based on multiple studies and data collection efforts that comprised the evaluation, a number of people played key roles in the project. We appreciate the efforts of these individuals in contributing to the success of the evaluation. Specifically, we want to thank: Erica Brown, Brett Foley, and Jennifer Paine of Alpine Testing Solutions; and Alycia Hardy of edCount, LLC.

Although we have received feedback from FLDOE and vendors during the investigation, the judgments expressed in this report are those of the authors. This fulfills the spirit and intent that this evaluation be independent.

Andrew Wiley
Tracey R. Hembry
Chad W. Buckendahl
Alpine Testing Solutions, Inc.

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Executive Summary

Alpine Testing Solutions (Alpine) and edCount, LLC (edCount) were contracted to conduct an Independent Verification of the Psychometric Validity of the Florida Standards Assessments (FSA). Collectively, this evaluation team's charge was to conduct a review and analysis of the development, production, administration, scoring and reporting of the grades 3 through 10 English Language Arts (ELA), grades 3 through 8 Mathematics, and Algebra 1, Algebra 2, and Geometry End-of-Course assessments developed and administered in 2014-2015 by American Institutes for Research (AIR). To conduct the work, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; *Test Standards*), along with other seminal sources from the testing industry including *Educational Measurement*, 4th ed. (Brennan, 2006) and the *Handbook for Test Development* (Downing & Haladyna, 2006) were the guidelines to which all work was compared and served as the foundation of the evaluation.

As articulated in the Request for Offers, this investigation was organized into six separate studies; each study contributed to the overall evaluation of the FSA. These studies focused on evaluating several areas of evidence: 1) test items, 2) field testing, 3) test blueprint and construction, 4) test administration, 5) scaling, equating and scoring, and 6) specific questions of psychometric validity. For each of the six studies, the evaluation used a combination of document and data review, data collection with Florida educators, and discussions with staff from the Florida Department of Education (FLDOE) and its testing vendors. Although organized into separate studies, the synthesis of the results formed the basis for our findings, commendations, recommendations, and conclusions that emerged in this report.

This Executive Summary provides a high-level summary of the evaluation work including results of each of the six studies along with the overall findings and recommendations. In the body of the report, further detail for each of the six studies is provided, including the data and evidence collected, the interpretation of the evidence relative to the *Test Standards* and industry practice, findings, commendations, and recommendations. Following the discussion of the studies individually, we provide a synthesis of recommendations along with conclusions from the evaluation regarding the psychometric validity of the FSA scores for their intended uses.

Summary of the Evaluation Work

"Evidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test" (Test Standards, 2014, p. 11).

The process of validation refers not to a test or scores but rather to the uses of test scores. By reviewing a collection of evidence gathered throughout the development and implementation of a testing program, an evaluation can provide an indication of the degree to which the available evidence supports each intended use of test scores. As such, the evaluation of the FSA program began with the identification of the uses and purposes of the tests. Per legislation and as outlined within FLDOE's *Assessment Investigation* (2015) document, FSA scores will contribute to decisions made regarding students, teachers, schools, districts, and the state. These uses across multiple levels of aggregation incorporate FSA data taken from a single year as well as measures of student growth from multiple years of data.

To consider the validity of each of these uses, the evaluation team worked with FLDOE and AIR to collect available documentation and information regarding each of the FSA program activities within the six studies. These materials were supplemented by regular communication via email and phone as well as interviews with relevant staff. Together, the evaluation team, FLDOE, and AIR worked together to identify key data points relevant to the evaluation. In addition, the evaluation team collected data related to the FSA items and the FSA administrations through meetings with Florida educators and a survey of district assessment coordinators.

This evidence was then compared to industry standards of best practice using sources like the *Test Standards* as well as other key psychometric texts. For each of the six studies, this comparison of evidence to standards provided the basis for the findings, recommendations, and commendations. These results were then evaluated together to reach overall conclusions regarding the validity evidence related to the use of FSA scores for decision-making at the levels of student, teacher, school, district, and state.

Evaluation of Test Items

This evaluation study is directly connected to the question of whether FSA follows procedures that are consistent with the *Test Standards* in the development of test items. This study included a review of test materials and included analyses of the specifications and fidelity of the development processes.

Findings

The review of FSA's practices allowed the evaluation team to explore many aspects of the FSA program. Except for the few noted areas of concern below, the methods and procedures used for the development and review of test items for the FSA were found to be in compliance with the *Test Standards* and with commonly accepted standards of practice*.*

Commendations

- Processes used to create and review test items are consistent with common approaches to assessment development.
- Methods for developing and reviewing the FSA items for content and bias were consistent with the *Test Standards* and followed sound measurement practices.

Recommendations:

Recommendation 1.1 Phase out items from the spring 2015 administration and use items written to specifically target Florida standards.

Every item that appears on the FSA was reviewed by Florida content and psychometric experts to determine content alignment with the Florida standards; however, the items were originally written to measure the Utah standards rather than the Florida standards. While alignment to Florida standards was confirmed for the majority of items reviewed via the item review study, many were not confirmed, usually because these items focused on slightly different content within the same anchor standards. It would be more appropriate to phase-out the items originally developed for use in Utah and replace them with items written to specifically target the Florida standards.

Recommendation 1.2 Conduct an independent alignment study

FLDOE should consider conducting an external alignment study on the entire pool of items appearing on future FSA assessments to ensure that items match standards. Additionally such a review could consider the complexity of individual items as well as the range of complexity across items and compare this information to the intended complexity levels by item as well as grade and content area. Further, the specifications for item writing relating to cognitive complexity should be revisited and items should be checked independently for depth of knowledge (DOK) prior to placement in the FSA item pool.

Recommendation 1.3 The FLDOE should conduct a series of cognitive labs

FLDOE should consider conducting cognitive laboratories, cognitive interviews, interaction studies involving the capture and analysis of data about how students engage with test items during administration, or other ways to gather response process evidence during the item development work over the next year.

Evaluation of Field Testing

Appropriate field testing of test content is a critical step for many testing programs to help ensure the overall quality of the assessment items and test forms. For this evaluation, the item development was started as part of the Utah Student Assessment of Student Growth and Excellence (SAGE) assessment program. Therefore, this study began with a review of the field testing practices that were followed for SAGE. The evaluation team also completed a review of the procedures that were followed once the SAGE assessments were licensed and the steps followed to identify items for the FSA.

Findings

For this study, the policies and procedures used in the field testing of test forms and items were evaluated and compared to the expectations of the *Test Standards* and industry best practices. While the FSA field testing was completed through a nontraditional method, the data collected and the review procedures that were implemented were consistent with industry-wide practices. The rationale and procedures used in the field testing provided appropriate data and information to support the development of the FSA test, including all components of the test construction, scoring, and reporting.

Commendations

- The field test statistics in Utah were collected from an operational test administration, thus avoiding questions about the motivation of test takers.
- During the Utah field testing process, the statistical performance of all items was reviewed to determine if the items were appropriate for use operationally.
- Prior to use of the FSA, all items were reviewed by educators knowledgeable of Florida students and the Florida Standards to evaluate whether the items were appropriate for use within the FSA program.
- After the FSA administration, all items went through the industry-expected statistical and content reviews to ensure accurate and appropriate items were delivered as part of the FSA.

Recommendations

Recommendation 2.1 Further documentation and dissemination on the review and acceptance of Utah state items.

The FLDOE should finalize and publish documentation that provides evidence that the FSA followed testing policies, procedures, and results that are consistent with industry expectations. While some of this documentation could be delayed due to operational program constraints that are still in process, other components could be documented earlier. Providing this information would be appropriate so that Florida constituents can be more fully informed about the status of the FSA.

Evaluation of Test Blueprints and Construction

This study evaluated evidence of test content and testing consequences related to the evaluation of the test blueprint and construction. This study focused on the following areas of review:

a) Review of the process for the test construction,
b) Review of the test blueprints to evaluate if the blueprints are sufficient for the intended purposes of the test,
c) Review of the utility of score reports for stakeholders by considering:
   i. Design of score reports for stakeholder groups
   ii. Explanatory text for appropriateness to the intended population
d) Information to support improvement of instruction

Findings

Given that the 2015 FSA was an adaptation of another state's assessments, much of the documentation about test development came from that other state. This documentation reflects an item development process that meets industry standards, although the documentation does not appear to be well represented in the body of technical documentation AIR offers. Likewise, the documentation of the original blueprint development process appears to have been adequate, but that information had to be pieced together with some diligence. The documentation about the process FLDOE undertook to adapt the blueprints and to select from the pool of available items reflects what would have been expected during a fast adaptation process.

The findings from the blueprint evaluation, when considered in combination with the item review results from Study 1, indicate that the blueprints that were evaluated (grades 3, 6, and 10 for English Language Arts, grades 4 and 7 for Math, and Algebra 1) do conform to the blueprint in terms of overall content match to the expected Florida standards. However, the lack of any cognitive complexity expectations in the blueprints mean that test forms could potentially include items that do not reflect the cognitive complexity in the standards and could vary in cognitive complexity across forms, thus allowing for variation across students, sites, and time.

In regards to test consequences and the corresponding review of score reporting materials, insufficient evidence was provided. The individual score reports must include scale scores and indicate performance in relation to performance standards. The performance level descriptors must be included in the report as must some means for communicating error. Currently, due to the timing of this study, this information is not included within the drafted FSA score reports.

Given the timing of this review, FLDOE and AIR have yet to develop interpretation guides for the score reports. These guides typically explicate a deeper understanding of score interpretation such as what content is assessed, what the scores represent, score precision, and intended uses of the scores.

Commendations

- FLDOE clearly worked intensely to establish an operational assessment in a very short timeline and worked on both content and psychometric concerns.

Recommendations

Recommendation 3.1 FLDOE should finalize and publish documentation related to test blueprint construction. Much of the current process documentation is fragmented among multiple data sources. Articulating a clear process linked to the intended uses of the FSA test scores provides information to support the validity of the intended uses of the scores.

Finalizing and publishing documentation related to test blueprint construction is highly recommended.

Recommendation 3.2 FLDOE should include standard specific cognitive complexity expectations (DOK) in each grade-level content area blueprint. While FLDOE provides percentage of points by depth of knowledge (DOK) level in the mathematics and ELA test design summary documents, this is insufficient to guide item writing and ensure a match between item DOK and expected DOK distributions.

Recommendation 3.3 FLDOE should document the process through which the score reports and online reporting system for various stakeholders was developed, reviewed, and incorporated usability reviews, when appropriate. Given the timing of this evaluation, the technical documentation outlining this development evidence for the FSA score reports was incomplete.

Recommendation 3.4 FLDOE should develop interpretation guides to accompany the score reports provided to stakeholders. The guides should include information that supports the appropriate interpretation of the scores for the intended uses, especially as it relates to the impact on instruction.

Evaluation of Test Administration

Prior to beginning the FSA evaluation, a number of issues related to the spring 2015 FSA administration were identified. These issues ranged from DDoS attacks, student login issues, and difficulty with the test administration process. The evaluation team gathered further information about all of these possible issues through reviews of internal documents from the FLDOE and AIR, data generated by the FLDOE and AIR, and focus groups and surveys with Florida district representatives.

Findings

The spring 2015 FSA administration was problematic. Problems were encountered on just about every aspect of the administration, from the initial training and preparation to the delivery of the tests themselves. Information from district administrators indicate serious systematic issues impacting a significant number of students, while statewide data estimates the impact to be closer to 1 to 5% for each test. The precise magnitude of the problems is difficult to gauge with 100% accuracy, but the evaluation team can reasonably state that the spring 2015 administration of the FSA did not meet the normal rigor and standardization expected with a high-stakes assessment program like the FSA.

Commendations

- Throughout all of the work of the evaluation team, one of the consistent themes amongst people the team spoke with and the surveys was the high praise for the FLDOE staff members who handled the day-to-day activities of the FSA. Many individuals took the time to praise their work and to point out that these FLDOE staff members went above and beyond their normal expectations to assist them in any way possible.

Recommendations

Recommendation 4.1 FLDOE and its vendors should be more proactive in the event of test administration issues.

Standard 6.3 from the *Test Standards* emphasizes the need for comprehensive documentation and reporting anytime there is a deviation from standard administration procedures. It would be appropriate for the FLDOE and its vendors to create contingency plans that more quickly react to any administration-related issues with steps designed to help ensure the reliability, validity, and fairness of the FSAs.

Recommendation 4.2 FLDOE and its FSA partners should engage with school districts in a communication and training program throughout the entire 2015-16 academic year.

The problematic spring 2015 FSA administration has made many individuals involved with the administration of the FSA to be extremely skeptical of its value. Given this problem, the FLDOE and its partners should engage in an extensive communication and training program throughout the entire academic year to inform its constituents of the changes that have been made to help ensure a less troublesome administration in 2016.

Recommendation 4.3 The policies and procedures developed for the FSA administration should be reviewed and revised to allow the test administrators to more efficiently deliver the test, and when required, more efficiently resolve any test administration issues.

Test administration for all FSAs should be reviewed to determine ways to better communicate policies to all test users. The process for handling any test administration issues during the live test administration must also be improved. Improved Help desk support should be one essential component.

Evaluation of Scaling, Equating, and Scoring

This study evaluated the processes for scaling, calibrating, equating, and scoring the FSA. The evaluation team reviewed the rationale and selection of psychometric methods and procedures that are used to analyze data from the FSA. It also included a review of the proposed methodology for the creation of the FSA vertical scale.

Findings

Based on the documentation and results available, acceptable procedures were followed and sufficient critical review of results was implemented. In addition, FLDOE and AIR solicited input from industry experts on various technical aspects of the FSA program through meetings with the FLDOE's Technical Advisory Committee (TAC).

Commendations

- Although AIR committed to the development of the FSA program within a relatively short timeframe, the planning, analyses, and data review related to the scoring and calibrations of the FSA (i.e., the work that has been completed to date) did not appear to be negatively impacted by the time limitations. The procedures outlined for these activities followed industry standards and were not reduced to fit within compressed schedules.

Recommendation

Recommendation 5.1 - Documentation of the computer-based scoring procedures, like those used for some of the FSA technology-enhanced items as well as that used for the essays, should be provided in an accessible manner to stakeholders and test users.

AIR uses computer-based scoring technology (i.e., like that used for the FSA technology-enhanced items and essays). Therefore, for other programs in other states, the documentation around these scoring procedures should already exist and be available for review (e.g., scoring algorithms for FSA technology-enhanced items was embedded within patent documents).

Specific Psychometric Validity Questions

This study evaluated specific components of psychometric validity that in some instances aligned with other studies in the broader evaluation. The evaluation team considered multiple sources of evidence, including judgmental and empirical characteristics of the test and test items, along with the psychometric models used. This study also included a review of the methodology compiled for linking the FSA tests to the FCAT 2.0.

Findings

During the scoring process, the statistical performance of all FSA items were evaluated to determine how well each item fit the scoring model chosen for the FSA and that the items fit within acceptable statistical performance. In regards to the linking of scores for grade 10 ELA and Algebra 1, FLDOE and AIR implemented a solution that served the purpose and requirement determined by the state. While some concerns about the requirements for linking the FSA to the FCAT were raised, the methodology used was appropriate given the parameters of the work required.

Commendations

- Given an imperfect psychometric situation regarding the original source of items and the reporting requirements, AIR and FLDOE appear to have carefully found a balance that delivered acceptable solutions based on the FSA program constraints.

Recommendation

Recommendation 6.1 The limitations of the interim passing scores for the grade 10 ELA and Algebra 1 tests should be more clearly outlined for stakeholders.

Unlike the passing scores used on FCAT 2.0 and those that will be used for subsequent FSA administrations, the interim passing scores were not established through a formal standard setting process and therefore do not represent a criterion-based measure of student knowledge and skills. The limitations regarding the meaning of these interim passing scores should be communicated to stakeholders.

Conclusions

As the evaluation team has gathered information and data about the Florida Standards Assessments (FSA), we note a number of commendations and recommendations that have been provided within the description of each of the six studies. The commendations note areas of strength while recommendations represent opportunities for improvement and are primarily focused on process improvements, rather than conclusions related to the test score validation question that was the primary motivation for this project.

As was described earlier in the report, the concept of validity is explicitly connected to the intended use and interpretation of the test scores. As a result, it is not feasible to arrive at a simple Yes/No decision when it comes to the question "Is the test score valid?" Instead, the multiple uses of the FSA must be considered, and the question of validity must be considered separately for each. Another important consideration in the evaluation of validity is that the concept is viewed most appropriately as a matter of degree rather than as a dichotomy. As evidence supporting the intended use accumulates, the degree of confidence in the validity of a given test score use can increase or decrease. For purposes of this evaluation, we provide specific conclusions for each study based on the requested evaluative judgments and then frame our overarching conclusions based on the intended uses of scores from the FSA.

Study-Specific Conclusions

The following provide conclusions from each of the six studies that make up this evaluation.

Conclusion #1 – Evaluation of Test Items

When looking at the item development and review processes that were followed with the FSA, **the policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the test items were determined to be error free, unbiased, and were written to support research-based instructional methodology, use student- and grade-appropriate language as well as content standards-based vocabulary, and assess the applicable content standard.

Conclusion #2 – Evaluation of Field Testing

Following a review of the field testing rationale, procedure, and results for the FSA, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the field testing design, process, procedures, and results support an assertion that the sample size was sufficient and that the item-level data were adequate to support test construction, scoring, and reporting for the purposes of these assessments.

Conclusion #3 – Evaluation of Test Blueprint and Construction

When looking at the process for the development of test blueprints, and the construction of FSA test forms, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards*.** The initial documentation of the item development reflects a process that meets industry standards, though the documentation could be enhanced and placed into a more coherent framework. Findings also observed that the blueprints that were evaluated do reflect the Florida Standards in terms of overall content match, evaluation of intended complexity as compared to existing complexity was not possible due to a lack of specific complexity information in the blueprint. Information for testing consequences, score reporting, and interpretive guides were not included in this study as the score reports with scale scores and achievement level descriptors along with the accompanying interpretive guides were not available at this time.

Conclusion #4 – Evaluation of Test Administration

Following a review of the test administration policies, procedures, instructions, implementation, and results for the FSA, **with some notable exceptions, the intended policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, some aspects of the test administration, such as the test delivery engine, and the instructions provided to administrators and students, were consistent with other comparable programs. However, for a variety of reasons, the spring 2015 FSA test administration was problematic, with issues encountered on multiple aspects of the computer-based test (CBT) administration. These issues led to significant challenges in the administration of the FSA for some students, and as a result, these students were not presented with an opportunity to adequately represent their knowledge and skills on a given test.

Conclusion #5 – Evaluation of Scaling, Equating, and Scoring

Following a review of the scaling, equating, and scoring procedures and methods for the FSA, and **based on the evidence available at the time of this evaluation, the policies, procedures, and methods are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the measurement model used or planned to be used, as well as the rationale for the models was considered to be appropriate, as are the equating and scaling activities associated with the FSA. Note that evidence related to content validity is included in the first and third conclusions above and not repeated here. There are some notable exceptions to the breadth of our conclusion for this study. Specifically, evidence was not available at the time of this study to be able to evaluate evidence of criterion, construct, and consequential validity. These are areas where more comprehensive studies have yet to be completed. Classification accuracy and consistency were not available as part of this review because achievement standards have not yet been set for the FSA.

Conclusion #6 – Evaluation of Specific Psychometric Validity Questions

Following a review of evidence for specific psychometric validity questions for the FSA, **the policies, methods, procedures, and results that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry with notable exceptions**. Evidence related to a review of the FSA items and their content are noted in the first conclusion above and not repeated here. The difficulty levels and discrimination levels of items were appropriate and analyses were conducted to investigate potential sources of bias. The review also found that the psychometric procedures for linking the FSA Algebra 1 and Grade 10 ELA with the associated FCAT 2.0 tests were acceptable given the constraints on the program.

Cross-Study Conclusions

Because validity is evaluated in the context of the intended uses and interpretations of scores, the results of any individual study are insufficient to support overall conclusions. The following conclusions are based on the evidence compiled and reviewed across studies in reference to the intended uses of the FSAs both for individual students and for aggregate-level information.

Conclusion #7 – Use of FSA Scores for Student-Level Decisions

With respect to student level decisions, **the evidence for the paper and pencil delivered exams support the use of the FSA at the student level. For the CBT FSA, the FSA scores for some students will be suspect. Although the percentage of students in the aggregate may appear small, it still represents a significant number of students for whom critical decisions need to be made. Therefore, test scores should not be used as a sole determinant in decisions such as the prevention of advancement to the next grade, graduation eligibility, or placement into a remedial course**. However, under a “hold harmless” philosophy, if students were able to complete their tests(s) and demonstrate performance that is considered appropriate for an outcome that is beneficial to the student (i.e., grade promotion, graduation eligibility), it would appear to be appropriate that these test scores could be used in combination with other sources of evidence about the student’s ability. This conclusion is primarily based on observations of the difficulties involved with the administration of the FSA.

Conclusion #8 – Use of Florida Standards Assessments Scores for Group-Level Decisions

In reviewing the collection of validity evidence from across these six studies in the context of group level decisions (i.e., teacher, school, district or state) that are intended uses of FSA scores, **the evidence appears to support the use of these data in the aggregate**. **This conclusion is appropriate for both the PP and the CBT examinations.** While the use of FSA scores for individual student decisions should only be interpreted in ways that would result in student outcomes such as promotion, graduation, and placement, the use of FSA test scores at an aggregate level does appear to still be warranted. Given that the percentage of students

with documented administration difficulties remained low when combining data across students, schools and districts, it is likely that aggregate level use would be appropriate.

The primary reason that aggregate level scores are likely appropriate for use is the large number of student records involved. As sample sizes increase and approach a census level, and we consider the use of FSA at the district or state level, the impact of a small number of students whose scores were influenced by administration issues should not cause the mean score to increase or decrease significantly. However, cases may exist where a notably high percentage of students in a given classroom or school were impacted by any of these test administration issues. It would be advisable for any user of aggregated test scores strongly consider this possibility, continue to evaluate the validity of the level of impact, and implement appropriate policies to consider this potential differential impact across different levels of aggregation.

Florida Standards Assessment Background

At the beginning of 2013, the state of Florida was a contributing member to the *Partnership for Assessment of Readiness for College and Careers* (PARCC) consortia. However, in August of 2014, Governor Rick Scott convened a group of the state's leading educators who completed a review of the Common Core State Standards and its application to Florida schools. Shortly after this summit, Governor Scott announced that that Florida would remove itself from the PARCC consortia and pursue an assessment program focused solely on Florida standards.

In February of 2014, changes to the Florida Standards were approved by the Florida State Board of Education. These new standards were designed to encourage a broader approach to student learning and to encourage deeper and more analytic thinking on the part of students.

In March of 2014, Florida began a contract with the American Institutes for Research (AIR) for the development of the Florida Standards Assessments (FSA) program. AIR was selected through a competitive bidding process that began in October of 2013 with the release of an Invitation to Negotiate by the Florida Department of Education (FLDOE).

The FSA program consists of grades 3-10 English Language Arts (ELA; grade 11 ELA was originally included as well), grades 3-8 Math, and end-of-course (EOC) tests for Algebra 1, Geometry, and Algebra 2. The ELA assessments consist of Reading and Writing assessments which are administered separately but combined for scoring and reporting, except for Grade 3 which only includes Reading. The FSA program consists of a combination of both paper-and-pencil (PP) and computer-based tests (CBT) depending on the grade level and the content area. Additionally accommodated versions of the tests were also prepared for students with disabilities (SWD).

In April of 2014, it was announced that the items that would comprise the 2014-15 FSA would be licensed from the state of Utah's Student Assessment of Growth and Excellence (SAGE) program. All items would be field tested with Utah students as part of their 2014 operational test administration. The process of reviewing and approving the items began immediately, and culminated later in 2014 with the creation of the first FSA test forms.

Throughout the 2014-15 academic year, FLDOE in collaboration with AIR and Data Recognition Corporation (DRC), the vendor responsible for the scoring of FSA Writing responses as well as the materials creation, distribution and processing for the PP tests, provided training materials to Florida schools and teachers. These materials were provided through a combination of materials on the FLDOE website, webinars, and in-person workshops.

The administration of the FSA tests began on March 2, 2015 with the Writing tests and concluded on May 15, 2015 with the EOCs.

Legislative Mandate

Florida House Bill 7069, passed in April 2015, mandated an independent evaluation of the FSA program and created a panel responsible for selecting the organization for which Florida would partner for the work. The panel is comprised of three members: one appointed by the Governor of Florida, one appointed by the President of the Florida Senate, and the third appointed by the Speaker of the Florida House of Representatives. The charge for this project was to conduct a review of the development, production, administration, scoring and reporting of the grades 3-10 ELA, grades 3-8 Math, and Algebra 1, Algebra 2, and Geometry EOC assessments.

Florida Standards Assessment Timeline

Table 1 outlines the major milestones that led up to or were part of the development of the FSA assessments, including those related to the legislative mandate the outlined the current evaluation work.

Table 1. Timeline of Florida Standards Assessment-Related Activities.

December 2010	Florida is announced as one of 13 states acting as governing states for the Partnership for Assessment Readiness for College and Careers (PARCC) consortium.
September 2013	Using input from the summit, Governor Scott issued Executive Order 13-276, which (among other requirements): • Tasked the Commissioner of Education to recommend to the State Board of Education the establishment of an open process to procure Florida's next assessment by issuing a competitive solicitation; • Initiated Florida's departure from the national PARCC consortium as its fiscal agent, to ensure that the state would be able to procure a test specifically designed for Florida's needs without federal intervention.

Date	Action
February 2014	State Board of Education approved changes to the standards that reflected the input from public comments about the standards, which resulted from public hearings around the state and thousands of comments from Floridians.
March 2014	An evaluation team reviewed five proposals and narrowed the choice to three groups. Subsequently, a negotiation team unanimously recommended the not-for-profit American Institutes for Research (AIR).
May 2014	Commissioner of Education releases the 2014-2015 Statewide Assessment Schedule
June 3, 2014	AIR Contract executed
December 1-19, 2014 and January 5-February 13, 2014	Grades 4-11 CBT Writing Component Field test
February 24, 2015	Governor Rick Scott signs Executive Order 15-31 to suspend the Grade 11 Florida Standards Assessment for English Language Arts
March 2, 2015	Operational FSA Testing begins with grades 8-10 Writing
April 14, 2015	House Bill 7069 is signed by Governor Rick Scott. It creates a panel to select an independent entity to conduct a verification of the psychometric validity of the Florida Standards Assessments.
May 15, 2015	Operational FSA testing concludes
May 15, 2015	Request for Offers for the Independent Verification of the Psychometric Validity for the Florida Standards Assessment is issued
May 18, 2015	FLDOE announces that districts are to calculate final course grades and make promotion decisions for Algebra 1, Algebra 2, and Geometry without regard to the 30% requirement for the FSAs.
May 29, 2015	Alpine Testing Solutions and edCount LLC are selected to perform independent validation study
June 5, 2015	Alpine Testing Solutions contract executed
August 31, 2015	Alpine and edCount deliver final report to FLDOE

Evaluation Design

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (Test Standards, 2014, p. 11)

As requested for the project, our approach to the independent investigation of the FSA was framed by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; *Test Standards*). For assessment programs, the *Test Standards* require that test sponsors develop not only an explicit definition of the intended uses and interpretations of the test scores, but also a comprehensive collection of evidence to support these inferences and interpretations. "It is not the test that is validated, and it is not the test scores that are validated. It is the claims and decisions based on the test results that are validated" (Kane, 2006, pp. 59-60). For assessment programs like FSA, validity evidence that links the assessment development and program activities to the intended uses of the scores is critical.

Validity is evaluated by considering each of the intended uses of test scores separately along with the evidence that has been collected throughout the lifespan of a program in support of such test uses. "The test developer is expected to make a case for the validity of the intended uses and interpretations" (Kane, 2006, p. 17). As such, the role of this investigation is to consider the validity evidence available in support of each use of the FSA test scores, as outlined by FLDOE, and to compare this evidence to that required by the *Test Standards* and other significant works within the field of psychometrics. Based on this comparison of available FSA-related evidence to that prescribed by industry standards, the evaluation team provides recommendations, commendations, and conclusions about the validity of the intended uses of the 2014-15 FSA test scores.

It is important to emphasize that validity is a matter of degree and is not an inherent property of a test. Validity is evaluated in the context of the intended interpretations and uses of the test scores and the capacity of the evidence to support the respective interpretation.

Intended Uses of the Florida Standards Assessments

Developing or evaluating an assessment program begins with an explicit determination of the intended interpretations and uses of the resultant scores. For this evaluation, the intended uses and interpretations of FSA scores serve as the context for integrating the sources of evidence from the evaluation to then form recommendations, commendations, and conclusions. To lay the groundwork for readers to better understand and interpret the findings that are reported in the remaining sections of the report, we provide an overview of the intended uses of the FSA scores as well the source for the associated mandates for each use.

"Standard 1.2: A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation." (Test Standards, 2014, p. 23)

The process of evaluating an assessment and its associated validity evidence is directly related to the intended uses of the scores. Validity refers to these specific uses rather than a global determination of validity for an assessment program. As such, it is possible that the validity evidence supports one specific use of scores from an assessment while is insufficient for another.

Like many state assessment programs, FSA includes a number of intended uses of scores with varying stakes for individuals or groups. The FSA is intended to be used to make decisions related to students. In addition, student-level results, both for the current year as well as for progress across years, are then to be aggregated to make decisions related to teachers, schools, districts, and the state.

More information related to the details of these uses at varying levels, as well as the associated state statutes that outline and mandate these uses can be found in FLDOE's *Assessment Investigation February 2015* document which can be accessed at http://www.fldoe.org/core/fileparse.php/12003/urlt/CommAssessmentInvestigationReport.pdf

Table 2 provides a summary of these intended uses of the FSA and notes the uses for which modifications have been made for 2014-15 as the first year of the program.

Table 2. Intended Uses of the Florida Standards Assessments (FSA) Scores

Content Area	Grade	Grade Promotion	Graduation Eligibility	Course Grade	Teacher Evaluation	School Grade	School Improvement Rating	Opportunity Scholarship	District Grade	State Accountability
English/ Language Arts	3	✓			✓	✓	✓	✓	✓	✓
	4				✓	✓	✓	✓	✓	✓
	5				✓	✓	✓	✓	✓	✓
	6				✓	✓	✓	✓	✓	✓
	7				✓	✓	✓	✓	✓	✓
	8				✓	✓	✓	✓	✓	✓
	9				✓	✓	✓	✓	✓	✓
	10		✓		✓	✓	✓	✓	✓	✓
Mathematics	3				✓	✓	✓	✓	✓	✓
	4				✓	✓	✓	✓	✓	✓
	5				✓	✓	✓	✓	✓	✓
	6				✓	✓	✓	✓	✓	✓
	7				✓	✓	✓	✓	✓	✓
	8				✓	✓	✓	✓	✓	✓
Algebra 1			✓	✓	✓	✓	✓	✓	✓	✓
Geometry				✓	✓	✓	✓	✓	✓	✓
Algebra 2				✓	✓	✓	✓	✓	✓	✓

Studies within the Evaluation

In accordance with the Request for Offers, the investigation of the psychometric validity of the FSA has been organized to include six separate studies. These studies include an evaluation of 1) test items, 2) field testing, 3) test blueprint and construction, 4) test administration, 5) scaling, equating, and scoring, and 6) specific questions of psychometric validity. Table 3 outlines the framework for these studies as they relate to the various sources of validity evidence cited within the *Test Standards.*

While these studies are presented separately within this report, the combination of the evidence gathered from each study provides the basis of the evaluation of the uses of the FSA. Determinations of sufficient validity evidence cannot be based on single studies. Rather, each study captures a significant group of activities that were essential to the development and delivering of the FSA program, and therefore ample validity evidence from each individual study can be viewed as necessary but not sufficient to reach a final determination of adequate validity evidence related to specific score uses.

Table 3. Validation Framework for Independent Verification of Psychometric Validity of Florida Standards Assessments

	AERA et al. (2014) Source of Validity Evidence				
Evaluation Target Areas	**Test Content**	**Response Processes**	**Internal Structure**	**Relations to other Variables**	**Testing Consequences**
Evaluation of Test Items	Review test development and review processes Review sample of assessment items for content and potential bias	Review student and grade level language; cognitive levels			
Evaluation of Field Testing			Review rationale, execution, and results of sampling		Review whether results support test construction
Evaluation of Test Blueprint and Construction	Review test blueprint for sufficiency to support intended purposes				Review the utility of score reports for stakeholders to improve instruction
Evaluation of Test Administration		Review of test accommodations		Review of delivery system utility and user experience Review of third-party technology and security audit reports	Review of test administration procedures Review of security protocols for prevention, investigation, and enforcement
	Review evidence of content validity produced by the program	Review evidence of content validity produced by the program	Review choice of model, scoring, analyses, equating, and scaling.	Review evidence of construct validity collected by the program	Review evidence of testing consequences

	AERA et al. (2014) Source of Validity Evidence				
Evaluation Target Areas	Test Content	Response Processes	Internal Structure	Relations to other Variables	Testing Consequences
			Subgroup psychometric characteristics Subscore added value analyses, decision consistency, and measurement precision	Review criterion evidence collected by the program	produced by the program
Specific Evaluation of Psychometric Validity	Review a sample of items relative to course descriptions and for freedom from bias	Review of a sample of items for intended response behavior as opposed to guessing	Review of item difficulty, discrimination, potential bias Review the linking processes for Algebra 1 and Grade 10 ELA relative to 2013-14 results.		

Evaluation Procedure

The majority of the work focused on reviewing evidence produced by FLDOE and the FSA vendor partners. This focus of the evaluation is consistent with the expectations of the *Test Standards* that indicate

> Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of any test score interpretations for specified uses intended by the developer. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. (2014, p. 13)

To supplement the document, policy, and material review, the evaluation team also collected additional information through interviews with key personnel during in-person meetings. This two stage approach to testing program evaluation is more fully described in Buckendahl and Plake (2006).

The evaluation team also collected supplemental evidence for the evaluation directly from Florida educators. This evidence included information regarding the alignment of the FSA to Florida academic content standards. It also included surveys and focus groups with Florida district representatives regarding the spring 2015 FSA test administrations.

In addition, the evaluation team worked with the FLDOE and with AIR to identify key data points that could be used to evaluate the magnitude and impact of the test administration issues from spring FSA administration. This included data summarizing the test administration behavior of students as well as analyses to look further at impact on student performance. All analyses completed were reviewed by the FLDOE and by the evaluation team.

Together, information collected from the testing vendors and FLDOE, both through documentation and interviews, as well as the data collected during the alignment meeting, online survey, and focus group meetings provided a great deal of information related to the development of and processes used within the FSA program.

Limitations of the Evaluation

Several factors limited the comprehensiveness of the evaluation design and its implementation. Given the size of the FSA program and the number of intended uses for its scores, our greatest limitation was a constraint regarding time to collect and review evidence. The findings, recommendations, and conclusions of this evaluation are limited by the availability of information during the evaluation. Similar to an organization conducting a financial audit, the quality of the documentation and supporting evidence influences an independent auditor’s judgment. The concept is analogous for assessment programs.

A primary source for evidence of development and validation activities for assessment programs is the documentation provided in a program’s technical manual and supporting technical reports. A technical manual will generally document the qualifications of the individuals engaged in the process, processes and procedures that were implemented, results of these processes, and actions taken in response to those results.

Because the FSA were administered in the spring of 2015, some of the development and validation activities are ongoing and a comprehensive technical manual was not yet available. Nonetheless, the evaluation team was able to access technical reports, policy documents, and other process documents, along with interviews with key staff, student data files, and vendor produced analyses, to inform the evaluation. Instances where collection of evidence was in progress or not available are noted in the respective study. A list of the documents and materials reviewed for the project is included as Appendix B.

Study 1: Evaluation of Test Items

Study Description

The design and implementation of this study focused on how the assessments were developed along with a review of FSA test items. The evaluation team reviewed the documentation of the development processes using criteria based on best practices in the testing industry. In addition, the team conducted in-person and virtual interviews with FLDOE and partner vendor staff to gather information not included in documentation or to clarify evidence. The study was planned to include the following:

- Test development and review processes including:
  - The characteristics and qualifications of subject matter experts used throughout the process
  - The review processes that were implemented during the development process along with quality control processes
  - The decision rules that were implemented throughout the item development and review process
  - The consistency of the results with expected outcomes of the processes and with any changes that were recommended during the review processes

- A review of a minimum of 200 operational assessment items across grades and content areas. The review was led subject matter experts and included a sample of Florida teachers. The item review evaluated test items for the following characteristics:
  - Structured consistently with best practices in assessment item design
  - Consistent with widely accepted, research-based instructional methods
  - Appropriate cognitive levels to target intended depth of knowledge (DOK)
  - Review for potential bias related to sex, race, ethnicity, socioeconomic status
  - Appropriate student and grade-level language
  - Targeting the intended content standard(s)

Sources of Evidence

The following documents served as the primary sources of evidence for this study:

- Utah State Assessment Technical Report: Volume 2 Test Development
- Test Development Staff Resumes (UT item development)
- SAGE Item Development Process Draft
- Writing and Reviewing Effective Items PowerPoint (UT item development)
- Bias and Sensitivity Review Training PowerPoint (UT item development)
- Item Writing Specifications
- Fall 2014 Bias and Sensitivity Review Summary Comments (per grade/content area)
- Content Committee and Bias and Sensitivity Report for SAGE

- SAGE Parent Review Committee Report
- FSA Test Construction Specifications

In addition to document and process review, the evaluation of test items also included additional reviews and data collection by the evaluation team. First, data related to item content and DOK match were collected July 20-21, 2015 in Tampa, Florida. During this period, the evaluation team conducted item reviews with Florida stakeholders from the Test Development Center (TDC), as well as classroom teachers and content coaches/instructional specialists at the district level to gather information directly from Florida stakeholders about the items on the FSA. Panelists (n=23) were selected via a list of names provided by FLDOE as individuals recommended by the TDC with Mathematics or ELA content experience. The panelists served on panels to review one form for each of ELA grades 3, 6, and 10 and Math grades 4, 7, and Algebra 1. The grades were selected purposefully to represent 1) one grade in each of the grade bands, 2) both paper-and-pencil (PP) and online administrations of the FSA, and 3) an end of course assessment. For the purpose of this study, all the items on the forms were reviewed, including field test items. The item review study focused on 1) the content match between the intended Florida standard for each item and the Florida standard provided by panelists and 2) the match between the DOK rating provided by FLDOE for each of the items and the DOK rating provided by panelists for that grade-level/content area. Panelists were not told what the intended content or DOK ratings were for any of the items they reviewed.

Data from this study were analyzed in two ways: 1) computation of the percentage of exact match between panelists' ratings and intended ratings, and 2) computation of the difference between the average target DOK and the average rater DOK indices. The difference between the average target and rated DOK indices of less than or equal to .5 would be considered strong DOK consistency, a difference of less than 1 point but more than .5 points would be considered moderate, and a difference of 1 point or greater would represent weak evidence of DOK consistency.

Next, content/test development experts reviewed the same items for bias, sensitivity, and fairness considerations. Then, special education experts reviewed the items on these forms for accessibility considerations, especially in relation to students with visual and hearing impairments and students with mild-moderate disabilities. Finally, experts reviewed the items for purposeful item development to reduce the likelihood of guessing. Results from these studies/reviews provided additional evidence to evaluate the test content. Results from all studies and reviews are included within the interpretation section that follows. Confidential reports with item specific information for consideration will be delivered to FLDOE separately for item security purposes.

Study Limitations

The program documentation and activities permitted the completion of this study as intended and originally designed.

Industry Standards

A firm grounding in the *Test Standards* is necessary to the credibility of each study in this evaluation. With specific regard to Study 1, the following standards are most salient and were drivers in the study design and implementation.

Important validity evidence related to test content is often obtained from "an analysis of the relationship between the content of a test and the construct it is intended to measure" (*Test Standards*, p. 15). In regard to evidence based on test content, the *Test Standards* (1.1) first direct a clear specification of the construct(s) that the test is intended to assess. The *Test Standards* (4.12) also recommend that test developers "document the extent to which the content domain of a test represents the domain defined in the test specifications" (p. 89). Most often, test developers document the extent of this content representation by providing information about the design process in combination with an independent/external study of the alignment between the test questions and the content standards. Such documentation should address multiple criteria regarding how well the test aligns with the standards the test is meant to measure in terms of the range and complexity of knowledge and skills students are expected to demonstrate on the test.

In regard to evidence based on test content, the Test Standards (1.1) first direct a clear specification of the construct(s) that the test is intended to assess.

As evidence that a test is fair and free from bias, the *Test Standards* (4.0/3.9) recommend that test developers and publishers 1) "document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population" (p. 85) and 2) "are responsible for developing and providing accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs" (p. 67). These studies often include bias, sensitivity, and accessibility reviews with panelists who have expertise in issues related to students with disabilities, students who are English learners, as well as panelists who can provide sensitivity considerations for race, ethnicity, culture, gender, and socio-economic status.

The *Test Standards* recommend (1.12) "if the rationale for score interpretation for a given use depends on premises about the … cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided." Evidence related to response processes should be documented through consideration of student performance and characteristics 1) during item development (e.g., through a principled development process/approach), 2) during test administration and gathered from the digital platform, or 3) through cognitive laboratories or interviews during item development, administration, or post hoc.

Florida Standards Assessments Processes and Evaluation Activities

For the review of evidence of test content and response processes related to the evaluation of test items developed for the spring 2015 FSA Assessment, AIR and FLDOE provided substantial documentation. The evaluation team also gathered documentation via item reviews with Florida stakeholders and content/test design/and special education experts. Reviews and interpretation of the evidence in each of these areas is outlined below.

Test Content

Evidence of test content begins with a clear description of the construct(s) that the test is intended to measure and the extent to which the content domain of a test represents the domain defined in the test specifications.

Evidence of test content begins with a clear description of the construct(s) that the test is intended to measure and the extent to which the content domain of a test represents the domain defined in the test specifications.

The prioritization of content and explication of the content intended to be measured by the FSA was well documented by AIR and FLDOE. Experts engaged in the item development had the content expertise as would be expected of item writers and developers. Item development and review practices as well as the documentation of these practices met industry standards and followed the *Test Standards* guidelines. However, due to the limited time frame for developing the FSA, item reviews related to content, cognitive complexity, bias/sensitivity, etc. were not conducted by Florida stakeholders. Florida content and psychometric experts from FLDOE reviewed every item appearing on the FSA, but other Florida stakeholders were not involved.

As an external check on alignment of test items with the Florida Standards, the evaluation team conducted item reviews with Florida stakeholders recommended by the Test Development Center (TDC). Panelists were: 1) split into groups by grade-level/content expertise, 2) asked to complete a background questionnaire to describe the expertise and experience of the panelists, 3) trained on completing the Florida Standards match and rating DOK, 4) given an opportunity to conduct practice ratings using the Florida Standards to ground them in the standards and calibrate the ratings of DOK between panelists, 5) provided a panel facilitator to answer questions, monitor ratings between panelists to ensure high inter-rater agreement, and monitor security of materials, and 6) asked to rate the Florida Standards match and DOK of each of the items for that grade-level/content area (individually first, then asked to determine consensus ratings as a panel).

A total of 23 panelists were selected from a list of names provided by FLDOE as individuals recommended by the TDC with Math or ELA content experience. All panels included four participants except ELA grade 10 which had only three. About 70% of the panelists were females and 30% were males. Most panelists were white (67%), 25% were African-American, and Hispanic and Native American panelists each represented 4% of the panel make-up. The highest level of education represented was at the Masters level (80% of panelists). Almost 80% of the participants had more than 10 years of experience, with half of those having more than 20 years of experience. More than 90% of educators had experience conducting and leading professional development and all had experience in curriculum planning for the content area panel on which they served.

Florida Standards Comparisons

After panelists' ratings had been collected, researchers compared the intended Florida Standards designated to be assessed by each item with the Florida Standards ratings provided by content experts on each panel. The outcomes of the content match analyses are presented in Table 4.[1]

Table 4. Item Content Match with Intended Florida Standards

Content Area/Grade	Standard Match	Partial Standard Match	No Standard Match
ELA Grade 3	65%	2%	33%
ELA Grade 6	76%	6%	17%
ELA Grade 10	65%	15%	20%
ELA Total	69%	8%	23%
Math Grade 4	94%	0%	6%
Math Grade 7	79%	0%	21%
Algebra 1	81%	0%	19%
Math Total	84%	0%	16%

Note: Some percentages do not equal 100% due to rounding.

English Language Arts Grade 3. Panelists reviewed a form of the grade 3 ELA test consisting of 60 items. The grade 3 ELA panelists' ratings matched the intended standards for the majority of items (65%). The single item that was rated as a partial match encompassed two parts; panelists matched the intended standard on the first part and added a standard for the second part, resulting in the partial alignment rating. Panelists selected a different standard than the intended standard for 33% of the items.

English Language Arts Grade 6. Panelists reviewed a form of the grade 6 ELA test consisting of 63 items. The grade six ELA panelists selected standards that agreed with the intended standards on the majority of items (76%). The panelists matched the intended standard on

[1] Specific information about item content cannot be provided in evaluation reports of this kind because these reports are or may be public. Information about specific item content cannot be made public as that would invalidate scores based in any part on those items.

three two-part items and added a standard for the second part of these items, resulting in a 6% partial match overall. Panelists selected a different standard than the intended standard for 17% of the items.

English Language Arts Grade 10. Panelists reviewed a form of the grade 10 ELA test consisting of 65 items. The grade ten ELA panelists selected standards that agreed with the intended standards on the majority of items (65%). The panelists partially matched the intended standard on 15% of the items. For four two-part items, they reported two standards, one of which matched the intended standard. The panelists added a second standard for six items: one that matched the intended standard and one in addition to that standard. Panelist selected a different standard than the intended standard for 20% of the items.

Summary of English Language Arts Florida Standards Comparison. The majority of the items in ELA had exact matches with the intended Florida Standards (65%-76%). However, for those that did *not* have exact matches for the Florida Standards ratings (31% of the total), the majority (64% of the 31%) actually represented a very close connection (e.g., alignment with slightly different content within the same anchor standard), while 36% of the 31% had no connection to the standard (n=16 items across all three grade levels). Specific information related to the items where panelists selected a different standard than the intended standard can be found in a separate, confidential report provided directly to FLDOE for consideration in future item revision and development processes.

Math Grade 4. Panelists reviewed a form of the grade 4 Math test consisting of 64 items. The grade four Math panelists matched the intended standards for a large majority of the items (94%). Panelists selected a different standard than the intended standard for 6% of the items.

Math Grade 7. Panelists reviewed a form of the grade 7 Math test consisting of 66 items. The grade seven Math panelists matched the intended standards for a large majority of the items (79%). Panelists selected a different standard than the intended standard for 21% of the items.

Algebra 1. Panelists reviewed a form of the Algebra 1 test consisting of 68 items. The Algebra 1 panelists matched the intended standards for a large majority of the items (81%). Panelists selected a different standard than the intended standard for 19% of the items.

Summary of Math Florida Standards Comparison. The majority of the items (79-94%) in Math had exact matches with the intended Florida Standards. However, for those few items that were not rated as exact matches with the intended Florida Standards (16% of the total), the majority (81% of the 16%) actually represented a very close connection (e.g., alignment with slightly different content within the anchor standard) while 19% of the 16% (n=6 items) had no connection to the standard. There were instances where a different Math area was identified, but the concepts and contexts overlapped. Specific information related to the items where panelists selected a different standard than the intended standard can be found in a separate, confidential report provided directly to FLDOE for consideration in future item revision and development processes.

Depth of Knowledge Comparisons

After panelists' ratings had been collected, researchers compared the intended Florida DOK assignments designated to be assessed by each item with the DOK ratings provided by content experts on each panel.

For this data collection, panelists used the same 4-level DOK rubric as was used by FLDOE to rate the Florida content standards. Panelists first rated DOK independently for all items on a reviewed form, using descriptions of DOK levels provided by FLDOE. The facilitator for each grade and content group then led a discussion resulting in consensus ratings for the DOK for each item. Researchers compared the DOK ratings provided by FLDOE to the consensus DOK ratings provided by the content expert panels. (Note: For items with multiple parts, the state provided DOK for the item as a whole. Researchers used panelist ratings at the overall item level for comparisons.) Panelists rated the DOK level the same as that provided by the state 43-65% of the time for the ELA tests and 50-59% of the time for the Math tests. With few exceptions, the two DOK judgments that were not in exact agreement were, adjacent, or within one DOK rating. For example, on the scale of 1-4, rater X rated an item as 3 and the assigned rating by FLDOE was 2. In this case, the ratings were adjacent, or off by just one level. As another example, rater X rated an item as 1 and the FLDOE rating was 2. Again, the ratings were adjacent, or off by just one level. For ELA, panelist ratings that differed tended to be at a higher DOK level than that provided by the state. The opposite was true for Math. To clarify, the ELA items were rated as more cognitively complex (higher DOK) than the FLDOE assigned DOK and the Math items were rated less cognitively complex (lower DOK) than the FLDOE assigned DOK.

For DOK rating analyses, panelists' ratings are compared with the intended DOK ratings. Weighted averages are calculated for each DOK level, by multiplying the number of items in a level by that level number and then averaging those products. For example, if 6 items of the 20 items on a test are rated as DOK 1, 10 items are rated as DOK 2, and 4 items as DOK 3, the average DOK would be:

$$\frac{(6*1) + (10*2) + (4*3)}{20} = \frac{6 + 20 + 12}{20} = \frac{38}{20} = 1.9$$

This average can be calculated for intended DOK and rated DOK and the averages can be compared.

A difference between the target and rated DOK indices of less than or equal to .5 would be considered strong DOK consistency, a difference of less than 1 point but more than .5 points would be considered moderate, and a difference of 1 point or greater would represent weak evidence of DOK consistency. This methodology and studies have been used by the evaluation team in a number of studies conducted with other states, have been approved by their Technical Advisory Committees (TAC), and have been accepted in United States Peer Review documentation for those states.

English language arts grade 3. Panelists provided DOK ratings in the range of one to three (out of four levels on the DOK rubric), which coincided with the range of intended DOKs provided by FLDOE (see Table 5). Panelists rated 55% of the items with the same DOK level.

Level by level, DOK ratings were much higher on average than intended for level 1, slightly higher than intended for level 2, and lower than intended for level 3. Of the 13 items intended to reflect level 3 DOK, panelists concurred for only four items. However, panelists determined that seven of the 32 items intended to reflect level 2 DOK actually reflected level 3. In total, the average rated DOK across items (2.1) is slightly higher than intended (2.0) which indicates strong DOK consistency.

Table 5. DOK Ratings for English Language Arts Grade 3

Panelists'	FLDOE/AIR Ratings			
Ratings	**1**	**2**	**3**	**Total**
1	4			4
2	11	25	9	45
3		7	4	11
Total	15	32	13	60

English language arts grade 6. As described in Table 6, panelists provided DOK ratings in the range of one to four. Panelists rated 65% of the items with the same DOK level. Further, panelists rated 11 of the 14 items the state rated a DOK level one as DOK level two; 8 of the 38 items the state rated a DOK level two as DOK level three; 1 item the state rated a DOK level two as DOK level one; and 2 of the 10 items the state rated a DOK level three as DOK level two. Both entities rated the writing item a DOK level 4. Overall, the DOK ratings were slightly higher than intended (2.2 vs. 1.9) indicating strong DOK consistency.

Table 6. DOK Ratings for English Language Arts Grade 6

Panelists'	FLDOE/AIR Ratings				
Ratings	**1**	**2**	**3**	**4**	**Total**
1	3	1			4
2	11	29	2		42
3		8	8		16
4				1	1
Total	14	38	10	1	63

English language arts grade 10. Panelists provided DOK ratings in the range of two to four, which was narrower than the range of one to four indicated by FLDOE. As shown in Table 7, panelists rated 43% of the items with the same DOK. Further, panelists rated all 16 items the state rated a DOK level one as DOK level two (n=12) or DOK level three (n=4); 17 of the 32 items the state rated a DOK level two as DOK level three; and 4 of the 16 items the state rated a DOK level three as DOK level two. Both entities rated the writing item a DOK level 4. Overall, the DOK ratings were somewhat higher than intended (2.5 vs. 2.0) indicating strong DOK consistency.

Table 7. DOK Ratings for English Language Arts Grade 10

Panelists'	FLDOE/AIR Ratings				
Ratings	**1**	**2**	**3**	**4**	**Total**
1					0
2	12	15	4		31
3	4	17	12		33
4				1	1
Total	16	32	16	1	65

Mathematics grade 4. Panelists provided DOK ratings in the range of one to three, which coincided with the range provided in the standards by FLDOE. As described in Table 8, panelists rated 52% of items with the same DOK level. Further, panelists rated 6 of the 14 items the state rated a DOK level one as DOK level two. Of the 45 items the state rated a DOK level two, 1 was rated as DOK level three and 21 as DOK level one. Three of the 5 items the state rated a DOK level three as DOK level two. Overall, the rated DOK level was slightly lower than intended (1.6 v. 1.9) but still with strong DOK consistency.

Table 8. DOK Ratings for Mathematics Grade 4

Panelists'	FLDOE/AIR Ratings			
Ratings	1	2	3	Total
1	8	21		29
2	6	23	3	32
3		1	2	3
Total	14	45	5	64

Math grade 7. Panelists provided DOK ratings in the range of one to three, which coincided with the range provided by FLDOE. As shown in Table 9, panelists rated 59% of the items with the same DOK level. In addition, panelists rated 1 of the 9 items the state rated a DOK level one as DOK level two; 21 of the 51 items the state rated a DOK level two as DOK level one; and 5 of the 6 items the state rated a DOK level three as DOK level two. Overall, the DOK ratings indicated somewhat lower DOK than what was intended for this test (1.6 v. 2.0) but still indicating strong DOK consistency.

Table 9. DOK Ratings for Math Grade 7

Panelists'	FLDOE/AIR Ratings			
Ratings	1	2	3	Total
1	8	21		29
2	1	30	5	36
3			1	1
Total	9	51	6	66

Algebra 1. Panelists provided DOK ratings in the range of one to three, which coincided with the range provided by FLDOE. As described in Table 10, panelists rated 34 of the 67 (51%) items at the same DOK level as was intended. Level by level, DOK ratings were slightly higher on average than intended for level 1, somewhat lower than intended for level 2, and lower than intended for level 3. Of the 7 items intended to reflect level 3 DOK, panelists concurred for only one item. However, panelists determined that four of the 47 items intended to reflect level 2 DOK actually reflected level 3. In total, the average rated DOK across items is slightly lower than intended (1.7 v 1.9) but as with the other grades reviewed, still indicates strong DOK consistency.

Table 10. DOK Ratings for Math Algebra 1

Panelists’	FLDOE/AIR Ratings			
Ratings	**1**	**2**	**3**	**Total**
1	9	19		28
2	4	24	6	34
3		4	1	5
Total	13	47	7	67

In summary, a difference between the target and rated DOK indices of less than or equal to .5 would be considered strong DOK consistency. Each grade and content area reviewed in this study resulted in DOK indices of less than or equal to .5. However, as with any review of alignment, average DOK ratings varied somewhat from what was intended. Delving deeper into the data and reviewing the three Math grades in total, rated DOK was slightly lower than intended for all three grades evaluated. These differences were mostly due to the significant number of items that were intended to reflect level 2 DOK but were rated as DOK 1. In contrast and reviewing the three ELA grades in total, average DOK ratings were slightly or somewhat higher than intended. These differences were due to the significant number of items that were intended to reflect level 3 DOK but were rated as DOK 2. As indicated below in Table 11, 37% of the ELA DOK ratings were above the intended DOK while 36% of the Math DOK ratings were below the intended DOK. These patterns could indicate that DOK may not be as closely attended to during item construction or item writer training as would be best practice and that additional external reviews of DOK may be necessary to align items to intended DOK levels as they are being developed. Given the intent of FLDOE to write new items aligned with the Florida Standards and to phase out the items included on the FSA that were originally developed for use in Utah, FLDOE should ensure tight content and cognitive complexity alignment in these newly developed items.

Table 11. Relationship between Intended DOK and Panelists’ DOK Ratings

	ELA		Math	
Comparison with Intended DOK	**N**	**%**	**N**	**%**
Higher	70	37	16	8
Match	102	54	110	56
Lower	16	9	71	36
Total number of items	188		197	

Fairness, Bias, Sensitivity, Accessibility, and Purposeful Item Development to Reduce the Likelihood of Guessing

Evidence of test content related to fairness, bias, and sensitivity was heavily documented during the development of the items for use in Utah. AIR and Utah Department of Education staff conducted and documented multiple rounds of committee reviews focusing on fairness, bias, sensitivity, and parent/community input. However, due to the limited time frame for developing the FSA, reviews by Florida stakeholders were not conducted. FLDOE did conduct content reviews with Florida content experts at the state level and psychometric reviews with psychometricians at the state level, but Florida stakeholders such as classroom teachers, content coaches/instructional specialists at the district level, and parents and other community representatives, as noted previously, did not review the items appearing on the FSA. To evaluate fairness, bias, and accessibility concerns, the evaluation team conducted item reviews with content/test development specialists to specifically review the FSA items for racial/ethnic/cultural considerations, sex and gender bias considerations, and socio-economic considerations.

Fairness, Bias, and Sensitivity Review

The evaluation team reviewed the same grade and content area forms as the item review panelists (grades 3, 6, and 10 in ELA and Math grades 4, and 7, and Algebra 1). Experts noted a concern in grade 6 ELA with a passage posing a negative presentation or stereotype of a female which was later dispelled in the passage. In Math, experts did not find any specific considerations, but did note that of the protagonists presented in items, 70% were male. Experts determined that the items reviewed for this evaluation suggested the FSA was fair and free from bias.

Finally, this review included two additional considerations: 1) is the assessment accessible or does it pose barriers for students with vision, hearing or mild-moderate intellectual disabilities, and 2) do particular design characteristics of items reduce the likelihood that the student answers the question correctly by guessing (e.g., no cue in stem or answer choices, appropriate and quality distractors for answer choices).

English Language Arts Content Area Review for Accessibility

The evaluation team reviewed the accommodated paper-based English Language Arts items at grades three, six, and ten to identify possible barriers for students with vision, hearing, or intellectual disabilities. These accommodated forms contain all of the same items in grades 3 and 4 but due to the computer-based administration in the remaining grades, the accommodated forms include a small number of items that differ from the online administration for the purposes of ensuring access, in particular for students with unique vision needs. In addition to the individual items, the evaluation team reviewed test procedures for all students and allowable accommodations for students with disabilities.

Students who are blind or deaf-blind can access items using the accommodations of braille (contracted or uncontracted), enlarged text, magnification devices, color overlays, one-item-per-page, special paper (e.g., raised line) or masking. In the braille versions of the tests, items may be altered in format (e.g., long dash to indicate first blank line) and may provide description of graphics, provide tactile graphics, and/or omit graphics. Students who have vision and hearing impairments are able to access writing items using a scribe.

Students who have mild-moderate intellectual disabilities can access the majority of the items using allowable accommodations such as oral reading/signing of items and answer options, one-line-per-page, special paper (e.g., raised line) and masking. Students may receive verbal encouragement (e.g., "keep working," "make sure to answer every question") which increases some students' ability to complete the test. Students can use alternative augmentative communication systems, including eye-gaze communication systems and signing (ASL/SEE) to respond to reading and writing items. Students are able to access writing items using a scribe (including ASL/SEE).

Given the interpretation of "reading" by FLDOE, use of a human reader is not an allowable accommodation to ensure the construct remains intact. Students who have mild-moderate intellectual disabilities and limited reading skills will have limited access to the passages without the use of a human reader. Students with vision or hearing impairments who also have limited ability to read, including reading braille, will have limited access to the passages without the use of a human reader. When required to read independently, these groups of students will not have the ability to demonstrate their understanding of the text beyond the ability to decode and read fluently. For example, without access to the passage, the students will be unable to demonstrate their ability to draw conclusions, compare texts, or identify the central/main idea.

Mathematics Content Area Review for Accessibility

The evaluation team reviewed the accommodated paper-based Math items at grades four and seven and for Algebra 1 to identify possible barriers for students with vision, hearing, or intellectual disabilities. In addition to the individual items, the evaluation team reviewed test procedures for all students and allowable accommodations for students with disabilities.

The accommodated paper-based test lacked some features that allow full access for students with vision impairments and mild-moderate intellectual disabilities. The computer-based features for all students allow the use of color contrast, however, there is no reference to same or similar allowances other than color overlays for the paper version of the test. The color contrast provides the option of inverted colors of the text and background and may be important for students with certain types of visual impairments such as Cortical Visual Impairment (CVI) to clearly view the items.

Students who are blind or deaf-blind can access the items using the accommodations of braille (contracted or uncontracted), enlarged text, magnification devices, color overlays, one-item-per-page, abacus, or masking. Students are able to respond to items through the use of a scribe; however, special care on constructed response items should be taken if a student with visual impairments does not use this accommodation as the response mode may increase the likelihood of "writing" errors for these students.

Students who have mild-moderate intellectual disabilities can access the majority of the items using allowable accommodations such as oral reading/signing of items and answer options, one-line-per-page, and masking. As with the ELA review, students may receive verbal encouragement (e.g., “keep working,” “make sure to answer every question”) which increases some students’ ability to complete the test. Students can use alternative augmentative communication systems, including eye-gaze communication systems and signing (ASL/SEE) to respond to Math items. Students can use a scribe as needed.

The paper-based test includes several items with graphics (e.g., coordinate grids, graphs, etc.), that include a description that can be read to or by the student or a tactile graphic. However, several graphics are visually complex, especially for students with visual impairments even with accommodations (e.g., tactile, description of graphic), as they require large amounts of information that must be stored in the students’ short-term memory.

Purposeful Item Development to Reduce the Likelihood of Guessing

This review included consideration of particular design characteristics of items that reduce the likelihood that the student answers the question correctly by guessing (e.g., no cuing in stem or answer choices, appropriate and quality distractors for answer choices). In both content areas, the reviews indicated item development included appropriate and quality distractors for answer choices and the stem or answer choices were free from language that would cue students to the correct answer choice. Further, the item writer training highlighted effective stem, effective options, and effective distractor development. Together, this information suggests items were developed to intentionally reduce the likelihood of guessing.

Response Processes

The *Test Standards* recommend (1.12) “if the rationale for score interpretation for a given use depends on premises about the … cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.” Evidence related to response processes should be documented through consideration of student performance and characteristics 1) during item development (e.g., through a principled development process/approach), 2) during test administration and gathered from the digital platform, or 3) through cognitive laboratories or interviews during item development, administration, or post hoc. During this review, AIR documented a principled item development approach but the only specific reference to response processes was in regard to acceptable response mechanisms designated as part of the item writing specifications. The response mechanisms more closely highlighted response formats acceptable for measuring the content rather than actual response processes used as expectations for the cognitive operations for students.

AIR provided the Smarter-Balanced Assessment Consortium (SBAC) Cognitive Laboratories Final Report for review, but it was not considered in this evaluation because there is no evidence indicating that any of the items reviewed in that study were ones that contributed to scores for Florida students. Studies conducted with items "similar to" those on the Florida tests do not offer any evidence regarding the quality of the items that did appear on Florida tests. We have no information about the definition of "similar" and the questions addressed in the SBAC study may, or may not, be ones of most importance for the assessments as administered in Florida. Further, while the item types on the FSA may be similar to those administered during the SBAC study, how similar or different those technology enhanced items play out via the platform for the FSA along with the interaction of the content within the platform is inconclusive.

Findings

Based on the documentation available and the studies/reviews completed related to the evaluation of the test items, the evaluation team did not find any evidence to question the validity of the FSA scores for the intended purposes. FLDOE and AIR made efforts to describe, document, and ensure content alignment, reduce item bias related to race, ethnicity, culture, sex/gender, and socio-economic considerations, increase accessibility of the test items especially for students who are deaf, blind, and have mild-moderate intellectual disabilities, and have adhered to industry standards as well as recommendations of the *Test Standards* in completing this work.

Based on the documentation available and the studies/reviews completed related to the evaluation of the test items, the evaluation team did not find any evidence to question the validity of the FSA scores for the intended purposes.

While a review of the items by stakeholders in Florida would be expected based on typical practice and the *Test Standards,* given the rapid development timeline and policy requirements, there was insufficient time to complete the review for the 2015 administration of the FSA assessment. FLDOE made substantial efforts to conduct a careful review of the items with content and psychometric experts to ensure the items matched Florida Standards. The majority of the items in ELA and Math had exact matches with the intended Florida Standards. When there was not an exact match, many of the items had matches with slightly different content within the same anchor standard.

As indicated earlier, for the three Math grades in total, rated DOK was slightly lower than intended for all three grades evaluated. These differences were mostly due to the significant number of items that were intended to reflect level 2 DOK but were rated as DOK 1. In contrast and reviewing the three ELA grades in total, average DOK ratings were slightly or somewhat higher than intended. These differences were due to the significant number of items that were intended to reflect level 3 DOK but were rated as DOK 2. These patterns could indicate that DOK may not be as closely attended to during item construction or item writer training as would be best practice and that additional external reviews of DOK may be necessary to align items to intended DOK levels as they are being developed. Given the intent of FLDOE to write new items aligned with the Florida Standards and to phase out the items included on the FSA that were originally developed for use in Utah, FLDOE should ensure tight content and cognitive complexity alignment in these newly developed items. Without conducting a Florida-specific stakeholder review of all the items appearing on the FSA test forms, FLDOE and AIR completed, at a minimum, the review necessary to safeguard the quality of the items and test forms used on the spring 2015 administration of the FSA.

Commendations

- AIR provided substantial documentation outlining the item development and review process for the items, as intended for Utah.
- FLDOE spent considerable time reviewing each and every item that appeared on the FSA with a content and psychometric lens.
- The majority of items reviewed by the evaluation team were
    - free from bias related to race, ethnicity, culture, sex/gender, and socio-economic considerations,
    - developed to be accessible for students with vision, hearing, and mild-moderate intellectual disabilities, and
    - developed to reduce the likelihood of guessing with effective stems, options, and distractors.

Recommendations

Recommendation 1.1 FLDOE should phase out the Utah items as quickly as possible and use items on FSA assessments written specifically to target the content in the Florida Standards. While every item appearing on the FSA was reviewed by Florida content and psychometric experts to determine content alignment with the Florida Standards, the items were originally written to measure the Utah standards rather than the Florida Standards. The standards in these two states are very similar, but do vary within some shared anchor standards. Thus, while alignment to Florida Standards was confirmed for the majority of items reviewed via the item review study, many were not confirmed, usually because these items focused on slightly different content within the same anchor standards. As such, in these areas it would be more appropriate to use items written to specifically target the Florida Standards.

Recommendation 1.2 FLDOE should conduct an external alignment study on the entire pool of items appearing on the future FSA assessment with the majority of items targeting Florida Standards to ensure documentation and range of complexity as intended for the FSA items across grades and content areas. Further, the specifications for item writing relating to cognitive complexity should be revisited and items should be checked independently for DOK prior to placement in the item pool for administration.

Recommendation 1.3 FLDOE should conduct cognitive laboratories, cognitive interviews, interaction studies involving the capture and analysis of data about how students engage with test items and the content within each of the items during administration, and/or other

ways in which to gather response process evidence during the item development work over the next year.

Study 2: Evaluation of Field Testing

Study Description

For this study, the evaluation team reviewed documentation and data from the field test activities, supplementing this information with an in-person meeting with FLDOE and partner vendor staff. The planned field test study activities included:

- A review of the sampling plan for the following:
    - Design characteristics that are consistent with intended purpose(s)
    - Processes for creating the sampling plan
    - Extent to which the sampling plan was executed as expected
    - Processes and procedures to ensure evidence of sufficient sample size and population representation
- A review of the ability of field test results to support test form construction
- A review of whether the field test results yield results that support a range of raw scores that would be transformed into scale scores relative to cut scores
- A review of the decision rules that were applied to the results of the field test

Sources of Evidence

To conduct the review of the FSA field testing, AIR supplied the primary sources of data and information for the procedures for the field testing in the form of technical reports for the 2013-14 Utah state assessment program. These documents were:

- Utah State Assessment, 2013-14 Technical Report: Volume 1 Annual Technical Report
- Utah State Assessment, 2013-14 Technical Report: Volume 2 Test Development
- Utah State Assessment, 2013-14 Technical Report: Volume 3 Test Administration
- Utah State Assessment, 2013-14 Technical Report: Volume 4 Reliability and Validity

For the review of the Florida-based field testing activities, many of the analogous documents and data that were available for the Utah-based field testing were not yet available at the time of this evaluation. Instead, this review was conducted using a variety of internal memos written specifically for this evaluation, conversations with key staff involved in the procedures, and working documents used to track work activities.

Study Limitations

As is mentioned in the previous section, formal documentation related to the processes used to evaluate items in place of a field test with Florida students were not yet available. This is not surprising given that formal technical manuals are commonly generated after the completion of the program year and therefore likely won't be ready until fall 2015 for the first year of FSA. AIR and FLDOE were able to provide the needed information to complete the evaluation of FSA field testing as it was originally designed.

Industry Standards

Appropriate field testing of test content is a critical step for testing programs to evaluate the empirical characteristics that contribute to the overall quality of the assessment items and test forms. Even after the most rigorous item development process, field testing of items by exposing the items to large groups of students under standardized conditions allows for statistical and content reviews that eliminate possibly problematic items and help ensure the reliability, validity, and fairness of the assessments. With respect to field testing, the *Test Standards* state that:

> The purpose of a field test is to determine whether items function as intended in the context of the new test forms and to assess statistical properties. (p. 83)

While the *Test Standards* do not provide prescriptive methods for how and when field testing should be completed, they do provide important guidelines that need to be considered when looking at any field testing. Specifically, *Test Standards* (4.9) discuss the importance of gathering a sufficient and representative sample of test takers for the field testing. The sample size also needs to be sufficient to support intended psychometric analysis procedures, such as Differential Item Functioning (DIF) methods that are designed to help evaluate empirical evidence of the fairness of the examination across student groups.

The *Test Standards* (4.10) also discuss the importance of documenting any assumptions of the scoring model that have been adopted when reviewing the field test results. For example, any data screening rules for the items and students should be clearly documented for all phases of the work; clear rationales for these rules should also be provided. Similarly, if multiple Item Response Theory (IRT) scoring models are considered and evaluated, the assumptions for each model should be documented, and the data and evidence to support the models selected should be provided.

"The process by which items are screened and the data used for screening… should also be documented."
(Test Standards, 2014, p. 88-89)

In addition to considering the types of evidence for which we expect to evaluate compliance with the *Test Standards*, our review also focused on industry best practices and the current state of research in the field. One of the persistent problems in field testing items is student motivation. If students are informed that an assessment is solely for field testing purposes (i.e., little or no stakes for students, their teachers, and their schools) students have limited motivation to perform their best. Therefore, the assessment community recommends that, when feasible, field testing be conducted by embedding items within operational test forms where the student is unaware of which items are being field tested and which are operational items (Haladyna & Rodriguez, 2013; Schmeiser & Welch, 2006).

However, in cases where new assessment programs are being introduced, it is not normally feasible to embed items into an existing assessment program; this make it more challenging to field test items. In some scenarios, field testing can be conducted as stand-alone events, solely for the purposes of trying out items and/or test forms (Schmeiser & Welch, 2006).

> "The items were screened for DIF with the groups including ethnicity, gender, English Language Proficiency, and income status."

Florida Standards Assessments Processes and Evaluation Activities

Although most field tests occur with samples of the intended population, the FSA field testing was completed with students in another state; the item bank used for the spring 2015 FSA administration was licensed from Utah's Student Assessment of Growth and Excellence (SAGE) assessment program. This method for gathering items for 2015 was primarily necessitated due to the limited timeframe available to develop and review test items for the FSA. Because the 2015 FSA items were licensed from the state of Utah, the review of the FSA field testing started with a review of the field testing methods, procedures, and results that occurred with students from Utah. After this step, the policies and procedures that were followed to transition from the Utah item bank to the FSA were also reviewed.

Utah-Based Field Testing Activities

The policies and procedures that were followed to develop test items is reviewed as part of Study #1 in this evaluation, and are not repeated here. This section focuses on how items were field tested and the appropriateness of these processes relative to the *Test Standards* and best practices. All items that were considered viable items for Utah were field tested during the operational 2014 test administration of the Utah state assessments. Prior to scoring the assessments, all items were screened for appropriate statistical performance. The statistical performance of all items was reviewed. Items with any of the criteria listed below were flagged for further content based reviews.

- Proportion correct value is less than 0.25 or greater than 0.95 for multiple-choice and Constructed-response items; proportion of students receiving any single score point greater than 0.95 for constructed-response items (see *Item Difficulty* in Appendix A).
- Adjusted biserial/polyserial correlation statistic is less than 0.25 for multiple-choice or constructed-response items (see *Item Discrimination* in Appendix A).
- Adjusted biserial correlations for multiple-choice item distractors is greater than 0.05.
- The proportion of students responding to a distractor exceeds the proportion responding to the keyed response for MC items (i.e., option analysis).
- Mean total score for a lower score point exceeds the mean total score for a higher score point for constructed-response items. (Utah State Assessment, Volume 1: p. 15).

The items were also screened using DIF (see *Differential Item Functioning [DIF]*, in Appendix A) with these analyses completed for groups defined by ethnicity, gender, English Language

Proficiency, and income status. For the DIF analyses, any item classified at the C level of DIF (i.e., the most significant level) was flagged and sent for further review (see Camilli, 2006, at pp. 237-238). Each of the SAGE assessments were taken by approximately 37,000 to 47,000 students for English Language Arts, and approximately 17,000 to 44,000 students in Math, depending upon the grade level.

Florida-Based Field Test Activities

One critical point that must be considered when looking at the FSA field testing is the actual purpose of using items from the Utah item bank. For Florida, the items that were licensed from Utah presented an opportunity to identify items that were appropriate to measure Florida's academic content standards and that had been previously field tested and had demonstrated appropriate statistical performance. This selection of items did not guarantee that all of the items from Utah would be appropriate for the FSA. Instead, it allowed Florida to select from items that FLDOE could be reasonably confident would demonstrate acceptable statistical performance when used on the FSA.

While the statistical performance of the items provided some assurance that the items would behave appropriately if used as part of the FSA, it did not guarantee that the items were appropriate for Florida students. To address these concerns, FLDOE, in collaboration with AIR, completed an item review to determine if the items were appropriate with respect to content in addition to statistical qualities. The reviews started with an available pool of approximately 600 items per grade level and test. These items were evaluated for their statistical performance as well as other characteristics, such as word count, passage length, and content alignment with Florida's academic content standards. After this review, approximately 180 to 200 items remained as part of the pool of items for each test.

To finalize the item pool, in July and August of 2014, FLDOE and AIR worked together to conduct a final review of all items. From these items, test forms were constructed to meet the psychometric, content, and blueprint requirements for each test form. Throughout this process, the range of items available and the performance of the items provided sufficient data and information for all test forms to be constructed so that full range of test scores could be supported in the 2015 spring test administration. After constructing each test form, staff members from FLDOE completed a final review of all items and test forms to ensure that met all documented requirements. Finally, as described in Study #5, all items on the FSA were screened after the 2015 spring administration using data collected from Florida students before being used as operational test items. For any items where concerns remained after post-administration reviews, the items were removed from the scorable set, meaning that they did not impact student scores.

"Prior to use on the FSA, all items were reviewed by FLDOE staff who were familiar with Florida students and the Florida standards."

Findings

For this evaluation, the policies and procedures used in the field testing of test forms and items were evaluated and compared to the expectations of the *Test Standards* and industry best practices. While the FSA field testing was completed through a nontraditional method, the data collected and the review procedures that were implemented were consistent with industry-wide practices. The rationale and procedures used in the field testing provided appropriate data and information to support the development of the FSA test, including all components of the test construction, scoring, and reporting.

Commendations

- The field test statistics in Utah were collected from an operational test administration, thus avoiding questions about the motivation of test takers that normally accompany traditional field testing methods.
- During the Utah field testing process, the statistical performance of all items was reviewed to determine if the items were appropriate for use operationally.
- Prior to use of the FSA, all items were reviewed by educators knowledgeable of Florida students and the Florida Standards to evaluate whether the items were appropriate for use within the FSA program.
- After items were administered on the FSA, the statistical performance was evaluated again; items were only used after the statistical performance of the items was evaluated and items with problematic statistics were reviewed based on Florida data and excluded from student scoring if needed.

Recommendations

Recommendation 2.1 FLDOE should provide further documentation and dissemination of the review and acceptance of Utah state items.

FLDOE should finalize and publish documentation that provides evidence that the FSA field testing policies, procedures, and results are consistent with industry expectations. While some of this documentation could be delayed due to operational program constraints that are still in process, other components could be documented earlier. Providing this information would be appropriate so that Florida constituents can be fully informed about the status of the FSA. Some misconceptions existed about the FSA being a Utah-based test and therefore not appropriate for Florida students. The lack of documentation and information for the public regarding the use of Utah items and the review processes that FLDOE employed may have helped support some of these misconceptions.

Further public documentation for the field testing process is highly recommended.

Study 3: Evaluation of Test Blueprint and Construction

Study Description

This study focused on the consistency of the test blueprint and construction process with the intended interpretations and uses of test scores. Along with a review of the documentation of the test development process, the evaluation team conducted in-person and virtual interviews with FLDOE and AIR to gather information not included in documentation or to clarify evidence. The following elements were planned for inclusion within this study:

- Review of the process for the test construction to evaluate its consistency with best practices
- Review of the test blueprints to evaluate if the blueprints are sufficient for the intended purposes of the test
- Review the utility of score reports for stakeholders by considering the following:
    - Design of score reports for stakeholder groups
    - Explanatory text for appropriateness to the intended population
    - Information to support improvement of instruction

Sources of Evidence

The following documents served as the primary sources of evidence for this study:

- FSA Test Construction Specifications (Draft 2015)
- Description of the Blueprint Development Process
- ELA and Mathematics Test Design Summary Documents
- PLD Development Summary Report
- Item Form Selection Process Report
- Item Data Review action/approval logs
- Student Report Mock-ups
- Online Reporting System Mock-ups

Study Limitations

The second focus of this study involved the review of FSA score reports. Given the timing of this study and ongoing program development activities, actual reports were not available and FLDOE and AIR provided mock reports for the FSA for this review. FLDOE and AIR did not provide samples of the interpretive guides that are to accompany score reports and aid in score interpretation and use because these documents are still under development. The findings here represent statements about what the score reports and interpretive guides should include to meet ESEA requirements and to support the uses of test information by educators.

Industry Standards

Common questions such as, “What’s on the test?” and “How well are my students doing in relation to the standards?” rely on evidence related to test content. A large-scale standardized test designed to help answer these questions must be built to do so for every student in the testing population and in ways that support comparable interpretations across students, sites, and time.

Developers are also to "document the extent to which the content domain of a test represents the domain defined in the test specifications" (Test Standards, p. 89).

With regard to test content, the *Test Standards* state that "the domain definition should be sufficiently detailed and delimited to show clearly what dimensions of knowledge, skills, cognitive processes, attitudes, values, emotions, or behaviors are included and what dimensions are excluded" (*Test Standards,* p. 85). Developers are also to "document the extent to which the content domain of a test represents the domain defined in the test specifications" (*Test Standards,* p. 89). These standards are meant to ensure that each instance of a test administration yields information that is interpretable in relation to the knowledge and skills domain the test is meant to measure. A test blueprint is, in many cases, the de facto definition of the knowledge and skill domain in the context of the test. As such, the blueprint should clearly reflect the external-to-the-test domain definition, which is the case of the FSA and the Florida Standards. In addition to demonstrating a clear relationship to a domain definition, evidence related to test content should include support for comparable interpretations of student performance in relation to that domain across students, sites, and time. While comparability is often thought of in the sense of reliability, here we focus on the validity concern that a test must be constructed in ways that allow for comparability in score interpretations about the target knowledge and skill domain.

Testing consequences encompass a broad range of considerations, from an individual student's cognitive or emotional take-aways from a testing situation to educators determining how to use information from tests to reflect upon their curricula and instructional practices to policy-makers deciding via accountability systems how to distribute resources. In this study, we focus on the second of these examples. Educators' use of test information to support reflection upon their curricula and instructional practices relies upon the receipt of information that is (a) meaningful in relation to the academic standards that guide their curricular and instructional decisions and (b) communicated in clear terms.

In regard to evidence related to testing consequences, the *Test Standards* (12.19) state that "in educational settings, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided" (p. 201). Further, the *Test Standards* (12.18) state that score reports must provide clear information about score interpretation, including information on the degree of measurement error associated with a score or classification. The *Test Standards* (6.8) emphasize that test users (in the present case, FLDOE) should use simple language that is appropriate to the audience and provide information on score interpretation such as what a test covers, what scores represent, the errors associated with scores, and intended score uses.

Florida Standards Assessment Processes and Evaluation Activities

For the review of the test blueprint and construction, AIR and FLDOE provided documentation similar to what is expected under industry standards and recommendations in the *Test Standards.* Evidence about the item development process was extensive and clear. However, information necessary to conduct the alignment analyses, including information about the intact forms provided for review, was neither timely nor readily accessible to evaluators. The first part of this study involved the collection of ratings of FSA items by Florida stakeholders. It is important to note AIR and FLDOE provided access to grade-level intact forms for each of the grades and content areas reviewed during the item review study. The forms included both vertical linking items and field test items. The field test items were removed for the purpose of the review of the match to the blueprint. The vertical linking items were used as part of the vertical scaling process but were grade appropriate so those items were included for the purpose of the blueprint match analysis.

Pending conclusion of this evaluation, FLDOE will release the scores of the FSA prior to standard setting. As such, FLDOE will only report raw score and percentile rank information. The documentation for the review of score reports and interpretive guides did not meet industry standards because these documents are still under development. The status of development of these documents aligns with typical practice for a program in the first year of implementation.

Test Content

The content and skill areas a test is intended to measure must be sufficiently detailed to allow for the construction of a test that assesses those areas with fidelity in terms of breadth and depth. Such detail should be communicated in the form of a blueprint or other documents that articulate the characteristics of individual items that students encounter on a test and of the set of items that contribute to students' test scores. A blueprint of some sort is necessary to ensure that the test items individually and as a set target appropriately the intended content and skills; further a blueprint of some sort is necessary to ensure that tests can yield comparable results across students, sites, and time. The evaluation of a blueprint, its development, and its use in test construction involves both a qualitative capture of how a blueprint was developed in ways that meet industry standards and consideration of how it actually reflects the target content and skill area.

Given the abbreviated timeline to construct assessments for 2015, FLDOE did not have time to begin test- or item-development from 'scratch' or to implement a wide-reaching stake-holder involvement process prior to the first administration of the FSA. To ensure that the FSA items and forms could be ready for administration on the very short timeline, FLDOE staff established an intense review process that involved primarily internal content and psychometric experts in reviewing items and adjusting blueprints from those used in Utah to what would better fit the Florida context.

From the documentation provided, it is clear that content experts at FLDOE worked closely with AIR to make changes to the blueprint for each grade and content area. The intent of this process was to establish blueprints that better reflected the Florida Standards and FLDOE expectations for its tests forms. The content team flagged issues such as misalignment of content and then the flagged items were reviewed for inclusion on the test or replacement based on the FLDOE input. Florida psychometricians reviewed the performance characteristics of the items intended for use in Florida. The reviews started with an available pool of approximately 600 items per grade level and test. These items were evaluated for their statistical performance as well as other characteristics, such as word count, passage length, and content alignment with Florida’s academic content standards. After this review, approximately 180 to 200 items remained as part of the pool of items for each test. This low level of item survival suggests that the item review criteria were rigorous with regard to alignment with Florida’s standards and vision for the FSA.

During the item review process, discussions among FLDOE and AIR staff were documented through test summary construction sheets that mapped the pathway for placement of items on the final forms. FLDOE reviewers considered bias issues as they reviewed the items, specifically to ensure Utah-centric items were eliminated and did not appear on the FSA. The FSA ELA and Math test design summary documents include the percentage of items in each content category, cognitive complexity, and the approximate number of assessment items.

Although statewide stakeholder involvement was not an option under the first year of the FSA development timeline, ELA and Math content experts at the Test Development Center, a partner group of FLDOE that contributed to FSA development, conferred with content experts in the Florida Department of Education’s Bureau of Standards and Instructional Support and Just Read Florida office to solidify the content of the blueprints. These meetings and calls occurred during May and June, 2014.

In addition to the reviews of the items and the blueprints, FLDOE established reporting categories for the new FSA. The reporting categories for ELA were derived from the “domain” naming convention in the Florida Standards. Speaking and Listening standards were folded into the Integration of Knowledge and Ideas reporting category, and Text-Based Writing was added in grades 4-10 since the writing assessment occurs in those grades. Guidelines for the weight of each reporting category was determined by Florida’s Technical Advisory Committee (TAC) who suggested that to avoid “statistical noise” generated from the items scored in a small reporting category, a minimum of fifteen percent of the entire test should be derived from each reporting category. In some cases, “domains” may have been logically combined to adhere to the fifteen percent rule. The reporting categories for Math were also derived from the “domain” naming convention in the Florida Standards. Like ELA, if a Math domain had too few standards, two or more domains might be combined to make the reporting category fifteen percent of that grade's assessment.

Evaluation of the blueprint involved the use of the item ratings described in Study 1 (i.e., the same ratings were used for both Study 1 and Study 3), the published blueprints, and characteristics of the items in the item sets used for the item review. Only content was considered in the blueprint evaluation because the blueprints do not provide any indication of standard specific cognitive complexity expected of the items that make up the forms. Such information is clearly specified in the item writing and internal item review documents in ways that support the development of items that match the standards in both content and cognitive complexity terms.

The logic underlying the blueprint holds that the blueprint is the translation of the knowledge and skill domain defined in the standards for the purpose of test construction. The items, as compiled on a test form by the developer, should conform to the blueprint and independent, external reviewers should provide evidence that that is the case. If the Florida Standards are thought of as the large circle in the sense of a Venn diagram, the blueprint should represent a sample of that domain that is adequate in terms of content match and cognitive complexity as determined by content experts and adequate to support quality score production as determined by psychometricians. The items on any given test form are yet a sample of the items that could populate that form. The items that are reviewed must be considered representative of items that actually do appear on a typical test form. The evaluation considers whether those items were appropriately identified by the vendor to populate the form and whether they reflect the specific standards and cognitive complexity the vendor claims they do.

As noted above, we did not consider cognitive complexity in evaluating the blueprints because no relevant indicators were provided for each standard. However, in Study 1 we evaluated the cognitive complexity of the items in the review sets; the outcomes of that study indicated that the cognitive complexity of the items conformed well to the intended cognitive complexity established by the item writers.

This evaluation considered blueprints and item sets in grades 3, 6, and 10 for English Language Arts, in grades 4 and 7 for Math, and for the Algebra 1 End-of-Course (EOC) assessment. Panelists considered documentation about how the blueprints were adapted to reflect the Florida Standards as well as the structure and overall content of the blueprints in relation to the Florida Standards. Panelists used information about what the items were intended to measure in terms of content and cognitive complexity gleaned from vendor-provided files and ratings gathered from the content experts that served as panelists to evaluate fidelity of the items to the blueprint and of the item characteristics to the intended item characteristics.

Reviews of the items considered both content and cognitive complexity in analyses not involving the blueprint. Specific information about blueprints and items is not provided in this report to protect the security of these items.

The blueprints are organized by category as follows:

Grade 3 ELA	Grades 6 and 10 ELA	Grades 4 and 7 Math	Algebra 1
Key Ideas and Details Craft and Structure Integration of Knowledge and Ideas Language/Editing Task	Key Ideas and Details Craft and Structure Integration of Knowledge and Ideas Language/Editing Task Writing Task	Operations and Algebraic Thinking Numbers and Operations in Base Ten Numbers and Operations – Fractions Measurement, Data, and Geometry	Algebra and Modeling Functions and Modeling Statistics and the Number System

The results here are presented in terms of general overlap of standards on the blueprint and standards indicated for the items on the test forms. It is important to note that the set of items on any test do not necessarily have to address each and every standard on a blueprint. The FSA blueprints, like those in many states, indicate the possible range of item counts for a given category and standard within category; as long as the range of items within a category is somewhat balanced (e.g., items related to several of the standards within a category such as Key Ideas and Details) rather than clustered on only a small proportion of the standards in that category, leaving out some standards on a test form – which serves as an instance of the blueprint – is not of concern and meets industry standard.

For grade 3 ELA, the items covered all but five of the standards and did not reflect any standards not on the blueprint. The results were the same for grade 10 ELA. Only one standard in the blueprint was not in the grade 6 ELA item set; one standard in the item set was not on the blueprint (see Figures 1-3 below).

The fidelity of the item sets to the Math blueprints in terms of content match was similarly strong. In grade 4, three blueprint standards were not on the form and all of the form standards were on the blueprint. The grade 7 Math items represented all but two of the blueprint standards and included two standards not on the blueprint. For Algebra, five blueprint standards did not appear on the form and all of the items on the form reflected blueprint standards (see Figures 4-6 below).

These results indicate that the items selected to be on the form reflect the overall content of the blueprints with fidelity. That is, FLDOE and AIR selected items that conformed to the broad content of the blueprints. When considered in combination with the item review results from Study 1, these results further indicate that the forms, as reviewed by panelists, conform to the blueprints because of the strong degree of agreement between the intended content of the items and the panelists' ratings.

A second set of analyses compares the blueprints, intended item content, and item content as rated by panelists in terms of proportions of items across the level of the categories listed above. In Figures 1 through 6, results are presented in graphic form and numerically.

The results for Math are all strong and positive. The items selected to reflect the blueprints and the proportions indicated in the blueprints did reflect those proportions and panelists' ratings support this fidelity.

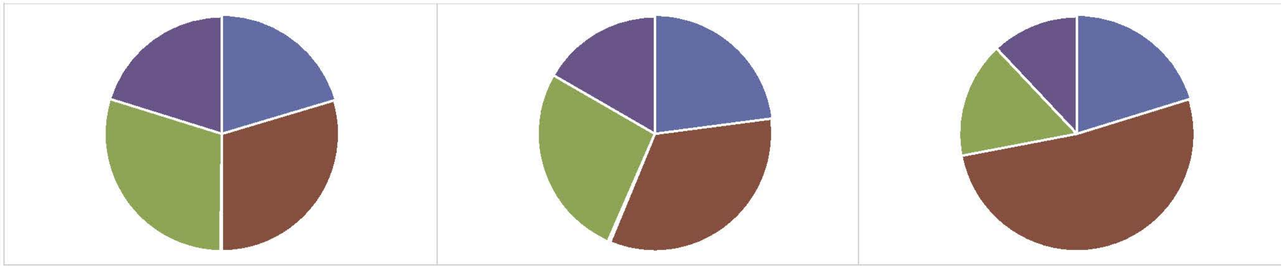
The results for ELA are generally positive, although a few of the categories were either under- or over-represented as indicated in the panelists' ratings. This result emerged even with the general agreement between the vendor ratings of the items and the panelist ratings described in Study 1. When there was not agreement between these ratings, the differences sometimes meant that the item was rated as reflective of a standard in a different category.

Even with these differences in proportion, however, the findings for ELA suggest the need to review the panelists' ratings and comments but do not raise critical concerns about the validity of the test score interpretations. The correlations among subscores, which would be scores for individual categories such as Key Ideas and Details, is typically very high within a content area and some variation in proportion from the blueprint and over time is common.

Grade 3 ELA

Standards on blueprint not on form = 5

Standards on form not on blueprint = 0



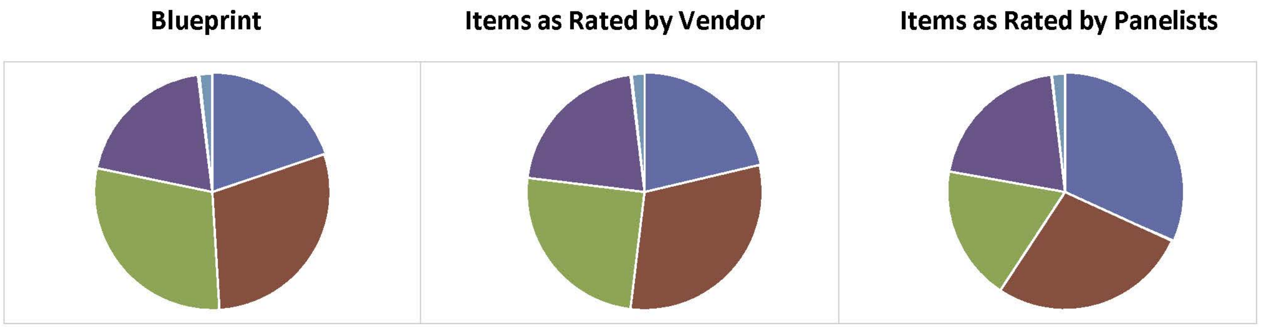
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Key Ideas and Details	0.21	0.23	0.20
● Craft and Structure	0.31	0.33	0.53
● Integration	0.31	0.27	0.16
● Language/Editing	0.21	0.17	0.12

Figure 1. Grade 3 ELA: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Grade 6 ELA

Standards on blueprint not on form = 1

Standards on form not on blueprint = 1



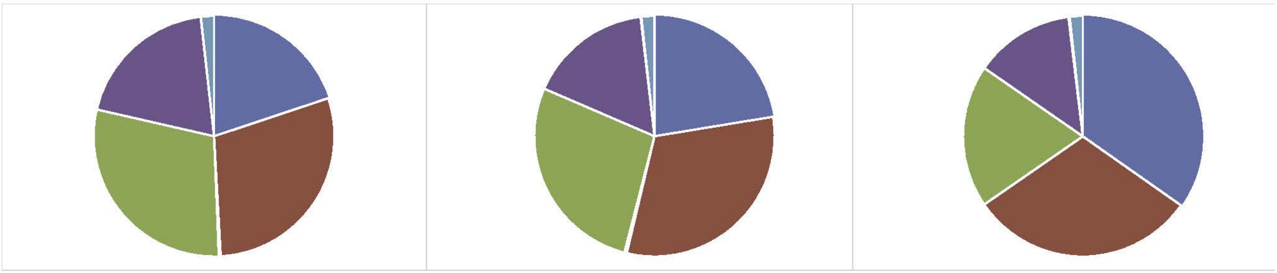
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Key Ideas and Details	0.20	0.21	0.31
● Craft and Structure	0.30	0.31	0.28
● Integration	0.30	0.25	0.19
● Language/Editing	0.20	0.21	0.20
● Text based writing	0.02	0.02	0.02

Figure 2. Grade 6 ELA: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Grade 10 ELA

Standards on blueprint not on form = 5

Standards on form not on blueprint = 0



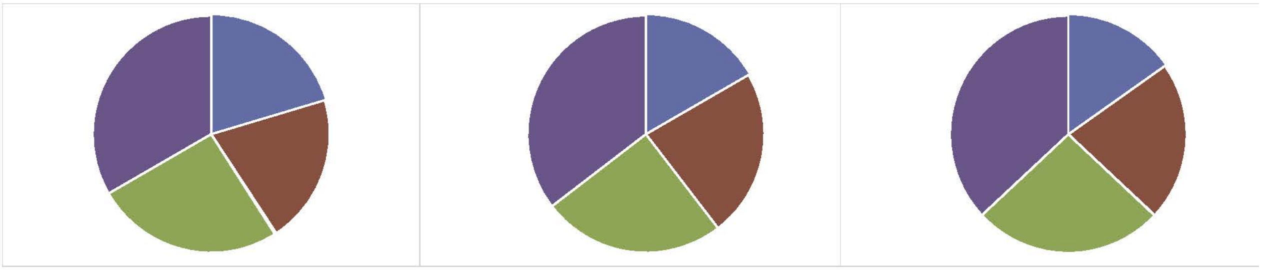
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Key Ideas and Details	0.20	0.22	0.35
● Craft and Structure	0.31	0.31	0.31
● Integration	0.31	0.28	0.19
● Language/Editing	0.20	0.17	0.13
● Text based writing	0.02	0.02	0.02

Figure 3. Grade 10 ELA: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Grade 4 Math

Standards on blueprint not on form = 3

Standards on form not on blueprint = 0



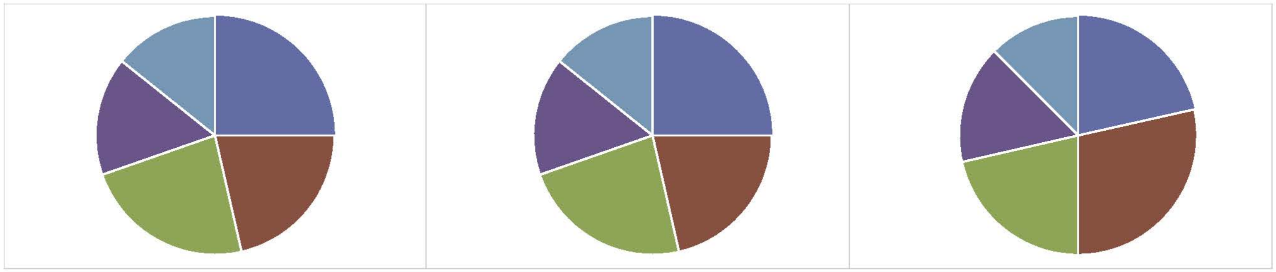
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Operations and Algebraic Thinking	0.20	0.17	0.15
● Numbers and Operations Base 10	0.20	0.23	0.22
● Numbers-Operations – Fractions	0.26	0.25	0.26
● Measurement, Data, Geometry	0.33	0.35	0.37

Figure 4. Grade 4 Math: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Grade 7 Math

Standards on blueprint not on form = 2

Standards on form not on blueprint = 2



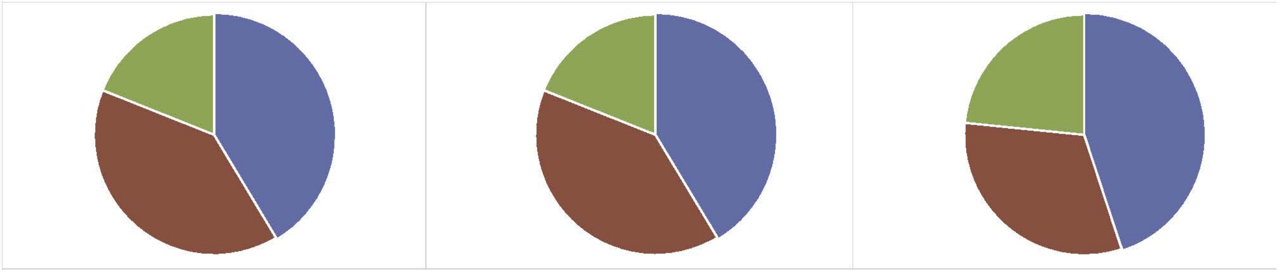
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Ratio & Proportional Relationships	0.25	0.25	0.21
● Expressions & Equations	0.21	0.21	0.29
● Geometry	0.23	0.23	0.21
● Statistics & Probability	0.16	0.16	0.16
● The Number System	0.14	0.14	0.13

Figure 5. Grade 7 Math: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Algebra 1 End of Course

Standards on blueprint not on form = 5

Standards on form not on blueprint = 0



	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Algebra and Modeling	0.41	0.41	0.45
● Functions and Modeling	0.40	0.40	0.32
● Statistics and the Number System	0.19	0.19	0.23

Figure 6. Algebra 1: Match between Standards on the Blueprint, Intended Standards of the Items, and Standards Rated by Panelists

Test Consequences

FLDOE and AIR provided mock-ups of the individual student reports they intend to use to communicate information about a student's test performance to students, parents, and teachers. These mock-up student reports were two pages in length and indicated the student's percentile rank and, for each of the reporting categories, the number of points the student earned, the number of points possible, and the average number of points earned statewide. Currently, the state does not plan to report scale score information or scores in relation to performance levels as required by ESEA given this is the first year of FSA implementation. However, the state does plan to provide a formula that can be used by districts to transform the t-score into a scale score so that districts can do their own analyses to retrofit scores for informational purposes. AIR and FLDOE evaluated several options to determine the interim standards and consulted with members of the Technical Advisory Committee (TAC) as well as an expert specializing in assessment and the law. Equipercentile linking of the cut scores from FCAT 2.0 to FSA was selected as the approach for establishing the interim cut scores for grade 3 ELA and Algebra 1.

FLDOE and AIR have yet to develop interpretive guides for the scores reports; therefore, this information could not be included within this evaluation. The status of development of these documents aligns with typical practice for a program in the first year of implementation.

Findings

FLDOE and AIR provided extensive documentation about the test development/adaptation process at the item and test blueprint levels. In the limited timeline available for FLDOE to establish a new assessment system, FLDOE took great care in adapting an existing test to meet the Florida Standards.

Given that the 2015 FSA was an adaptation of another state's assessment, much of the documentation about test development came from that other state. This documentation reflects an item development process that meets industry standards, although the documentation does not appear to be well represented in the body of technical documentation AIR offers, especially for an assessment that has been in place for more than one year. Likewise, the documentation of the original blueprint development process appears to have been adequate, but that information had to be pieced together with some diligence. The documentation about the process FLDOE undertook to adapt the blueprints and to select from the pool of available items reflects what would have been expected during a fast adaptation process. To facilitate stakeholders' understanding of the tests and the test scores, FLDOE should consider a review and reorganization of the information about how the FSA came to be. This is not a highly critical finding given the short FSA development timeline to date; the decision to prioritize activities related to development over documenting those activities this past year seems logical and reasonable.

The first set of blueprint analyses reviewed the general overlap of standards on the blueprint and standards indicated for the items on the test forms. Findings indicated that the blueprints that were evaluated (grades 3, 6, and 10 for English Language Arts, grades 4 and 7 for Mathematics, and Algebra 1) do reflect the Florida Standards in terms of overall content match. That is, FLDOE and AIR selected items that conformed to the broad content of the blueprints. When considered in combination with the item review results from Study 1, these results further indicate that the forms, as reviewed by panelists, conform to the blueprints because of the strong degree of agreement between the intended content of the items and the panelists' ratings. However, the lack of standard specific cognitive complexity expectations in the blueprints means that test forms could potentially include items that do not reflect the cognitive complexity in the standards and could vary in cognitive complexity across forms, thus allowing for variation across students, sites, and time. Given the extensive information in the item specifications, it would be possible to select items that meet cognitive complexity expectations when populating a test form if standard specific cognitive complexity were included on the blueprints. This exclusion of cognitive complexity from the blueprint does not meet industry standards.

A second set of analyses compared the blueprints, intended item content, and item content as rated by panelists in terms of proportions of items across the level of the categories listed above. The results for Math were all strong and positive. The results for ELA are generally positive, although a few of the categories were either under- or over-represented as indicated in the panelists' ratings. This result emerged even with the general agreement between the vendor ratings of the items and the panelist ratings described in Study 1.

In regard to test consequences and the corresponding review of score reporting materials, the individual score reports must include scale scores and indicate performance in relation to performance standards. The performance level descriptors must be included in the report as must some means for communicating error. Currently, this information is not included within the drafted FSA score reports given the timing of this evaluation and the intent to release reports prior to standard setting and consideration should be given to inclusion for subsequent years after standard setting is complete.

Given the timing of this review, FLDOE and AIR have yet to develop interpretation guides for the score reports. These guides typically explicate a deeper understanding of score interpretation such as what content is assessed, what the scores represent, score precision, and intended uses of the scores. These guides are critical to ensuring appropriate interpretation and intended use of the FSA scores. Given the use of FSA scores for promotion and graduation decisions as well as to improve instruction (FLDOE, 2015), it is important to document evidence outlining the impact on instructional practices and students' learning experiences and the appropriateness of this relationship between instruction and the FSA. As stated above, FLDOE and AIR have yet to develop interpretation guides for the FSA score reports. The status of development of these documents aligns with typical practice for a program in the first year of implementation. In subsequent years, specific information on the score reports and in the interpretation guides should be targeted directly at teachers and districts to support the improvement of instruction, especially in those areas related to the reporting categories. Further, technical documentation for the FSA outlining the validity of the intended uses of the scores should specifically document the rationale for and evidence supporting the relationship between instruction and the FSA.

Commendations

- FLDOE clearly worked intensely to establish an operational assessment in a very short timeline and considered on both content and psychometric concerns.

Recommendations

Recommendation 3.1 FLDOE should finalize and publish documentation related to test blueprint construction. Much of the current process documentation is fragmented among multiple data sources. Articulating a clear process linked to the intended uses of the FSA test scores provides information to support the validity of the intended uses of the scores.

> Finalizing and publishing documentation related to test blueprint construction is highly recommended.

Recommendation 3.2 FLDOE should include standard specific cognitive complexity expectations (DOK) in each grade-level content area blueprint. While FLDOE provides percentage of points by depth of knowledge (DOK) level in the mathematics and ELA test design summary documents, this is insufficient to guide item writing and ensure a match between item DOK and expected DOK distributions.

Recommendation 3.3 FLDOE should document the process through which the score reports and online reporting system for various stakeholders was developed, reviewed, and incorporated usability reviews, when appropriate. Given the timing of this evaluation, the technical documentation outlining this development evidence for the FSA score reports was incomplete.

Recommendation 3.4 FLDOE should develop interpretation guides to accompany the score reports provided to stakeholders. The guides should include information that supports the appropriate interpretation of the scores for the intended uses, especially as it relates to the impact on instruction.

Study 4: Evaluation of Test Administration

Study Description

Given many of the challenges that were publicly reported regarding administration of the Florida Standards Assessments (FSA) in 2015, this study of the test administration practices contributes important information about the design and implementation of the delivery platform, as well as the potential impact on the validity of scores for students in Florida. Information was gathered from multiple sources to ensure a comprehensive review of the FSA test administration.

The study included in-person and virtual interviews with staff at FLDOE and its partner vendors to gather information that was not included in the provided documentation and to clarify evidence. The work also included a survey and focus groups to gather information directly from Florida district assessment coordinators on the nature and degree of test interruptions within the test administration. The evaluation team also identified key data and information that was required for the study and was produced by AIR. Finally, numerous other pieces of data and reports from FLDOE and AIR were also reviewed to gain greater understanding of the nature and magnitude of the test administration issues. Planned activities for this study included:

- Review of the delivery system from local education agencies to consider the following:
  - Training and testing of the system prior to the exam administration
  - Technical specifications provided for the test administration and protocols for the test administration
- Review of third-party technology and security audit reports including any stress testing performed on the system prior to the administration
- Review of test administration practices, including the following:
  - Documented student interruptions or students who encountered difficulty initially entering into the system to begin an assessment
  - Procedures that were followed when administration issues were encountered and the process followed to resolve the issues

Sources of Evidence

As part of the investigation, the evaluation team worked with FLDOE, its vendors, and directly with school districts to gain a better understanding of the spring 2015 FSA administrations. The evaluation team collected information from district representatives through three different activities:

1. the Administration Debrief Meeting held by FLDOE in Tallahassee on June 15 and 16
2. an online survey of district assessment coordinators
3. three focus group meetings with district representatives held across Florida in July

The evaluation team also reviewed a number of documents and reports that were produced by the FL DOE and their vendors. The primary documents used as part of this review included:

- FSA Test Administrator User Guide 2014-2015
- FSA ELA, Mathematics and EOC Quick Guide Spring 2015
- 2015 Test Administration Manual
- Spring 2015 FSA Training Materials PPT
- 2014-15 Test Administration and Security Agreement
- AIR Secure Browser Installation Manual 2014-2015
- AIR Technical Specifications Manual for Technical Coordinators 2014-2015
- 2014-15 Certification Process Diagram and Memo
- Letter to Pam Stewart, Commissioner of Education FLDOE from John Ruis, President FADSS
- 2015 Spring FSA Superintendent Certifications (30 school district records)
- Calculator Policy and Supporting Documents
- Monthly Emails from FLDOE to DAC

In addition, the evaluation team identified multiple data points that were needed as part of the investigation and reviewed all data produced by both FLDOE and by AIR. These reports and data included:

- Number of students active in both sessions of Reading on the same day
- Number of students who completed Reading (all sessions) in one day
- Number of students who completed Mathematics (all sessions) on the same day
- Number of students active in a single session on multiple days
- Number of students who took Writing in the second and third window
- Number of tests reopened

Each of these data files included data for schools, districts, and statewide totals. The only exception was the number of tests reopened and the number of students taking Writing in the second and third window, which provided data on a statewide basis. This evaluation also included analyses performed by AIR that focused on the consistency of trends and the potential empirical impact of the administration on test and item performance. These analyses were delivered via the technical report titled *Impact of Test Administration on FSA Test Scores*.

Study Limitations

From the onset of this evaluation, issues related to the spring administration of the FSA were already known. AIR and FLDOE communicated these issues to the evaluation team. Many of the administration issues are complex and challenging to investigate. As such, the use of a single point or source of data to capture the impact of these issues would not be appropriate,. Quantitative student data such as test scores or counts of the number of students impacted were not necessarily sufficient because they may not discernibly reflect the impact on factors like motivation and student effort. To better understand the FSA administration issues, qualitative feedback from various district representatives across the state was also collected. This evidence is essential to this evaluation because it provides information related to the series of events that occurred during the test administrations. However, this qualitative feedback also has its limitations and does not provide a measure of the impacts that these issues had on student performance and test scores.

Some of the administration-related issues that have been raised are, by their nature, not easily measured. For example, if students are unable to login to the test administration system, there is not necessarily a record of student login attempts that can be used to evaluate how commonly this type of issue was encountered. Therefore, for some noted issues, there is minimal data available to gauge the number of students impacted and the degree of impact on student scores.

Industry Standards

One of the fundamental tenants of educational assessment is that the test administration must follow consistent, standardized practices to provide all students the opportunity to demonstrate their knowledge and skills. The *Test Standards* highlight the essential role of standardization; Chapter 6 on test administration begins as follows:

> The usefulness and interpretability of test scores require that a test be administered and scored according to the test developer's instructions. When directions, testing conditions, and scoring follow the same detailed procedures for all test takers, the test is said to be standardized. Without such standardization, the accuracy and comparability of score interpretations would be reduced. (*Test Standards*, p. 111)

The usefulness and interpretability of test scores require that a test be administered and scored according to the test developer's instructions. (Test Standards, p. 111)

For most educational assessments, the ability to make the intended inferences and comparisons is directly tied to the standardization of the test administration. For example, standardized, controlled conditions are required to compare student performance across students, teachers, schools, districts, and years.

Cohen and Wollack (2006) also discuss the importance of standardization in test administration by stressing that the standardization requirement is not met merely because students have received the same set of items, the same type of items, or scores on the same scale. Instead, "tests are standardized when the directions, conditions of administration, and scoring are clearly defined and fixed for all examinees, administrations, and forms" (p. 358).

A number of specific *Test Standards* address appropriate test administration procedures and their importance to the reliability, validity, and fairness of the tests. Standard 6.1 discusses the importance of test administration practices and that the test administration should "follow carefully the standardized procedures for administration …" (*Test Standards*, p. 114). This standard also stresses the need for appropriate training for all individuals involved with the administration to allow them to understand the value and importance of their role in the test administration.

Standard 6.3 focuses on the requirements for testing programs when any deviation from the standardized procedures are encountered by stating that “changes or disruptions to standardized test administration procedures or scoring should be documented and reported to the test user” (*Test Standards*, p. 115).

In addition to discussing the value and importance of administration practices and standardization of these practices, the *Test Standards* also focus on the need to develop a system that quickly and efficiently deals with any test administration difficulties that may arise. In Chapter 12, which focuses on educational assessment, the *Test Standards* state that “test developers have an obligation to support the test administration process and to provide resources to help solve problems when they arise” (*Test Standards*, p. 192).

The purpose of highlighting the relevant *Test Standards* at the outset of our discussion of this study is to emphasize that the standardization of test administration conditions is a prerequisite for subsequent data analyses and interpretation of scores. Deviations from the intended standardized conditions and environment can impact the comparability and interpretability of scores. Per the *Test Standards,* test administration issues must be addressed immediately to resolve the issue and investigate the impact of the issue on the scores and their uses.

Florida Standards Assessments Processes and Evaluation Activities

District Data Collection

As mentioned previously, the evaluation team used a combination of data and information collected directly from Florida district representatives and data and information from FLDOE and AIR to reach the most comprehensive understanding of the FSA administration as possible.

FLDOE invited members of the evaluation team to attend the Administration Debrief Meeting. Thirteen districts were represented at the meeting; district assessment coordinators provided feedback to FLDOE and testing vendors regarding the challenges and accomplishments of the 2014-15 administrations. This meeting provided valuable information and insight into the test administration difficulties that Florida schools and districts encountered. It also highlighted a number of critical areas where further information is needed.

After this meeting, the evaluation team developed a questionnaire; on July 1, 2015, this questionnaire was distributed via an email survey to district assessment coordinators or representatives from every district in the state. The survey closed on July 20; at that time, data were available from 55 respondents who represented 48 of the 76 Florida districts. Complete data on the survey and the responses received can be found in Appendix C.

In addition to the survey, three focus groups were held in Florida; these focus groups provided district representatives with the opportunity to share their experiences and to allow the evaluation team to ask follow-up questions and ensure accurate understanding of the events related to the test administrations. The focus group meetings were held on July 15 and 16 at schools within each of the following districts: Leon County, Miami-Dade County, and Orange County. District assessment coordinators or similar representatives from every district in Florida were invited to attend the meeting, but participation was limited to two representatives for each district. Across the three focus group meetings, a total of 56 participants from 33 districts attended the focus groups. Appendix D provides a complete listing of the data collected across these three focus group meetings.

Table 12 provides a summary of the districts from which the evaluation team received feedback regarding the FSA administrations. Between the Administration Debrief Meeting, the online survey, and the three focus group meetings, 53 of 76 districts (69.7%) provided input and data that were used for this evaluation.

53 of 76 districts (69.7%) provided input and data that were used for this evaluation.

Table 12: District representation across study-related activities

District		Study Participation		
Number	District Name	Debrief	Survey	Focus Group
1	ALACHUA			
2	BAKER		x	
3	BAY		x	x
4	BRADFORD		x	
5	BREVARD			x
6	BROWARD	x	x	x
7	CALHOUN		x	
8	CHARLOTTE			
9	CITRUS		x	x
10	CLAY			
11	COLLIER		x	
12	COLUMBIA			
13	MIAMI DADE	x	x	x
14	DESOTO		x	x
15	DIXIE		x	
16	DUVAL			
17	ESCAMBIA		x	x
18	FLAGLER			
19	FRANKLIN			
20	GADSDEN		x	x
21	GILCHRIST	x	x	
22	GLADES			
23	GULF			
24	HAMILTON		x	x
25	HARDEE			
26	HENDRY			
27	HERNANDO		x	
28	HIGHLANDS		x	
29	HILLSBOROUGH	x	x	x
30	HOLMES		x	
31	INDIAN RIVER			
32	JACKSON			
33	JEFFERSON		x	
34	LAFAYETTE		x	
35	LAKE	x	x	x
36	LEE	x	x	
37	LEON		x	x
38	LEVY		x	
39	LIBERTY		x	

District		Study Participation		
40	MADISON		x	
41	MANATEE		x	X
42	MARION		x	X
43	MARTIN		x	X
44	MONROE			
45	NASSAU			x
46	OKALOOSA		x	x
47	OKEECHOBEE		x	x
48	ORANGE	x	x	x
49	OSCEOLA			x
50	PALM BEACH	x	x	x
51	PASCO		x	x
52	PINELLAS		x	x
53	POLK		x	x
54	PUTNAM		x	
55	ST JOHNS			x
56	ST LUCIE	x	x	x
57	SANTA ROSA		x	x
58	SARASOTA		x	
59	SEMINOLE	x	x	x
60	SUMTER		x	x
61	SUWANNEE		x	x
62	TAYLOR			
63	UNION			
64	VOLUSIA	x	x	x
65	WAKULLA	x		
66	WALTON			
67	WASHINGTON	x	x	
68	FSDB		x	x
69	WCSP			
71	FL VIRTUAL		x	X
72	FAU LAB SCH			
73	FSU LAB SCH			
74	FAMU LAB SCH			
75	UF LAB SCH		x	
80	STATE COLLEGES			
98	AHFACHKEE SCHOOL			

Feedback from districts was used along with the documentation provided by FLDOE and its vendors, information collected during meeting and interviews with FLDOE and vendor staff, as well as various analyses provided by AIR related to the impact of the various administration issues investigated.

Test Administration Investigation by Test

In the remainder of this section, a number of issues or concerns that have been raised in regards to the FSA test administration are reviewed. The three primary issues that were encountered within each of the three content areas (Writing, Reading, and Math) are discussed first. District administrators identified each of these issues as the biggest challenge they faced this past year. While the Writing and Reading tests are combined for scoring and reporting of the English Language Arts (ELA) FSAs, the tests are administered in distinct sessions and are therefore addressed separately here. After reviewing the issues for Writing, Reading, and Math, the remaining sections outline additional issues that were encountered, some of which impacted all FSA administrations, others of which were relevant for specific tests. For each issue, after the nature of the issue is described, available evidence that describes the extent and nature of the issue is discussed.

Writing

Description of Administration Challenges. The FSA Writing test was comprised of one session; students were required to review multiple sources of evidence about a single topic. After reviewing the materials, students were required to respond to a prompt by organizing and providing information to support their opinion on the topic. For grades 4 to 7, the test was administered via a paper-and-pencil model (PP); for grades 8 to 10, a computer-based testing (CBT) modality was used.

Across the Administration Debrief Meeting, the online survey, and the focus groups, only minor issues related to materials distribution were noted regarding the PP-based Writing tests in grades 4 through 7. District assessment coordinators noted that these materials issues caused inconveniences; however, these inconveniences were manageable, typical of issues encountered during statewide assessment administrations, and not impactful for students.

For the CBT administrations in grades 8 to 10, considerably more reports of difficulty occurred with the test administration. The issues with the Writing test centered around two distinct issues. First, many schools reported that their students were unable to login to the testing system. Second, students appeared to be kicked out of the testing system without explanation, and possibly lost some of their work when it occurred.

Students were unable to login to the system because of two different problems. First, the login system had difficulties due to changes in the student database. Therefore, some students were unable to login at the time they were scheduled during the first two days of the testing window.

The problems on these two days were followed by a Distributed Denial of Services (DDoS) attack that occurred on Thursday, March 5 (DDoS attacks also occurred on March 2nd and 3rd, but were likely masked by the login difficulties that were encountered). The login issues and the DDoS attacks had much the same effect from the schools' perspectives; some students were unable to login to the system and begin their testing session. The extent of these problems is difficult to estimate because the AIR online delivery system only tracks activity after login. Data that might suggest ongoing challenges like multiple failed login attempts are not recorded.

The second issue for the CBT writing administrations related to students being removed from the testing system and in some cases losing work not saved in the last two minutes as a result. AIR explained that this issue primarily resulted from system settings related to an inactivity timer. While FLDOE and district test administrators were aware that an inactivity timer was in place for each test session that a test administrator created, they were not made aware that another inactivity timer, that monitored the activity of individual students, was also in place. This timer removed students from the testing system after 60 minutes of inactivity. After this time elapsed, students were inactive in the system. The student was not alerted to this condition until the next time the system tried to automatically save the student work, which happened every two minutes. Therefore, work completed after this 60 minutes of inactivity could have been lost. Some of the students who were timed out were unable to return immediately to their work, and needed to return either later that day or on subsequent days to finish their test.

Evidence. To investigate and better understand the various issues that occurred during the FSA writing administrations, the evaluation team sought both quantitative and qualitative information related to the prevalence of the issues and the type and degree of impact that they may or may not have had on student test scores. These data came from two sources: (1) both quantitative and qualitative feedback from district assessment coordinators and other representatives and (2) from AIR based on information compiled within their testing system.

Within the online survey of district assessment coordinators, several questions addressed the issues encountered during the FSA writing administration. Of the 55 survey responses, 94% indicated that their district experienced some type of technology issue during the administration of the CBT Writing tests. Of those impacted, 81% reported that students experienced difficulties logging into the system and 77% reported that some number of students lost work.

District assessment coordinators were also asked to estimate the percentage of students in their district that were impacted by the technology issues for the Writing test. As shown in Figure 7, 13 of the 53 respondents, or approximately 25%, estimated that 1-9% of students within their district were impacted by technology issues on the Writing FSAs while 12 respondents, or about 23%, estimated that 10-19% of students were impacted. Almost half of the respondents (27 of 53) estimated that 20% or more of the students in their district were impacted by the writing technology issues.

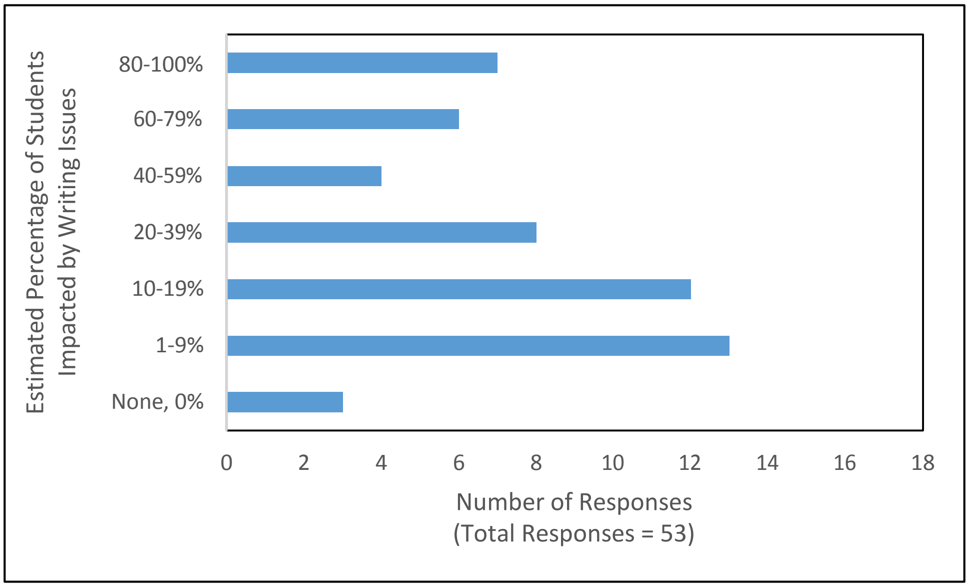


Figure 7: District Representatives' Estimated Percentage of Students Impacted by Writing Technology Issues

Based on the issues experienced, 38% of respondents reported that technology difficulties had a major impact on the Writing test administration, 36% characterized the impact as moderate, and 6% of respondents reported that the issues had no impact. All online survey data, including the data related to the writing administration, can be found in Appendix C.

Data from both the Administration Debrief Meeting and the three focus group meetings aligned with the data provided through the online survey. Preliminary survey data (i.e., responses received through July 13) were available for the focus group meetings; the evaluation team shared the initial findings with the focus groups and asked the district representatives to respond to the accuracy of the survey data and provide more details about their experiences with the Writing test administrations. At the focus group meetings, the district representatives provided additional information about the activities that occurred just prior to students losing work as well as the process and experiences for recovering student work. District representatives also emphasized the severity of issues related to students losing work, regardless of the number of students impacted. Finally, the district representatives also discussed and shared experiences related to the impact that the various system issues had both directly and indirectly on the student testing experience (e.g., students who experienced noisy and disruptive testing environments even when the individual student was not directly impacted by a testing issue).

In addition to the various sources of information from district representatives, AIR provided quantitative data to estimate the magnitude of the impact of the CBT writing administration issues on Florida students. AIR reported approximately 600 documented cases of students losing work on the Writing test across grades 8-10.

AIR also provided the evaluation team with data that summarized the number of students, by test, that were logged into the same test session on multiple days. This data provides insight into the magnitude of the problem of students being logged out of the system, being unable to log back in, and having to complete testing on a later date. As can be seen in Table 13, the number of students who were in the same test session across multiple days was less than 1% of the student population in each of the three grades.

Table 13. State-Level Occurrence of Students in the Writing Session on Multiple Days

Writing	Total Students Tested (Statewide)*	Students in Session on Multiple Days	
		Number	Percent of Total
Grade 8	201,700	678	0.33%
Grade 9	207,092	563	0.27%
Grade 10	197,072	456	0.23%

*These values are estimates based on data provided by AIR and do not represent final counts of students completing the test

In addition to reviewing this data at the state level, the information was also disaggregated to the school level and combined with estimates for the number of students who completed Writing at each school. It is important to note here that this data should not be treated as official state-certified data; instead, these data represent the estimates from the evaluation team to understand how the impact was felt at the school level. Aggregated to the school level, at least 1 student in approximately 17% to 19% of schools had students who had to test over more than one day to complete the Writing test. Within the schools that had at least one student impacted, the percent of students impacted was estimated to be between 1% and 2% as shown in Table 14.

Table 14. School-Level Occurrences of Students in the Writing Session on Multiple Days

Writing	Total Schools Administered Assessment	Schools with Students in Same Session on Multiple Days		
		Number	Percent of Total	Average Percent of Students Within School Impacted
Grade 8	1,303	226	17.34%	2.14%
Grade 9	992	180	18.14%	1.09%
Grade 10	921	175	19.00%	0.91%

In addition to data on the number of Florida students impacted, AIR conducted an analysis that was designed to determine if shifts in trends could be observed with this year's FSA results. The FSA score stability analysis first gathered the correlation between students' FCAT 2.0 Reading scores in 2012-13 and 2013-14. Correlations are statistical values that range from -1.0 to 1.0, and the statistic represents an estimate for how closely related two different set of number are. When you have two sets, and the numbers increase in approximately same fashion, the correlation between those two data sets will have a strong positive correlation. Values above 1.75 represent strong positive correlations between the test scores.

These correlations were calculated by gathering the same students' scores over two years. For every student included, their test scores from two consecutive years were gathered. For example, the data could have been from students who took Reading FCAT 2.0 in 5th grade in 2012-13, and the Reading FCAT 2.0 in 6th grade in 2013-14. For all of the data that linked the 2012-13 to the 2013-14 academic year, the correlations represent the baseline correlation values presented in Table 15. These values represent the relationship between students' scores across the two years.

After gathering these values for the baseline correlations, the same calculations were completed but using data from the 2013-14 Reading FCAT 2.0 and the 2014-15 FSA English Language Arts test score. These correlation values represent the current values provided in Table 15. The baseline and current correlations are nearly the same indicating that the relationship between students' scores from one year to the next was no different from 2013-14 to 2014-15 than those seen from 2012-13 to 2013-14. Issues encountered with the FSA Writing administrations in 2014-15 did not impact this relationship at the state level.

Table 15: Comparison of baseline and current correlations between two years' test scores in English Language Arts

Test	Baseline*	Current**
Grade 4 ELA test score to Grade 5 ELA test score	0.80	0.80
Grade 5 ELA test score to Grade 6 ELA test score	0.82	0.82
Grade 6 ELA test score to Grade 7 ELA test score	0.81	0.82
Grade 7 ELA test score to Grade 8 ELA Test Score	0.82	0.82
Grade 8 ELA test score to Grade 9 ELA test score	0.83	0.83
Grade 9 ELA test score to Grade 10 ELA test score	0.82	0.82

*** Baseline correlations were calculated between 2012-13 and 2013-14 test scores**

**** Current correlations were calculated between 2013-14 and 2014-15 test scores**

Reading

Description of Administration Challenges. For Reading, grades 3 and 4 FSAs were administered PP while grades 5 to 10 were administered via CBT. As with the Writing test, the PP test administrations did not cause significant issues with their test administration. In general, test administrators were able to complete the test administrations in a timely manner and without serious complications.

The CBT exams for Reading included two sessions; students were scheduled to complete one session on their first day and the second session on a following day. Students who completed session 1 should not have entered into session 2 until the next day, and students should have been restricted from access to session 2 unless they received approval from the test administrators to move forward. For Reading, the primary concern that was raised focused on this student transition from session 1 to session 2.

The student movement across testing sessions appears to have occurred for a number of different reasons. Some students had not yet finished session 1, but were merely scanning forward in the test form, and did not realize that they had entered into session 2. Other students had completed session 1 and moved forward unaware that they were entering into session 2. Once students entered into session 2, they were unable to go back to session 1. They needed to close out of their testing session and request it to be reopened through the test administration management system. This led to some serious administration delays because this reopening of tests required the involvement of the district assessment coordinator and AIR as well as FLDOE approval, actions that in some cases took several days to complete before the student could resume testing.

Evidence. The review of the Reading test administration began with the development and analysis of the survey results, as well as the information collected during the focus group meetings. On the survey, 91% of the respondents indicated that their district had experienced some type of technology issue associated with the Reading test. Of the respondents, 77% indicated that some students had difficulty logging into the system, and 83% indicated that some students were inadvertently logged out while completing the test.

District assessment coordinators were also asked to estimate the percentage of students in their district that were impacted by the technology issues for the Reading test. As shown in Figure 8, 13 of the 53 respondents, or approximately 25%, estimated that 1-9% of students within their district were impacted by technology issues on the Reading FSAs while 9 respondents, or approximately 17%, estimated that 10-19% of students were impacted. Approximately half of the respondents (27 of 53) estimated that 20% or more of the students in their district were impacted by the Reading technology issues.

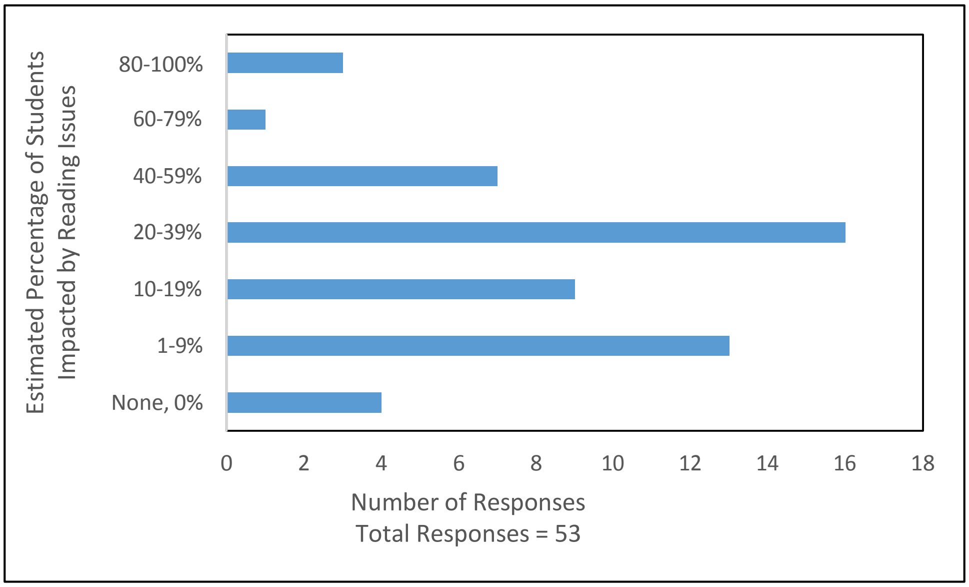


Figure 8: District Representatives' Estimated Percentage of Students Impacted by Reading Technology Issues

Based on the issues experienced, 25% of respondents reported that technology difficulties had a major impact on the Reading test administration, 47% characterized the impact as moderate, and 8% of respondents reported that the issues had no impact. All online survey data, including the data related to the Reading administrations, can be found in Appendix C.

During the focus group meetings, the district representatives described problems and issues that were consistent with the data from the survey. The problem with students entering session 2 was described by many of the focus group participants. Some participants said that after students inadvertently entered session 2 and had that session closed, students could not get back to session 1 to complete testing for that session on the same day.

In addition to the survey and focus group information, the evaluation team also identified other data that would be needed to estimate the magnitude of the empirical impact of these issues to the evaluation team. As with Writing, the first point of data summarized the number of students who completed a single test session on more than one day. As can be seen in Table 16, less than 1% of students in each grade had records of completing the same session on different days.

Table 16. State-Level Occurrence of Students in a Reading Session on Multiple Days

Reading	Total Students Tested (Statewide)*	Students in Session on Multiple Days	
		Number	Percent of Total
Grade 5	196,759	493	0.25%
Grade 6	195,746	1,296	0.66%
Grade 7	195,531	715	0.37%
Grade 8	201,348	625	0.31%
Grade 9	205,531	1,203	0.59%
Grade 10	194,985	666	0.34%

*These values are estimates based on data provided by AIR and do not represent final counts of students completing the test.

In addition to reviewing this data at the state level, the information was also disaggregated to the school level and combined with estimates for the number of students who completed Reading at each school. It is important to note here that this data should not be treated as official state-certified data; instead, these data represent the estimates from the evaluation team to understand how the impact was felt at the school level. Aggregated to the school level, at least 1 student in approximately 8% to 19% of schools had students who had to test over multiple days to complete a session for Reading. Within the schools that had at least one student impacted, the percent of students impacted was estimated to be between 3% and 6% as shown in Table 17.

Table 17. School-Level Occurrences of Students in a Reading Session on Multiple Days

Reading	Total Schools Administered Assessment	Schools with Students in Same Session on Multiple Days		
		Number	Percent of Total	Average Percent of Students Within School Impacted
Grade 5	2,233	180	8.06%	3.69%
Grade 6	1,301	215	16.53%	3.81%
Grade 7	1,237	150	11.96%	3.37%
Grade 8	1,303	138	12.13%	5.27%
Grade 9	992	192	19.35%	3.63%
Grade 10	921	159	17.26%	3.13%

The issue of students advancing test sessions earlier than intended is not unique to the 2015 FSA. This issue began prior to CBT delivery when students could move forward in PP test booklets without the permission or knowledge of the test administrator. FLDOE policy for students who enter into session 2 has been that once students enter into the second session, students must complete both sessions on that day. This policy was the intended policy again in 2015.

To help investigate student movement across test sessions, AIR provided two data points that focused on students who were active within both session 1 and 2 for Reading on the same day. All data was provided at the state, district, school, and test level. The first data point provided the number of students that were active within both sessions on the same day. The second data point was the number of students who completed both sessions on the same day per the administration policy.

As can be seen in Table 18, at the state level, between 2,079 and 5,138 students per grade level were active in both Reading sessions on the same day, which represents between 1% and 2% of students who completed each test. Across grades, between 41% and 60% of those students proceeded to finish their exam on that day.

Table 18. State-level Occurrence of Students Moving Across Sessions in Reading

		Students in Two Sessions on Same Day		Students Completing Two Session on Same Day	
Reading	Total Students Tested (Statewide)*	Number	Percent (of Total)	Number	Percent (of Students in Two Sessions)
Grade 5	196,759	2,079	1.05%	861	41.41%
Grade 6	195,746	4,328	2.21%	1,869	43.18%
Grade 7	195,531	3,301	1.69%	2,003	60.68%
Grade 8	201,348	3,258	1.62%	1,827	56.08%
Grade 9	205,531	5,138	2.50%	2,475	48.17%
Grade 10	194,985	4,123	2.11%	2,503	60.71%

*These values are estimates based on data provided by AIR and do not represent final counts of students completing the test.

At the school level, as can be seen in Table 19, between 35% and 53% of schools had at least one student impacted by the student movement across sessions. Within the schools impacted, between 5% and 15% of the students within the school appear to have had some issues with movement into session 2.

Table 19. School-Level Occurrence of Students Moving Across Sessions in Reading

		Schools with Students in Two Sessions on Same Day		
Reading	Total Schools Administered Assessment	Number	Percent of Total	Average Percent of Students Within School Impacted
Grade 5	2,233	800	35.82%	5.50%
Grade 6	1,301	677	52.03%	8.20%
Grade 7	1,237	577	46.64%	8.80%
Grade 8	1,303	572	43.90%	12.70%
Grade 9	992	520	52.42%	14.50%
Grade 10	921	490	53.20%	13.10%

As with the Writing test, the data provided by AIR designed to look at the correlation between last year's FCAT to this year's score is also applicable here. The ELA scores used in the analysis of the Writing test above uses student performance on both the Reading and Writing tests. As such, the stability of score correlations supports the concept of little to no change in the correlations being observed this year.

A regression analysis was also completed that focused on the test scores of students who mistakenly moved into session 2. A regression analysis is another way to estimate the relationship between two sets of variables. In this scenario, the 2013-14 FCAT 2.0 test scores can be used to predict student performance on the FSA. For this evaluation, two different groups were created; the first with all students who moved into session 2, and the other group all students who did not. Separate regression analyses were performed for the two groups across all grade levels. For 5 of the six grade levels, the prediction equation was the same across the two groups. For the one group that was different, it indicated student scores were slightly lower than predicted by the FCAT score.

AIR also completed work focused on the calibration of item response theory (IRT) item parameters. In the scaling of the FSA, one of the initial steps completed after screening the test data is to calibrate all items on the FSA. This process of calibrating the items produces item statistics for every item. Using the item statistics, a test characteristic curve (TCC) can be calculated. A test characteristic curve can be used to illustrate the relationship between the ability estimate for students, theta, and the proportion of items the students got correct. In the graph below, the percentage of items that a student got correct on the test is represented on the Y-axis, and labeled as TCC Proportion. The X-axis on the graph below represents the estimated score for students, *theta*, ranging from approximately -5 to 5, with -5 representing the lowest estimate and 5 representing the highest possible estimate. The Y-axis in Figure 9, *TCC Proportion*, represent the percent of items scored correctly on the exam.

In the analysis, the item parameters and TCC were calculated for all items using the complete sample of students used in the item calibration, including those students who appeared to have been impacted by these administration-related difficulties described in the sections on Writing and Reading. The calculation of item parameters was then repeated, excluding those students who were impacted. To illustrate these findings, the TCC for the Grade 10 ELA test is provided in Figure 9; the two curves almost perfectly overlap with one another. The same analyses were completed across all of the tests that comprise the FSA and consistent results were observed. These data provide evidence that the scores of students who were impacted by issues on the CBT administrations of Writing and Reading did not significantly affect the statistics of the FSA items and tests at the state level of analysis.

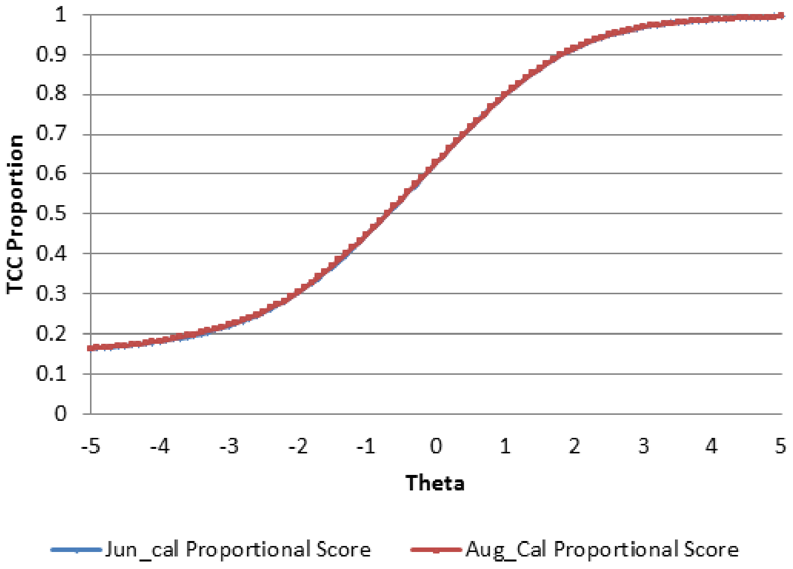


Figure 9. Test characteristic curve for Grade 10 ELA Florida Standards Assessments, with impacted students included and with impacted students removed.

Mathematics

Descriptions of Administration Challenges. The administration of the FSA Math test closely paralleled the Reading test administration model. Grades 3 and 4 were administered via PP. Grades 5 to 8, along with three end-of-course (EOC) tests, Algebra 1, Algebra 2, and Geometry, were CBT. One important distinction between the two is that Math FSAs grades 6 to 8 had three test sessions, whereas Reading had only two sessions. All other Math assessments also had only two sessions.

As with the other assessments, the PP test administrations were completed and delivered without much difficulty. Serious concerns were not raised about these administrations; test administrators were generally satisfied with the administration. For the CBT administrations, the difficulties described in moving across sessions were also encountered on the Math FSAs.

Evidence. The review of the Math test administration began with the development and analysis of the survey results, as well as the information collected during the focus group meetings. Approximately 91% of survey respondents indicated that they experienced some type of technology issue associated with the Math test. Of the respondents, 65% indicated that some students had difficulty logging into the system, and 75% indicated that some students were inadvertently logged out while completing the test.

District assessment coordinators were also asked to estimate the percentage of students in their district that were impacted by the technology issues for the Math test. As shown in Figure 10, 17 of the 52 respondents, or approximately 33%, estimated that 1-9% of students within their district were impacted by technology issues on the Math FSAs while 7 respondents, or approximately 13%, estimated that 10-19% of students were impacted. Approximately 44% of the respondents (23 of 52) estimated that 20% or more of the students in their district were impacted by the Math technology issues.

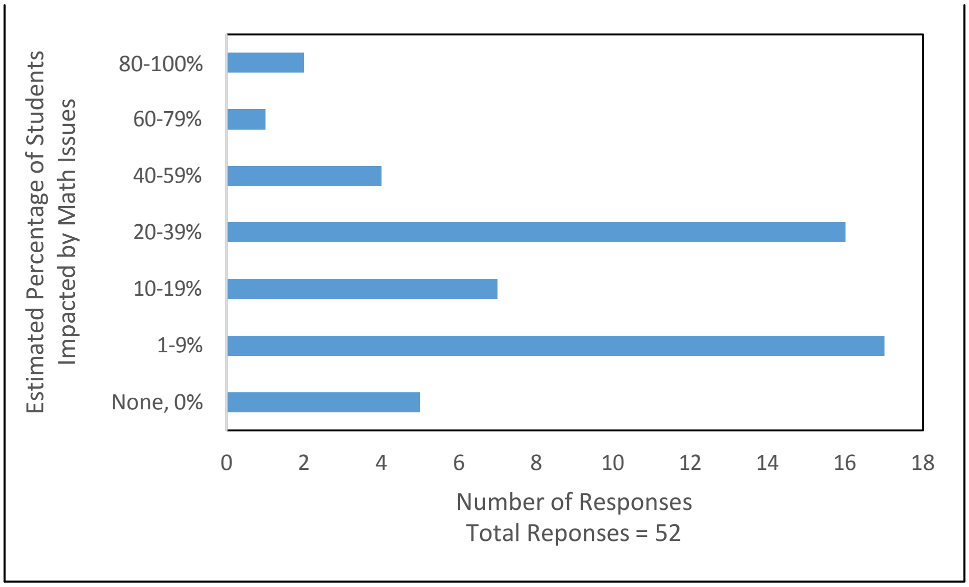


Figure 10: District Representatives' Estimated Percentage of Students Impacted by Math Technology Issues

Based on the issues experienced, 10% of respondents reported that technology difficulties had a major impact on the Math test administration, 48% characterized the impact as moderate, and 10% of respondents reported that the issues had no impact.

During the in-person focus groups, the test administrators described problems and issues that were consistent with the survey data. The problem with students moving into sessions 2 and 3 was described at length. As with other areas, the test administrators also raised the concern that the impact was felt by the students who were directly impacted as well as those students in the same classroom as administrators and other support staff needed to be in the testing room to resolve the various technology issues.

The number of students who appeared in the same session across multiple days was calculated. At the state level, as can be seen in Table 20, for almost every assessment, the percentage of students impacted was less than 1%. For Algebra 1, the number was closer to 2%.

Table 20. State-Level Occurrence of Students in a Math Session on Multiple Days

		Students in Session on Multiple Days	
Math	Total Students Tested (Statewide)*	Number	Percent of Total
Grade 5	196,970	457	0.23%
Grade 6	191,189	519	0.27%
Grade 7	179,595	557	0.31%
Grade 8	124,981	625	0.50%
Algebra 1	206,305	91	0.04%
Algebra 2	161,454	240	0.15%
Geometry	198,102	202	0.10%

*These values are estimates based on data provided by AIR and do not represent final counts of students completing the test.

Across schools, for grades 5 to 8, approximately 4% to 11% of schools had at least one student in the same session across multiple days. Within the schools impacted, between 3% and 7% of students appeared to have been in the same session on multiple days.

One important caveat regarding the EOC data should be noted. Data were compiled for the number of students at each school that took the various Math FSAs. This data served as baseline data, allowing the evaluation team to estimate the percentage of students in a given school that were impacted by any of the test administration issues. In the original extraction of data for the Math tests, data for the EOC exams were only pulled for one grade level, which underestimated the number of schools that administered the EOC exams and the number of students impacted within those schools. Because of this issue, accurate estimates for the percent of school impacted as well as the percent of students within schools is not available at this time for the three EOCs.

Table 21. School-Level Occurrences of Students in a Math Session on Multiple Days

		Schools with Students in Same Session on Multiple Days		
Reading	Total Schools Administered Assessment	Number	Percent of Total	Average Percent of Students Within School Impacted
Grade 5	2,229	94	4.17%	7.06%
Grade 6	1,322	130	9.76%	3.16%
Grade 7	1,230	132	10.57%	4.45%
Grade 8	1,209	87	7.20%	7.54%

The second data point that was investigated for the Math assessment was the number of students who completed all sessions of the Math FSA in one day. As a reminder, in Math, grades 6 to 8 are comprised of three sections, while all other grades and the EOC tests are comprised of two sessions. For grades 6 to 8, many schools scheduled testing to include the completion of two Math sessions on the same day. Therefore the completion of two sessions on the same day for Math in these grades is not indicative of an administration issue. Rather student activity in three sessions in one day would indicate an issue related to unintended movement across sessions. As can be seen in Table 22, across the entire state, less than 1% of students completed all Math sessions in one day for grades 5 to 8. The number does increase fairly dramatically for the EOC tests, ranging from 3% for Algebra 1 to 19% on Algebra 2.

Table 22: Number of students who completed all Math sessions in one day

Math	Total Students Tested (Statewide)*	Completed all sessions, 1 day Number of students who completed in 1 day	 Average Percent of Students within School Impacted
Grade 5	196,970	534	0.27%
Grade 6	191,189	921	0.48%
Grade 7	179,595	1,130	0.63%
Grade 8	124,981	1,352	0.67%
Algebra 1	206,305	2,628	1.27%
Algebra 2	161,454	2,135	1.32%
Geometry	198,102	2,490	1.26%

When looking at the percentage of schools with at least one student impacted, the same issue that was described above with the EOC exams data prevents us from providing accurate numbers for the percent of schools or the percent of students with schools for the EOC exams (see Table 23). For grades 5 to 8, a fairly wide range was observed; with 13% of schools had students who completed Math in one day in Grade 5, and approximately 30% of schools had at least one student impacted on the Grade 8 exam. Looking closer at the school level data, because of problems with the merging of multiple datasets, accurate estimates for the percentage of students within schools could not be calculated for the EOC exams. For grades 5 to 8, the percentage of students within the schools ranged from 5% to 13% impacted.

Table 23: Number of schools with students who completed all Math sessions in one day

		Schools with Students Who Completed Math Session in One Day		
Math	**Total Schools Administered Assessment**	**Number**	**Percent of Total**	**Average Percent of Students Within School Impacted**
Grade 5	2,229	297	13.32%	5.20%
Grade 6	1,322	283	21.41%	7.80%
Grade 7	1,230	331	26.91%	8.80%
Grade 8	1,209	368	30.44%	13.40%

AIR also completed IRT calibration analysis analyses as has already been described with the Writing and Reading assessments. The IRT parameters and the TCC were calculated using the total group of students, and then recalculated after the impacted students were removed. As with Reading and Writing, little to no difference in the IRT parameters was observed.

As with the Reading test, a regression analysis was also completed that focused on the test scores of students who mistakenly moved into session 2. Using last year's FCAT 2.0 Math score, a regression analysis was completed that used FCAT 2.0 Math test scores to predict the FSA Math scores for students. It also classified students into two groups; one group that did not mistakenly move into the second session, while the other group did mistakenly move into session 2. In this scenario, if students moved into session 2 and by being able to preview items were given some type of advantage, the regression equation between the two groups would be different. The regression analyses were completed for grades 5 to 8 on the Math FSA. For three of the four grades, the prediction equation was the same across the two groups. For the one group that was different, it indicated student scores were slightly lower than predicted by the FCAT score.

In addition to data on the number of Florida students impacted, AIR conducted an analysis that was designed to determine if shifts in trends could be observed with this year's FSA results. This was identical to the analyses described in the Writing section of this report using correlations of the same students' scores over two years. For every student included, their test scores from two consecutive years was gathered. For example, the data could have been from students who took FCAT 2.0 in 5th grade in 2012-13, and the FCAT 2.0 in 6th grade in 2013-14. For all of the data that linked the 2012-13 to the 2013-14 academic year, the correlations represent the *baseline* correlation values presented in Table 24. These values represent the relationship between students' scores across the two years.

After gathering these values for the baseline correlations, the same calculations were completed but using data from the 2013-14 FCAT 2.0 the 2014-15 FSA. These correlation values represent the *current* values provided in Table 24. The baseline and current correlations are very similar indicating that the relationship between students' scores from one year to the next was no different from 2013-14 to 2014-15 than those seen from 2012-13 to 2013-14. Issues encountered with the FSA Math administrations in 2014-15 did not impact this relationship at the state level.

Table 24: Comparison of baseline and current correlations between two years' test scores in Math

Test	Baseline*	Current**
Grade 4 Math test score to Grade 5 Math test score	0.76	0.79
Grade 5 Math test score to Grade 6 Math test score	0.79	0.82
Grade 6 Math test score to Grade 7 Math test score	0.80	0.82
Grade 7 Math test score to Grade 8 Math Test Score	0.74	0.71

* Baseline correlations were calculated between 2012-13 and 2013-14 test scores

** Current correlations were calculated between 2013-14 and 2014-15 test scores

Other Test Administration Issues Identified During the Investigation

In addition to the three issues described previously, a number of other issues were also identified; some of these issues were specific to one test, and other issues impacted the overall FSA administration.

External Technology Challenges

Description of Administration Challenges. Another issue that was encountered across the state of Florida was a number of Distributed Denial of Services (DDoS) attacks on the FSA delivery system. These are malicious attempts to interfere with technology or network availability during examination administrations. DDoS attacks were observed on the FSA delivery system on March 1, 2, 3, 5, 9, 11, and 12. As March 1 was the Sunday prior to the administration window, this DDoS attack did not impact students. The DDoS attacks on March 2 and 3 were likely masked to test users by the number of login issues that were encountered with the FSA system and therefore likely did not cause significant delays beyond those already being experienced. In comparison, the DDoS attacks observed on March 5 did receive a considerable amount of attention and did appear to cause some disruption of test delivery in schools. After some modifications were made to the security and monitoring of the system, the DDoS attacks March 9, 11, and 12 did not appear to cause any significant problems.

The DDoS attacks were designed to flood the FSA test delivery system which, in effect, caused the system to become so crowded with the handling of the DDoS-related traffic, that legitimate traffic (i.e., traffic from schools) was unable to properly connect with the testing log in system. The result for the end user was an inability to log into the FSA testing system. Not all students who attempted to login during a DDoS attack were denied access to the FSA delivery system, but a significant number of students were blocked from doing so. One fortunate characteristic of the FSA DDoS attacks is that once students were able to enter into the FSA testing system, they were able to complete the test in the manner intended.

Evidence. As with many components of this investigation, it is difficult to gauge the number of students impacted by the DDoS attacks as well as the degree of impact on students' testing experience. For example, the manner in which FSA registration is handled does not allow for an accurate estimate for the number of students who were scheduled to test on a given day. There are records for the total number of students who were registered to take a specific FSA, but this information does not reflect or include the day on which the tests were planned to be taken. Because of this limitation, it is not feasible to develop a reasonable estimate for the percentage of students, on any given day, that were scheduled to take a given test, but were unable to do so because of login system-related issues.

Another limitation is that the FSA login system does not track login attempts. Because of this limitation, we cannot compare the number of login attempts that occurred on any given day, and how many login attempts students needed to complete before they were successful.

One piece of evidence that can be compared is the number of users who accessed the system, on each day. A report on the number of users of the FSA delivery system throughout each day of the test administration window is included in Appendix E. The report provides a snapshot of the number of users every 30 minutes during the regular time period for the test administration for each date. For example, at 9:00 am on Monday 2, there were 29,779 users in the FSA system. While this data does not provide a perfect snapshot of the number of tests that were completed on each day, it does provide a general estimate for the amount of system activity each day.

In addition to looking at the overall level of activity, the maximum level of activity on each day can be determined. In Table 25, the maximum number of users for each day of the FSA test administration is provided which represents the peak number of students testing concurrently for each day. The days with reported DDoS attacks are highlighted in the table. Looking closer at the data, while there were reports of system disruption on these days, it does not appear to have had an impact on the maximum number of users on those days. The maximum number of users does decline when looking at March 11 and 12, but that appears to be a function of the Writing test administration window coming to a close. Also, it is worth noting that the number of users is less for the tests days from March 2 through March 13 as the only tests included in this window were Writing grades 8-10. In comparison, many more tests were being administered during the April and May dates and the Max Users values reflect this difference.

Looking at the overall trends that are included in Appendix E, a similar pattern is observed. Looking at the first week, there were three days that had reported DDoS attacks: the 2nd, 3rd, and 5th. On each of those days, despite the DDoS attacks, the amount of system-wide activities does not seem to have dramatically altered from the pattern of system use. The same pattern can be observed in the following week, when documented DDoS attacks occurred on March 9, 11, and 12. For each of those days, the documented activity observed within the FSA delivery system appears to be consistent with the pattern observed across the entire test administration window. For example, across all days during the week of March 2, peak activity appears to occur in the 9:30 to 10:30 range, with activity slowly decreasing for the remainder of the day. It also appears that Mondays are consistently one of the slower days, as many people report that schools prefer to allow students to test in the middle of the week.

It should also be noted here that on April 20, an issue with students being able to login to the system was encountered. The practical impact of these difficulties was fairly similar to the DDoS attacks, as students had difficulty logging into the system, though once they were able to do so, most were able to complete their test without any further difficulty. This issue did cause a decrease in the number of students who tested that day as can be seen in Table 25 as well as in the overall activity that day as can be seen in Appendix H. However, the login difficulties were not the result of a DDoS attack, but instead were the result of database issues with the FSA server.

Table 25: Maximum number of users by day of FSA test administration

Mon 3/2	Grades 8-10 Writing	31,832
Tues 3/3	Grades 8-10 Writing	38,930
Wed 3/4	Grades 8-10 Writing	33,389
Thurs 3/5	Grades 8-10 Writing	52,453
Fri 3/6	Grades 8-10 Writing	31,923
Mon 3/9	Grades 8-10 Writing	30,499
Tues 3/10	Grades 8-10 Writing	43,297
Wed 3/11	Grades 8-10 Writing	22,592
Thurs 3/12	Grades 8-10 Writing	11,432
Fri 3/13	Grades 8-10 Writing	3,469
Mon 4/13	(Grades 3-10 R, 3-8 M)	108,392
Tues 4/14	(Grades 3-10 R, 3-8 M)	140,092
Wed 4/15	(Grades 3-10 R, 3-8 M)	134,086
Thurs 4/16	(Grades 3-10 R, 3-8 M)	144,716
Fri 4/17	(Grades 3-10 R, 3-8 M)	82,140
Mon 4/20	(Grades 3-10 R, 3-8 M; EOC)	31,901
Tues 4/21	(Grades 3-10 R, 3-8 M; EOC)	170,132
Wed 4/22	(Grades 3-10 R, 3-8 M; EOC)	161,985
Thurs 4/23	(Grades 3-10 R, 3-8 M; EOC)	134,710
Fri 4/24	(Grades 3-10 R, 3-8 M; EOC)	111,426
Mon 4/27	(Grades 3-10 R, 3-8 M; EOC)	111,600
Tues 4/28	(Grades 3-10 R, 3-8 M; EOC)	143,299
Wed 4/29	(Grades 3-10 R, 3-8 M; EOC)	112,745
Thurs 4/30	(Grades 3-10 R, 3-8 M; EOC)	110,754

Date	Time	Max Users
Fri 5/1	(Grades 3-10 R, 3-8 M; EOC)	68,146
Mon 5/4	(Grades 3-10 R, 3-8 M; 8-10 W; EOC)	69,665
Tues 5/5	(Grades 3-10 R, 3-8 M; 8-10 W; EOC)	75,023
Wed 5/6	(Grades 3-10 R, 3-8 M; 8-10 W; EOC)	56,244
Thurs 5/7	(Grades 3-10 R, 3-8 M; 8-10 W; EOC)	44,518
Fri 5/8	(Grades 3-10 R, 3-8 M; 8-10 W; EOC)	25,328
Mon 5/11	(Grades 3-10 R, 3-8 M; EOC)	39,691
Tues 5/12	(Grades 3-10 R, 3-8 M; EOC)	17,886
Wed 5/13	Algebra 1, Geometry, Algebra 2	30,678
Thurs 5/14	Algebra 1, Geometry, Algebra 2	18,406
Fri 5/15	Algebra 1, Geometry, Algebra 2	5,974

Shifts in Administration Policy

Description of Administration Issues. During the focus group meetings, some district representatives shared their experiences related to changes in policy implementation that occurred over time as the FSA administrations continued. They specifically cited the rules and guidance related to students moving into test sessions inadvertently and earlier than scheduled. According to the Test Administrator Manual, students that advance to the next test session should then complete the test session on that day and be permitted the time necessary to do so. After the completion of testing, school staff needed to follow up with the student's parent to determine if the test score should be considered valid and used given the events of the test administration.

Early in the FSA administration windows, district representatives reported that their peers adhered closely to this policy because test administrators were acutely aware of the seriousness and consequences of test administration violations. As testing continued, the volume of students advancing across test sessions increased, which introduced significant test scheduling complications for many districts. Some districts reported that the administration rules were loosened in their district to facilitate getting as many students completed as possible.

Evidence. The evaluation team began their investigation into this issue by first sharing the feedback from the district representatives with FLDOE. Staff members from FLDOE stated that the official policy related to the movement across test sessions remained as it was stated within the Test Administrator Manual throughout the spring FSA administrations. However, feedback from FLDOE suggests that the Department regularly resolves this type of issue on a case-by-case basis after reviewing the extent and cause of the student moving into the next session. This year, on the first day when the issue was first brought to the attention of FLDOE, the instruction was to require students who entered session 2 to complete it that day. Later that day, the decision was made to allow students who entered the second session due to technological difficulties to complete testing on a later day. All subsequent cases were dealt with in the same manner and consistent with this decision.

As was previously discussed and is shown in Tables 18, a significant number of students advanced test sessions earlier than scheduled and did not complete the test session on that same day. Between 41% and 60% of students for Reading moved into the next test session completed the session on that same day.

In addition to information provided by FLDOE, AIR completed a set of analyses on the Reading and Math FSAs to determine if a consistent or prominent pattern of differential implementation of the administration policy could be detected. These analyses looked at the number of students who completed the entire test in 1 day across the entire testing window (either 2 sessions in one day for Reading or 2 or 3 sessions in one day for Math). Looking at Figure 11, a spike in the number of students who completed Reading on the first day of the administration can be observed; after that, no discernible pattern can be observed to indicate a widespread shift in how the policy was implemented across the state.

Figure 12 provides the same information for the Math testing window. A small increase in the latter part of the testing window can be observed; it is important to note that that the figure indicates a small increase of approximately 100 students over the time frame and that for most dates, the number of students actually taking the test ranges between 150,000 and 200,000 students. Therefore, these numbers indicate rather small percentages of the students tested.

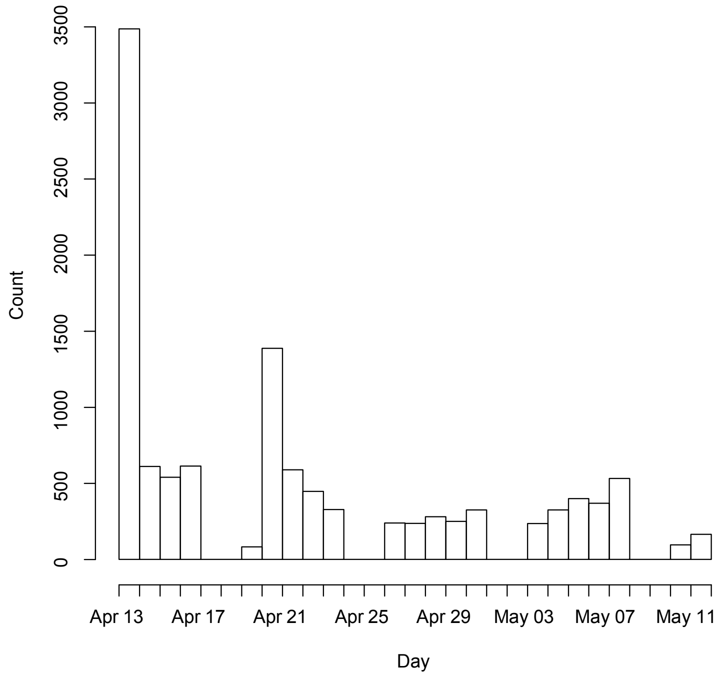


Figure 11: Number of students completing Reading in 1 day, by date.

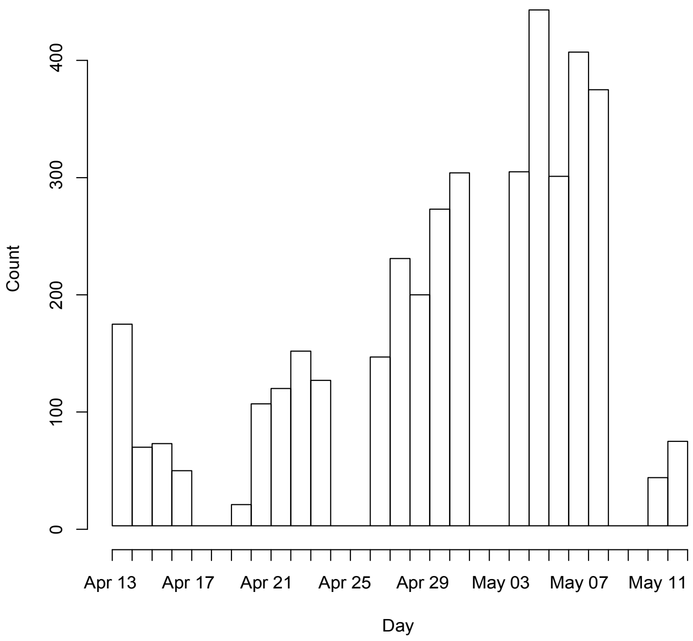


Figure 12: Number of students completing Mathematics in 1 day, by date.

Impact on Other Students

Description of Administration Issues. During the focus groups, many of the district representatives raised a concern that the issues encountered during the test administration could have impacted not only the immediate students encountering problems, but also the students in the same classrooms or testing sessions. District representatives also expressed a concern that mounting administration difficulties have a detrimental effect on the school as a whole as individuals may become frustrated. Such frustration could mean that students are not being placed in a situation that encourages their best performance.

Evidence. To evaluate this concern, AIR conducted a series of regression analyses that focused on predicting performance on the FSA using the prior years' FCAT 2.0 test scores. AIR completed this analysis at both the student and school level. At the student level, they did not find any meaningful differences in the ability of last years' test score to predict student performance. The school-level analyses was designed to evaluate if school-level impacts could be observed within schools that had students impacted by the difficulties with session movement in both Reading and Math. At the school level, no differences were observed in the prediction equation across the impacted and non-impacted schools.

Help Desk

Description of Administration Issues. One of the other persistent issues that arose during the investigation was concerns about the quality of the Help Desk assistance. As was described earlier, the *Test Standards* state that adequate support must be provided to help resolve any testing issues that may arise during the test administration. At the focus group meetings, district representatives were universally critical of the FSA Help Desk. Discussions included the difficulties getting through to the Help Desk, the poor preparation of the people who staffed the Help Desk, and the lack of follow through after questions were submitted to the Help Desk. Many district representatives also stated that as the test administration continued, they eventually stopped even using the FSA Help Desk because it was not beneficial and was perceived as a waste of time.

Many district representatives also indicated that the individuals staffing the Help Desk did not appear to have adequate training; many of these individuals were simply reading from a technical manual, and did not seem to understand the issues that were being encountered. Still other participants indicated that when they tried to resolve some issues with the Help Desk, the individuals staffing the Help Desk did not have the appropriate sign-on credentials, and were not able to work with the districts without "borrowing" the credentials from the district employee.

Evidence. While there is no way to gauge the impact of the Help Desk issues on student performance, the evaluation team did request feedback on the Help Desk as part of the online survey. On that survey, approximately 74% of respondents rated the Help Desk service as Poor or Exceptionally Poor. On that same question, only 2 of the 54 respondents rated the Help Desk service as Good, and none of the respondents rated the Help Desk as Excellent.

Training/Timeliness of Materials

Description of Administration Issues. One of the persistent issues that arose as a concern during the investigation was that many district representatives did not believe they were provided with sufficient training and information to support the implementation of the FSA. In some scenarios, this was described as information arriving too late for the district representatives to adequately respond or train staff members; in other cases, the feeling was that materials that were delivered were not sufficient or did not supply enough information.

As was mentioned at the beginning of this study description, the *Test Standards* stress that the sponsors of any testing program are responsible to provide appropriate training and support to individuals who will be responsible for administering the assessments. Poor or inadequate training can lead to significant issues within specific testing locations and can also possibly lead to serious differences in administration practices across testing locations. Some of the specific concerns that were mentioned by individuals were focused on 1) the use of calculators, 2) the text-to-speech feature that was supposed to be available for Reading and Math, 3) the late delivery of some training materials, and 4) and the proper administration of Listening items on the Reading test. A description of each of these issues is provided, along with the evidence available for each.

Calculator Use

Description of Administration Issues. Many districts reported a significant amount of confusion related to the calculator policy. At the beginning of the school year, districts were informed that students would not be able to use handheld calculators during the FSA administration; instead, students would need to use the on-screen calculator that would be supplied as part of the FSA administration system. However, after multiple complaints, FLDOE revised the policy in December 2014, and allowed some handheld calculators to be used. However, when the policy was changed, FLDOE did not release a list of approved calculators; instead, FLDOE released a list of prohibited functions that could not be present on calculators used during the administration. The decision not to provide a list of approved calculators was problematic because many schools had difficulty determining what function specific calculators did and did not have. Schools struggled with making those final decisions. The lateness of the decision to change the policy was also problematic because many students and schools had already purchased calculators; if the calculators had any of the prohibited functions, students could no longer use them.

Evidence. In the survey of district test administrators, approximately 60% of respondents indicated that the use of calculators caused some level of difficulty for them during the FSA administration. As can be seen in Table 26, the problems included test administrators allowing the use of calculators during the administration and difficulty identifying the appropriate handheld calculators.

Table 26: District Assessment Coordinators Survey Responses Related to Calculator Issues During the 2015 FSA Administration

Please indicate the types of [calculator] issues that were encountered (check all that apply).	
Test administrators permitted calculator use during non-calculator test sessions	66.67% (22)
The district had difficulties identifying approved handheld calculators	57.58% (19)
The district or schools had difficulties providing approved handheld calculators	51.52% (17)
Students had challenges using the onscreen calculator	27.27% (9)

Text-to-Speech Tool

Description of Administration Issues. At the beginning of planning for the spring 2015 FSA administration, schools and districts were informed that a text-to-speech feature would be available for all students who received an oral presentation accommodation on any of the Reading and Math assessments. However, just before the CBT administration window opened for Reading and Math, districts were informed that the text-to-speech would no longer be available.

FLDOE informed district by phone starting on Friday, March 27; the administration window was scheduled to start on Monday, April 13. School districts had limited time to adjust their schedule, develop resources, and prepare test administrators for this change, which led to considerable administrative difficulties for all parties involved.

Evidence. The difficulty with the text-to-speech feature was discussed at length during the focus group meetings with district representatives as well as at the Administration Debrief Meeting held in Tallahassee. One important issue here is that the guidelines for read-aloud accommodations for the FSA were different than what had been used with the FCAT 2.0, so adjustments were required of schools and districts, which made the last minute shift somewhat more difficult to manage. As this was primarily an administrative problem that negatively impacted schools and districts and their ability to prepare for the FSA administration, direct impacts on students would not be expected to be observed for the subgroup of students who were approved to use this accommodation.

Late Delivery of Training Materials

Description of the Administration Issues. Both FLDOE and its vendors are responsible for the delivery of a wide range of training materials and documents to districts in Florida, who are then responsible for the dissemination of these materials to their schools and the training of school representatives. For the 2014-15 academic year, some evidence suggests that some materials were delivered later than normal; district representatives were placed in the difficult position of completing training and setup with very limited timeframes, new system requirements, and many other unknowns that come with the first year of a new program. For example, the Writing Test Administration Manual was posted for districts more than a month later than in the 2013-14 academic year (January 15, 2015 in the 2014-15 academic year, as compared to November 27, 2013 in the 2013-14 academic year). Along the same lines, the EOC Training Materials for the CBT assessments were not delivered until January 30, 2015, whereas in the 2013-14 academic year, the materials were delivered on October 25, 2013.

Not all materials were delivered late; some materials were delivered at the same time as the previous year. Given that the 2014-15 academic year is the first year of the FSA, some administrative difficulties are not unexpected. In addition, the evaluation team considered the delivery of materials during the 2010-11 academic year, when the previous iteration of the Florida assessment program was introduced. In comparing the delivery of the FSA materials to those delivered in 2010-11, many of the materials were delivered earlier for the FSA. For example, the test item specifications for the FSAs were delivered in June and July of 2014. In comparison, while test item specifications for the Algebra exam for the FCAT 2.0 were delivered in July of 2010, the remaining Math specifications were delivered in December of 2010, and the Reading specifications were delivered in January of 2011. The Test Design Summary for the FSA was delivered on June 30 of 2014; in comparison, the Test Design summary for the FCAT 2.0 was delivered on September 9 of 2010.

Evidence. The difficulty with the late delivery of materials was discussed at length during the focus group meetings with district representatives as well as at the Administration Debrief Meeting held in Tallahassee. This was primarily an administrative problem that negatively impacted schools and districts and their ability to prepare for the FSA administration; therefore, direct impacts on students would not be expected to be observed.

Listening Items in Reading

Description of Administration Issues. Many school districts reported difficulties with the Listening items on the Reading test. The primary difficulty that was encountered was that if the headphones were not plugged into the computer being used prior to launching the secure browser for the test, the headphones would not work when Listening items were encountered. In this case, the test administrators had been instructed to test the headphones prior to the test starting. However, many administrators thought this only had to be completed once with a given computer, and were not aware that failing to plug in the headphones at the beginning of each test could interfere with the headphones functioning.

Further complicating these matters, not every Reading session actually contained Listening items. This left many students with headphones throughout the entire test, without ever needing the headphones. This caused even more disruption because many students were uncertain if they had missed the Listening items. For many test administrators, the exact reason why the headphones were required was unclear; these administrators reported that they had not received adequate information or training on how to properly use the headphones.

Evidence. The difficulty with the Listening items was discussed at length during the focus group meetings with district representatives as well as at the Administration Debrief Meeting held in Tallahassee. This issue alone was not a significant problem for schools and districts alone; as such, we would not expect to see significant impact on students from the Listening items.

However, it does highlight an important component of this evaluation. Like the Listening items, the other items listed here as individual issues around training and material may not rise to the level of a serious problem that solely compromises the integrity of the assessments; however, the cumulative effect should be considered as well. On the survey of district test administrators, more than 50% of the respondents estimated that 10% or more of their students were impacted by the various FSA technology challenges.

It is also important to note that many individuals raised concerns about the preparation of schools for the FSA administration prior to the administration. In February 2015, school districts were required to attest to the readiness of the schools in their district for the FSA. This had been done in previous years and was primarily focused on the systems and infrastructure of each school. This year, during that certification, 28 school districts included letters raising significant concerns about the ability of their school district to administer the FSA. The concerns raised by district superintendents ranged from needing more resources to administer the test, the negative impact on student learning as computer labs were occupied, and the ability to deliver the tests. Twenty of these letters raised concerns about the infrastructure of their school district or state to deliver the FSAs; 15 of these letters raised concerns about student familiarity with the CBT delivery system and that they had not received adequate time to understand the system, and 14 of these letters mentioned that schools had not had sufficient time to prepare for the FSA.

Findings

The 2014-15 FSA test administration was problematic; issues were encountered on just about every aspect of the computer-based test administrations, from the initial training and preparation to the delivery of the tests themselves. The review of test user guides and test administration guides indicate that the intended policies and procedures for the FSA were consistent with the *Test Standards*. However, as revealed throughout the survey and focus groups with district representatives, the administration difficulties led to a significant number of students not being presented with a test administration model that allowed them to demonstrate their knowledge and skills on the FSA.

Looking at the statewide data, a somewhat contradictory story emerges. The percentage of students that can be identified as directly impacted by any individual test administrations problem appears to be within the 1% to 5% range, depending on the specific issue and test. Because of these discrepancies, the precise number of students impacted by these issues is difficult to define, and will always be qualified by the precise definition of the term impact and on the data available. Despite these reservations, the evaluation team does feel like they can reasonably state that the spring 2015 administration of the FSA did not meet the normal rigor and standardization expected with a high-stakes assessment program like the FSA.

Commendations

- Throughout all of the work of the evaluation team, one of the consistent themes amongst people we spoke with and the surveys was the high praise for FLDOE staff members who handled the day-to-day activities of the FSA. Many district representatives took the time to praise their work and to point out that these FLDOE staff members went above and beyond their normal expectations to assist them.

Recommendations

Recommendation 4.1 FLDOE and its vendors should be more proactive in the event of test administration issues.

Standard 6.3 from the Test Standards emphasizes the need for comprehensive documentation and reporting anytime there is a deviation from standard administration procedures. It would be appropriate for FLDOE and its vendors to create contingency plans that more quickly react to any administration-related issues. These steps could include policies such as consultation with state TAC members, enhanced communication with its constituents, and validity agendas that directly address any possible administration related issues. In addition, when issues are encountered during an administration, it would be advantageous of FLDOE and its vendors to begin explorations into the related impacts immediately.

Recommendation 4.2 FLDOE and its FSA partners should engage with school districts in a communication and training program throughout the entire 2015-16 academic year.

Given the extensive nature of the problems with the 2014-15 FSA administrations, there is now a loss of confidence in FLDOE, its vendors, and the FSA program. Many individuals expressed extreme frustration at the difficulties that were encountered and the apparent lack of action despite their extensive complaints. The individuals who have expressed these concerns are not individuals who could be classified as "anti-testing" or individuals who do not support the FLDOE. Instead, these individuals have worked on the ground of the Florida statewide testing program and now have serious doubts that must be addressed.

Recommendation 4.3 FLDOE should review and revise the policies and procedures developed for the FSA administration to allow the test administrators to more efficiently deliver the test, and when required, more efficiently resolve any test administration issues.

Test administration manuals and other training materials for all FSAs should be reviewed to determine ways to more clearly communicate policies such as the transition from one test session to the next. In addition, test administrators need to be provided with more time to review and understand the procedures prior to the administration.

The process for handling any test administration should also be addressed. Many individuals with whom the evaluation team spoke described an onerous process to submit any request to the FSA Help desk, involving the test administrator, the school administrator, and finally the district administrator. In addition, many others described needing to be in the room itself where the test administration was occurring to resolve certain issues, which disrupted not only the immediate student(s) impacted, but other students in the room as well.

The FSA Help Desk also needs to be evaluated and procedures need to be put in place to make it more productive. Help Desk employees should be more familiar with the FSA and should be equipped with the appropriate access to efficiently work with schools and districts that have encountered a problem.

Study 5: Evaluation of Scaling, Equating, and Scoring

Study Description

In conducting this study, the evaluation team planned to review seven sources of evidence through a review of documentation and conducting in-person and virtual interviews with staff at FLDOE and partner vendors. These sources of evidence were:

- Review evidence of content validity collected by the program for the following:
  - Qualified subject matter experts
  - Appropriate processes and procedures
  - Results that support claims of content validity
- Review rationale for scoring model, analyses, equating, and scaling for the following:
  - Evidence that supports the choice of the scoring model
  - Implementation and results of the psychometric analyses
  - Design, implementation, results, and decision rules for equating
  - Design, implementation, results, and decision rules for scaling for total scores and domain or subscores
- Review psychometric characteristics of the assessments for the following:
  - Analyses of reliability, inclusive of standard error of measurement
  - Decision consistency and accuracy
  - Subscore added value analyses
- Review psychometric characteristics of subgroups for the following:
  - Psychometric performance of assessment items for reporting subgroup performance (e.g., reliability of subgroups, differential item functioning)
- Review evidence of construct validity collected by the program
- Review evidence of criterion validity collected by the program for the following:
  - Identified criterion variables and related studies
- Review evidence of testing consequences collected by the program

Sources of Evidence

The following documents served as the primary sources of evidence for this study:

- Florida Standards Assessment 2014-2015 Scoring and Reporting Specifications Version 1.0
- 2015 Calibration and Scoring Specifications
- Handscoring Specifications: Florida Standards Assessments ELA Writing Spring 2015 & Fall 2015
- Mathematics Test Design Summary – Updated 11-24-14
- ELA Test Design Summary – Updated 11-24-14
- Summary of Daily Calibration Call Process

- Proposed Plan for Vertical Linking the Florida Standards Assessments
- FSA Assessments Approval Log 7-2-15
- Florida Department of Education Early Processing Sample Design
- Constructed Response Scoring Patents
- Automated Essay Scoring information from AIR FSA proposal communications
- Master Data Files for each test (includes calibration data) files

Study Limitations

Information needed to fully evaluate the processes and data included in this study was not available. Areas for which analyses and development of related documentation is ongoing includes:

- Subgroup psychometric characteristics
- Subscore added value analyses, decision consistency, and measurement precision

Areas for which analyses and development of related documentation is not available includes:

- Criterion evidence collected by the program
- Evidence of testing consequences produced by the program

Additionally, the evaluation studies related to the test items (Studies #1 and #6), and the test blueprints (Study #3) focused on a review of the evidence related to content validity. Therefore, the majority of the work for this study focused on a review of psychometric model, scoring, analyses, equating and scaling.

Industry Standards

The activities included in this study take raw student data, assign score values to them and, then translate that information into readily used information for the various uses of the assessments. These activities are essential to the program's accuracy, reliability, fairness, and utility.

As is true of each aspect of this evaluation, the *Test Standards* served as a primary source when considering the scoring, calibrations, equating, and scaling of the FSA assessments. These activities are technical in nature, and the *Test Standards* do not provide much detail related to the various psychometric methods that can be used; therefore, other source documents were utilized as well. These sources include books devoted to each of the activities that are included in this study like Kolen and Brennan's *Test Equating, Scaling, and Linking: Methods and Practice* (2004).

While the *Test Standards* do not provide preference or evaluation of various psychometric or statistical models, several standards call out the importance of processes, protocols and documentation related to the scoring, calibrations, equating, and scaling of assessments. Specifically, Standards 6.8, 6.9, and 12.6 state the need for formal and well-documented scoring practices, including information related to accuracy and quality. Standard 5.2 notes the need for thorough documentation related to the selection and creation of score scales.

These Standards, their accompanying narratives, and various seminal texts from the field of measurement were used to evaluate the processes and, where possible, the results of the FSA program related to scoring, calibrations, equating, and scaling. The following section describes this evaluation effort.

Florida Standards Assessments Processes and Evaluation Activities

Scoring

Depending on the item types administered, scoring can consist of a variety of procedures. For multiple-choice items and some technology-enhanced item types where students select responses from given options or manipulate stimuli, scoring is typically done in a straightforward manner using computer systems. For other item types that require students to generate an answer rather than select an answer from options provided, scoring is done by computer, through human raters, or a combination of scoring methods (Williamson, Mislevy, & Bejar, 2006). FSA employs each of these types of scoring as described below:

- Multiple-choice items on FSA Reading and Mathematics tests are computer scored.
    - For the computer-based tests (CBT), student responses are passed from the test administration system to the scoring system.
    - For the paper-based tests, student responses are scanned from the answer documents into the scoring system.
- Technology-enhanced items on FSA computer-based Reading and Math tests are computer scored. In some cases, a Math-driven algorithm is used to score some items (e.g., those that require students to plot on a coordinate plane).
- The essay items on the FSA Writing test were scored by trained human raters. Each student response received two scores. For most grades, both scores were provided by human raters. In grades 8 and 9, student responses received one score from a human rater and one score from an automated computer-based scoring engine.

For the evaluation activities, FLDOE, along with the FSA testing vendors AIR and DRC, provided a number of documents that describe the scoring-related activities. This included some information related to the computer-based scoring algorithms and scoring engine, specifically from patents and FSA proposal communications. In addition, DRC provided the hand-scoring specifications for the human rater scoring process, which outlined the training, processes, and quality control procedures related to the human scoring of student essay responses. Alpine reviewed these documents and discussed details of these procedures during several meetings, including an in-person meeting with FLDOE, AIR, and DRC on July 13 and 14 in Washington, D.C.

Calibrations

An important step in the analyses procedures is to complete calibrations (i.e., psychometric analyses to determine empirical performance) of the administered items. These analyses are conducted by applying one or several statistical models to the data and using these models to provide a variety of information including the difficulty level of items and the degree to which the items distinguished between high and low performing students (i.e., item discrimination). Data from these calibrations are then used to evaluate the performance of items using statistical criteria. Any items that are identified based on these statistical criteria are reviewed by psychometricians and content experts. If needed, items may be removed from the scored set meaning that they would not impact students' scores.

Ideally, data from all students across the state would be used to conduct calibration activities. As is commonly observed in practice, the FSA administration and scoring schedules required that a sample of student data be used for calibrations for some tests. For these grades and content areas, the samples were created to represent the full population of students by considering variables like geographic region, school size, gender, and ethnicity. AIR and FLDOE provided documentation related to the sampling plans and implementation as part of the evaluation.

For the FSA, three different item response theory (IRT) models were used for the calibrations, depending on the item types as follows:

- For multiple-choice items, the 3-parameter logistic (3PL) model was used.
- For dichotomous items, (i.e., those scored right or wrong) where student guessing was not relevant, the 2-parameter logistic (2PL) model was used.
- For polytomous items (i.e., those with multiple score points), the generalized partial credit (GPC) model was used.

Results of these model applications were reviewed by AIR and FLDOE staff to evaluate model fit by item. Model choice adjustments were made, as needed, based on the results.

Calibrations were completed primarily by AIR staff and then verified by FLDOE as well as Human Resources Research Organization (HumRRO) and Buros Center for Testing, two independent organizations contracted by FLDOE to provide quality assurance services. Once the results of calibrations from each of these groups matched, AIR and FLDOE reviewed the item statistics, specifically considering statistics related to model fit, item difficulty, item discrimination, distractor analyses, and differential item functioning (DIF). AIR and FLDOE then met regularly to review these statistics, flag items for review, rerun calibrations, meet with content experts as part of the review process, and make final item-level scoring decisions. AIR and FLDOE provided Alpine with the specifications for the calibration analyses, a summary of the review activities, as well as a log of the items that were flagged and the associated follow-up actions.

Calibration activities were done in several stages in support of different program aspects. These activities included calibrations for the scorable (as opposed to unscored or field test) items, for the development of the vertical scale, and for the field test items that will be considered for use on forms in future years. The calibrations for the scorable items were completed early enough in the study to be included within the evaluation. Other calibration work was ongoing or not completed in time for inclusion.

Equating

Equating is commonly done when multiple forms of the same test are used either within the same administration or over time. Through statistical processes, equating assures that scores across test forms can be compared and that student performance can be interpreted relative to the same performance or achievement standard regardless of the individual items they experience.

"Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably" (Kolen & Brennan, 2004, p.2).

Because 2014-15 was the first year of the FSA program and because only one form was developed and administered for most grades and content areas, equating was not needed for most tests. In a few areas, specifically Algebra 1 and accommodated test forms, equating was employed.

Unlike other grades and content areas that only had one FSA test form, three forms were developed and administered for Algebra 1. In addition to Algebra 1, equating was also needed for paper-based accommodated test forms. For those tests where the primary test administration mode was computer, the creation of accommodated forms included the review and consideration of the item functionality in a paper-based format. Some items required modifications to adjust for the differing administration modes. Some other items, primarily technology-enhanced items, could not be adapted for paper-based administration without modifying the content or skills assessed. Because of these differences in items across the computer-based and paper-based accommodated forms, equating is needed to adjust the scores and make them comparable across these forms.

Specific steps within the equating process are related to the score scale on which results are reported as well as the performance standards on the test. As is described in the next section, the scaling work is ongoing for FSA. In addition, standard setting meetings, which are used to set performance standards, had not yet been completed. Because the scaling and standard setting activities were ongoing, additional work related to equating remains to be completed. Therefore, a full evaluation of this work was not available for this study.

Scaling

Raw scores, or number correct scores, "are often transformed to scale scores… to enhance the interpretability of scores" (Kolen & Brennan, 2004, p. 4). This creation of score scales can be done in a wide variety of ways depending on the intended purpose and uses of the scores. FLDOE has chosen to place FSA scores for grades 3-10 ELA and grades 3-8 Math on vertical scales. With a vertical scale, student performance across grade levels is reported on one continuous scale in an attempt to support cross-grade interpretability of scores. This contrasts to horizontal scales, which do not connect performance across grade levels. The benefit of a vertical scale is that it is intended to provide a readily interpretable metric to consider students' development and progression over time.

As is common in vertical scale development, considerations for the FSA vertical scale began during the construction of test forms. In addition to the set of items used to generate student scores, FSA test forms also included a small subset of embedded items for the purpose of field testing or other development activities (e.g., the development of the vertical scale). While students received the same set of scorable items (except for Algebra 1 and accommodated paper-based test forms), the items used for field testing or development activities varied.

Some students completed the embedded items whose purpose was the development of the vertical scale. These vertical scale items included items that were on-grade level as well as those from the grade level above and below that of the test. For example, the grade 5 vertical scale items included items from grades 4, 5, and 6. The student performance on these vertical scale items served as the basis of the FSA vertical scale development. The selection of vertical scale items included review of content and statistical criteria. After the administration, these items were again reviewed based on item statistics. AIR and FLDOE provided the vertical scale development plan for the FSA, and through several meetings, Alpine gained additional information related to the details of the plan's implementation. AIR also provided a summary of preliminary results for the Math vertical scale.

Findings

Based on the documentation and results available, acceptable procedures were followed and sufficient critical review of results was implemented. In addition, FLDOE and AIR solicited input from industry experts on various technical aspects of the FSA program through meetings with the FLDOE's Technical Advisory Committee (TAC). In addition to formal meetings with the full TAC, FLDOE and AIR also sought input from individual TAC members related to specific program details and results as data analyses were ongoing.

Using the *Test Standards*, as well as other prominent texts like Kolen and Brennan (2004), FSA policies and procedures for scoring, calibrations, and scaling were compared to industry practice.

It is worth noting that a good deal of work related to these activities is ongoing or yet to be conducted.

Commendations

- Although AIR committed to the development of the FSA program within a relatively short timeframe, the planning, analyses, and data review related to the scoring, calibrations of the FSA (i.e., the work that has been completed to date) did not appear to be negatively impacted by the time limitations. The procedures outlined for these activities followed industry standards and were not reduced to fit within compressed schedules.

Recommendations

Recommendation 5.1 Documentation of the computer-based scoring procedures, like those used for some of the FSA technology-enhanced items as well as that used for the essays, should be provided in an accessible manner to stakeholders and test users.

Standard 12.6:
Documentation of design, models and scoring algorithms should be provided for tests administered and scored using multimedia or computers.

It was expected that the documentation for the scoring, calibration, equating, and scaling activities would be hampered by the timing of the evaluation and the ongoing program activities. For example, it was not a surprise to the evaluation team to receive complete planning documents but no formal technical report related to these activities as they were occurring concurrently to the study. However, computer-based scoring technology that AIR implemented for FSA has been used elsewhere with other states and assessment programs. Therefore, the documentation around these scoring procedures should already exist and be available for review in formats that are readily accessible to stakeholders (e.g., scoring algorithms for FSA technology-enhanced items was embedded within patent documents). The limited availability of this information only serves to introduce questions and speculation about the procedures that are used and their quality.

Study 6: Specific Evaluation of Psychometric Validity

Study Description

To evaluate the specific elements of psychometric validity requested by FLDOE, the evaluation team reviewed documentation regarding development activities using criteria based on best practices in the industry. To supplement the information contained in documentation, the team conducted in-person and virtual interviews with FLDOE and partner vendors to gather information not included in documentation or to clarify evidence. The following elements were planned for inclusion within this study:

- Review a sample of items from each grade and subject for the following:
    - Content, cognitive processes, and performance levels of items relative to standards as described in course descriptions
    - Design characteristics of items that reduce the likelihood that the student answers the question correctly by guessing
    - Evidence of fairness or bias review
- Review psychometric characteristics of items for the following:
    - Item difficulty results with an acceptable range of parameters
    - Item discrimination results with an acceptable range of parameters
    - Option analyses for functional item response characteristics
    - Empirical evidence of potential bias such as differential item functioning
- Review the linking processes for Algebra 1 and Grade 10 ELA to 2013-14 results for the following:
    - Assumptions for the linking studies
    - Design of the linking studies
    - Results and associated decision rules applied in the linking studies
    - Communication reports regarding the linking and the information to schools and other Florida constituents

Sources of Evidence

The following documents served as the primary sources of evidence for this study:

- Florida Standards Assessment 2014-2015 Scoring and Reporting Specifications Version 1.0
- Mathematics Test Design Summary – Updated 11-24-14
- ELA Test Design Summary – Updated 11-24-14
- 2015 Calibration and Scoring Specifications
- Master Data Files for each test (include calibration data)
- FSA Assessments Approval Log

Study Limitations

The program documentation and activities permitted the completion of this study as intended and originally designed.

Industry Standards

In the review of item statistics and the resulting decision-making, the various criteria used, the process of the item evaluation, the student sample from which the data were obtained, and evidence of the appropriateness of the analysis procedures should all be well documented in adherence to Standard 4.10.

When scores from different tests or test forms are linked, as was done for FSA grade 10 ELA and Algebra 1 scores to those of FCAT 2.0, Standard 5.18 highlights the importance of documenting the procedures used, appropriate interpretations of the results, and the limitations of the linking. In addition to this guidance from the Test Standards, recommendations provided by Kolen and Brennan (2004) were also used, specifically in the evaluation of the linking procedure implemented.

Standard 5.18: When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.

Standard 5.18: When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.

Florida Standards Assessments Processes and Evaluation Activities

As outlined by the state, the focus of this study is psychometric validity, specifically related to the FSA item content, the item statistics and technical qualities, and the procedure used to link the grade 10 ELA and Algebra 1 scores to those from FCAT 2.0 in support of the mandated graduation requirement. There is significant overlap between the evaluation of the item content as requested for this study and the evaluation activities for Study 1. Rather than repeat that information, the reader should refer to Study 1 for the Sources of Evidence, FSA Processes, and Evaluation Activities related to FSA test item review. The following sections separately describe the remaining two aspects of the this study, the review of item statistics and qualities and the procedure used to link FSA and FCAT 2.0 scores, and the associated evaluation activities.

Item Statistics

In addition to reviewing item statistics pre-administration based on field test data (see Study #2 for more detail on how this was done for FSA), it is also typical to review item statistics after the operational administration of the test forms and prior to the completion of scoring activities.

For FSA, this step was of increased importance, as it was the first occasion to review statistics based on Florida student data as the field test was conducted in Utah.

After the spring 2015 FSA administration, AIR and FLDOE scored the items and ran a number of analyses to permit review of the psychometric characteristics and performance of the items. The review of item statistics included consideration of item difficulty, distractor analyses, item discrimination, differential item functioning (DIF) by ethnicity, gender, English language learners (ELLs), and students with disabilities (SWD). The criteria used for flagging items are as follows:

- P value < 0.20 (item difficulty, see Appendix A for a definition)
- P value > 0.90 (item difficulty, see Appendix A for a definition)
- Point biserial for distractor > 0 (distractor analysis, see Appendix A for a definition)
- Point biserial for correct answer < 0.25 (item discrimination)
- DIF classification = C

In addition to these statistics, the statistical model fit was also evaluated for each item. Flagged items were reviewed together by AIR and FLDOE staff, including both psychometricians and content experts, to determine if the items could be included for scoring.

The details of this post-administration review process were outlined within the 2015 Calibration and Scoring Specifications document. Additionally, FLDOE provided a description of the process that was used to review flagged items during daily phone calls between AIR and FLDOE throughout the review period. AIR and FLDOE also provided the evaluation team with the FSA Assessment Approval Log which lists the flagged items, the reasons for flagging, the final decision regarding the item use, and the justification for this decision.

Based on the criteria and processes used to review the statistical qualities of the items, the evaluation team found no cause for concern regarding the FSA items. The procedures implemented by AIR and FLDOE to review items post-administration follow those commonly used in similar assessment programs and adhere to the guidance provided by industry standards.

Linking of Florida Standards Assessments to FCAT 2.0

Per Florida statute 1003.4282, students must pass the statewide assessments for grade 10 ELA and Algebra 1 in order to earn a standard high school diploma.

As is common in assessment development, the passing scores or standard setting activities were scheduled to permit time for post-administration analyses and incorporation of data into the process. This schedule meant that the FSA standard setting activities would not occur until late summer/early fall 2015, months after the administration of the grade 10 ELA and Algebra 1 assessments in the spring. To meet legislative requirements, an interim standard for the spring 2015 administration was used based on the linking of the FSA and FCAT 2.0 tests.

AIR and FLDOE evaluated several options to determine the interim standards and consulted with members of the Technical Advisory Committee (TAC) as well as an expert specializing in assessment and the law. Equipercentile linking of the cut scores from FCAT 2.0 to FSA was selected as the approach for establishing the interim cut scores. Described simply, this process uses the percentile rank associated with the passing score on the FCAT 2.0 test in 2014 and finds the score on the FSA that corresponds with that same percentile rank (Kolen & Brennan, 2004).

Per Florida statute 1003.4282, students must pass the statewide assessments for grade 10 ELA and Algebra 1 in order to earn a standard high school diploma.

AIR and FLDOE provided the evaluation team with the calibration and scoring specifications which outlined the planned procedures for conducting the linking. In addition, during a meeting on July 13 and 14 in Washington, D.C., the groups discussed the steps taken to evaluate the available options, seek technical guidance from experts in the field, and select the equipercentile linking method.

From a psychometric perspective, this method of linking the two assessments is less than ideal because it is based on important assumptions that both tests are constructed using on the same framework and test specifications in order to support interpretations of equivalency of the resulting scores. The most apparent violation of this assumption, although not the only one, is the difference in content between the FCAT grade 10 Reading test and FSA grade 10 ELA test which includes both Reading and Writing. The alternative and preferred solution would be to reset the passing standard given the differences between the previous and new assessments. While this action will be taken, Florida legislation required that an interim passing score, based on the link of FSA to FCAT 2.0, be used for the spring 2015 FSA administration rather than delay reporting until after standard setting activities. Given this decision, the methodology applied in this instance was implemented out of necessity. FLDOE and AIR chose a process that met the needs of the FSA program using an acceptable, although less than ideal, solution given the state requirements.

Findings

Based on a review of both the item statistics and the score linking procedures, FLDOE and AIR appropriately and responsibly managed the psychometric activities of the FSA within the given program requirements. The post-administration review of the technical qualities of the FSA items adhered to industry standards and therefore does not present cause for concern. In regards to the linking of scores for grade 10 ELA and Algebra 1, FLDOE and AIR implemented a solution that served the purpose and requirement determined by the state. Concerns stemming from the psychometric approach and the soundness of the results were openly communicated and discussed with FLDOE.

The findings related to the review of FSA items, specifically regarding content, can be found in Study 1. While areas of improvement were noted as part of the evaluation, there was no significant cause for concern based on this review.

Commendations

- The operational application of psychometric standards and processes can be challenging given the political environment and the requirements placed upon a test program. AIR and FLDOE appear to have carefully navigated this path by openly discussing psychometric best practice and seeking alternatives, where needed, to fit the needs of the FSA requirements. Industry guidance from publications and psychometric experts was sought in support of this effort. Given an imperfect psychometric situation, both regarding the original source of items and the reporting requirements, AIR and FLDOE appear to have carefully found a balance that delivered acceptable solutions based on the FSA program constraints.

Recommendations

Recommendation 6.1 FLDOE should more clearly outline the limitations of the interim passing scores for the grade 10 ELA and Algebra 1 tests for stakeholders. Unlike the passing scores used on FCAT 2.0 and those that will be used for subsequent FSA administrations, the interim passing scores were not established through a formal standard setting process and therefore do not represent a criterion-based measure of student knowledge and skills. Since the results based on these interim standards have already been released, there may not be much that can be done about the misinterpretations of these data.

Recommendations related to the review of the FSA items can be found within Study 1.

Compilation of Recommendations

For ease of reading, the complete list of the recommendations, as identified within the previous sections for the individual studies, is provided here.

Recommendation 1.1: FLDOE should phase out the Utah items as quickly as possible and use items on FSA assessments written specifically to target the content in the Florida standards.

Recommendation 1.2: FLDOE should conduct an external alignment study on the entire pool of items appearing on the future FSA assessment with the majority of items targeting Florida standards to ensure documentation and range of complexity as intended for the FSA items across grades and content areas.

Recommendation 1.3: FLDOE should conduct cognitive laboratories, cognitive interviews, interaction studies involving the capture and analysis of data about how students engage with test items and the content within each of the items during administration, and/or other ways in which to gather response process evidence during the item development work over the next year.

Recommendation 2.1: FLDOE should provide further documentation and dissemination of the review and acceptance of Utah state items.

Recommendation 3.1 FLDOE should finalize and publish documentation related to test blueprint construction.

Recommendation 3.2 FLDOE should include standard specific cognitive complexity expectations (DOK) in each grade-level content area blueprint.

Recommendation 3.3 FLDOE should document the process through which the score reports and online reporting system for various stakeholders was developed, reviewed, and incorporated usability reviews, when appropriate.

Recommendation 3.4 FLDOE should develop interpretation guides to accompany the score reports provided to stakeholders.

Recommendation 4.1: FLDOE and its vendors should be more proactive in the event of test administration issues.

Recommendation 4.2: FLDOE and its FSA partners should engage with school districts in a communication and training program throughout the entire 2015-16 academic year.

Recommendation 4.3: FLDOE should review and revise the policies and procedures developed for the FSA administration to allow the test administrators to more efficiently deliver the test, and when required, more efficiently resolve any test administration issues.

Recommendation 5.1: Documentation of the computer-based scoring procedures, like those used for some of the FSA technology-enhanced items as well as that used for the essays, should be provided in an accessible manner to stakeholders and test users.

Recommendation 6.1: FLDOE should more clearly outline the limitations of the interim passing scores for the grade 10 ELA and Algebra 1 tests for stakeholders.

Conclusions

As the evaluation team has gathered information and data about the Florida Standards Assessments (FSA), we note a number of commendations and recommendations that have been provided within the description of each of the six studies. The commendations note areas of strength while recommendations represent opportunities for improvement and are primarily focused on process improvements, rather than conclusions related to the test score validation question that was the primary motivation for this project.

As was described earlier in the report, the concept of validity is explicitly connected to the intended use and interpretation of the test scores. As a result, it is not feasible to arrive at a simple Yes/No decision when it comes to the question "Is the test score valid?" Instead, the multiple uses of the FSA must be considered, and the question of validity must be considered separately for each. Another important consideration in the evaluation of validity is that the concept is viewed most appropriately as a matter of degree rather than as a dichotomy. As evidence supporting the intended use accumulates, the degree of confidence in the validity of a give test score use can increase or decrease. For purposes of this evaluation, we provide specific conclusions for each study based on the requested evaluative judgments and then frame our overarching conclusions based on the intended uses of scores from the FSA.

Study-Specific Conclusions

The following provide conclusions from each of the six studies that make up this evaluation.

Conclusion #1 – Evaluation of Test Items

When looking at the item development and review processes that were followed with the FSA, **the policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the test items were determined to be error free, unbiased, and were written to support research-based instructional methodology, use student- and grade-appropriate language as well as content standards-based vocabulary, and assess the applicable content standard.

Conclusion #2 – Evaluation of Field Testing

Following a review of the field testing rationale, procedure, and results for the FSA, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the field testing design, process, procedures, and results support an assertion that the sample size was sufficient and that the item-level data were adequate to support test construction, scoring, and reporting for the purposes of these assessments.

Conclusion #3 – Evaluation of Test Blueprint and Construction

When looking at the process for the development of test blueprints, and the construction of FSA test forms, **the methods and procedures that were followed are generally consistent with expected practices as described in the *Test Standards*.** The initial documentation of the item development reflects a process that meets industry standards, though the documentation could be enhanced and placed into a more coherent framework. Findings also observed that the blueprints that were evaluated do reflect the Florida Standards in terms of overall content match, evaluation of intended complexity as compared to existing complexity was not possible due to a lack of specific complexity information in the blueprint. Information for testing consequences, score reporting, and interpretive guides were not included in this study as the score reports with scale scores and achievement level descriptors along with the accompanying interpretive guides were not available at this time.

Conclusion #4 – Evaluation of Test Administration

Following a review of the test administration policies, procedures, instructions, implementation, and results for the FSA, **with some notable exceptions, the intended policies and procedures that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, some aspects of the test administration, such as the test delivery engine, and the instructions provided to administrators and students, were consistent with other comparable programs. However, for a variety of reasons, the 2014-15 FSA test administration was problematic, with issues encountered on multiple aspects of the computer-based test (CBT) administration. These issues led to significant challenges in the administration of the FSA for some students, and as a result, these students were not presented with an opportunity to adequately represent their knowledge and skills on a given test.

Conclusion #5 – Evaluation of Scaling, Equating, and Scoring

Following a review of the scaling, equating, and scoring procedures and methods for the FSA, and **based on the evidence available at the time of this evaluation, the policies, procedures, and methods are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry**. Specifically, the measurement model used or planned to be used, as well as the rationale for the models was considered to be appropriate, as are the equating and scaling activities associated with the FSA. Note that evidence related to content validity is included in the first and third conclusions above and not repeated here. There are some notable exceptions to the breadth of our conclusion for this study. Specifically, evidence was not available at the time of this study to be able to evaluate evidence of criterion, construct, and consequential validity. These are areas where more comprehensive studies have yet to be completed. Classification accuracy and consistency were not available as part of this review because achievement standards have not yet been set for the FSA.

Conclusion #6 – Evaluation of Specific Psychometric Validity Questions

Following a review of evidence for specific psychometric validity questions for the FSA, **the policies, methods, procedures, and results that were followed are generally consistent with expected practices as described in the *Test Standards* and other key sources that define best practices in the testing industry with notable exceptions**. Evidence related to a review of the FSA items and their content are noted in the first conclusion above and not repeated here. The difficulty levels and discrimination levels of items were appropriate and analyses were conducted to investigate potential sources of bias. The review also found that the psychometric procedures for linking the FSA Algebra 1 and Grade 10 ELA with the associated FCAT 2.0 tests were acceptable given the constraints on the program.

Cross-Study Conclusions

Because validity is evaluated in the context of the intended uses and interpretations of scores, the results of any individual study are insufficient to support overall conclusions. The following conclusions are based on the evidence compiled and reviewed across studies in reference to the intended uses of the FSAs both for individual students and for aggregate-level information.

Conclusion #7 – Use of FSA Scores for Student-Level Decisions

With respect to student level decisions, **the evidence for the paper and pencil delivered exams support the use of the FSA at the student level. For the CBT FSA, the FSA scores for some students will be suspect. Although the percentage of students in the aggregate may appear small, it still represents a significant number of students for whom critical decisions need to be made. Therefore, test scores should not be used as a sole determinant in decisions such as the prevention of advancement to the next grade, graduation eligibility, or placement into a remedial course**. However, under a “hold harmless” philosophy, if students were able to complete their tests(s) and demonstrate performance that is considered appropriate for an outcome that is beneficial to the student (i.e., grade promotion, graduation eligibility), it would appear to be appropriate that these test scores could be used in combination with other sources of evidence about the student’s ability. This conclusion is primarily based on observations of the difficulties involved with the administration of the FSA.

Conclusion #8 – Use of Florida Standards Assessments Scores for Group-Level Decisions

In reviewing the collection of validity evidence from across these six studies in the context of group level decisions (i.e., teacher, school, district or state) that are intended uses of FSA scores, **the evidence appears to support the use of these data in the aggregate**. **This conclusion is appropriate for both the PP and the CBT examinations.** While the use of FSA scores for individual student decisions should only be interpreted in ways that would result in student outcomes such as promotion, graduation, and placement, the use of FSA test scores at an aggregate level does appear to still be warranted. Given that the percentage of students

with documented administration difficulties remained low when combining data across students, schools and districts, it is likely that aggregate level use would be appropriate.

The primary reason that aggregate level scores are likely appropriate for use is the large number of student records involved. As sample sizes increase and approach a census level, and we consider the use of FSA at the district or state level, the impact of a small number of students whose scores were influenced by administration issues should not cause the mean score to increase or decrease significantly. However, cases may exists where a notably high percentage of students in a given classroom or school were impacted by any of these test administration issues. It would be advisable for any use of aggregated scores strongly consider this possibility, continue to evaluate the validity of the level of impact, and implement appropriate policies to consider this potential differential impact across different levels of aggregation.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brennan, R. L. (Ed.) (2006). *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger.

Buckendahl, C. W. and Plake, B. S. (2006). Evaluating tests. In S. M. Downing and T. M. Haladyna (eds.). *Handbook of Test Development* (pp. 725–738). Mahwah, NJ: Lawrence Erlbaum Associates.

Camilli, G. (2006). Test Fairness. In R.L. Brennan (ed.). Educational Measurement (pp 221-256). Westport, CT: American Council on Education and Praeger.

Cohen, A. S. and Wollack, J.A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (ed.), *Educational measurement* (4th ed., 17–64). Westport, CT: American Council on Education and Praeger.

Downing, S. M. and Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah: NJ: Lawrence Erlbaum Associates.

Florida Department of Education (2015). Assessment Investigation: February 18, 2015. Retrieved from http://www.fldoe.org/core/fileparse.php/12003/urlt/CommAssessmentInvestigationReport.pdf

Haladyna, T. M. and Rodriguez, M. C. (2013). *Developing and validating testing items*. New York, NY: Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational measurement* (4th ed., 17–64). Westport, CT: American Council on Education and Praeger.

Kolen, M.J. and Brennan, R.L. *Test equating, scaling, and linking: Methods and practice* (2nd ed.). New York, NY: Springer.

Schmeiser, C.B. and Welch, C.J. (2006). Test Development. In R.L. Brennan (ed.). Educational Measurement (pp 221-256). Westport, CT: American Council on Education and Praeger.

Williamson, D. M., Mislevy, R. J., and Bejar, I. I. (Eds.) (2006). *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum Associates.

Impact of Test Administration on FSA Test Scores

Prepared by Harold Doran and Monica Patrin
American Institutes for Research

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

BACKGROUND

The Florida Standards Assessment (FSA) was administered to students across the State of Florida during the spring of 2015 in grades 3–8 mathematics, 3–10 English language arts (ELA), and end-of-course (EOC) tests for mathematics for eligible students enrolled in courses. Assessments in grades 3 and 4 were administered on paper with all other tests administered online.

During the spring 2015 legislative session, Florida House Bill 7069 was signed into law and required an independent study of the psychometric validity of FSA as an assessment of student performance with respect to Florida's academic standards. Alpine Testing Solutions was awarded the contract to conduct the independent investigation. While completing their independent study, Alpine requested that the Florida Department of Education (FLDOE) conduct additional analyses to support their collection of empirical validity evidence. FLDOE coordinated with AIR to conduct and replicate these analyses.

This report is submitted to FLDOE in response to the request to provide validity evidence regarding student test scores from the spring 2015 administration of the FSA. The studies and results presented in this report align with test administration (TA) issues identified by Alpine, and are designed to provide a quantitative exploration of the potential impact of TA issues on student test scores. The primary objective of this report is to assess the degree to which any of the TA issues identified by Alpine may have impacted the quality and validity of student test scores arising from the spring 2015 test administration.

It is important to define what is meant when using the term validity in the context of this report. Test score validity is commonly used to mean the test measures what it purports to measure and that scores arising from the test can be used to support inferences related to the measured construct (ASA, 2004; Kane, 2004; Messick, 1989). This study investigates the degree to which TA issues may have impeded, or potentially even advantaged, a student from achieving the score he or she otherwise would have achieved had the TA issue not occurred.

In the context of this report, validity inferences drawn from FSA test scores regarding the measured construct would be limited if evidence exists suggesting that the TA issues caused a material interference that led to a systematic difference in the test scores for the affected students relative to what would have been achieved had the TA issue not occurred. Plainly stated, the question at hand is whether any of the TA issues interfere with our ability to use test scores as a measure of student performance in mathematics or ELA as measured by the FSA test.

The issues described by Alpine are primarily related to tests administered online, and so this document is focused on test scores derived from the online conditions. Further validity studies are scheduled to be completed by the American Institutes for Research (AIR) during the fall of 2015; hence this is only one of several reports that will contribute to the overall body of evidence as it relates to validity. All such evidence will be compiled into the annual Technical Report that will be made publicly available in December 2015.

Research Questions and General Approach

The preliminary Alpine report identified known issues that occurred during test administration that may have impacted the student test scores. Hence, this report centers on the key TA issues identified by Alpine and provides quantitative evidence of the degree to which test scores differ between students that experienced TA issues and those who did not.

This impact analysis is concentrated on the following key research questions, each of which is aligned with TA issues raised by Alpine.

1. Are the item parameters used to generate student ability estimates impacted by students experiencing test administration issues?
2. Are the observed FSA test scores trends in 2015 consistent with the historical trends observed throughout the state?
3. Is there evidence, that toward the end of the testing window, the Department changed its enforcement of policy, which required that all sessions entered must be completed in a single day?
4. Do students with inadvertent exposure to Session 2 or 3 items perform differently than students with no early exposure to those items?
5. Are the scores of students that completed both test sessions within a single day different from students who completed both sessions on different days?
6. Are other students in a school indirectly affected by any test administration issues experienced by other students within the school?

The analyses and results presented here examine these research questions by comparing differences in test scores or other psychometric characteristics of the FSA between the "affected" students and the "non-affected" students. The term affected is used here to denote that AIR has data indicating which specific students reported an issue or experienced one of the TA issues listed by Alpine. Non-affected students are those who did not experience or report a TA issue.

In our impact analysis, we begin with a statement of the treatment and the counterfactual. In this study, the term treatment is used to mean the test score that student i in the affected condition received, otherwise denoted as y_{it}. The counterfactual is the score that same student otherwise would have achieved had the TA issue never occurred, y_{ic}. We clearly cannot observe a student under both conditions; we have only one of the two potential outcomes for any given student. Students that were affected only have a score after treatment and unaffected students have only one score and never received the treatment.

This problem is framed within the potential outcomes framework (Rubin, 2005; McCaffrey, Ridgeway & Morral, 2004) which defines the treatment effect as $T = y_{it} - y_{ic}$.. Given that we do not observe both sets of potential outcomes we need to establish a basis by which we can infer what score students in the treatment condition would have achieved had the TA issue not occurred. If there is evidence that the scores for students in the treatment condition differ significantly from the counterfactual, we can assume a systematic effect.

With observational data, we cannot simply compare differences in means between affected and non-affected groups on the 2015 FSA outcome score and assume the non-affected group mean can serve as the replacement for the counterfactual. Such an analysis would conflate potential differences related to the test administration issues with any other real, systematic differences due to preexisting heterogeneity between students and schools. Such a simple comparison in outcomes between groups will yield biased results due to the confounding effects of other unobserved characteristics affecting students related to non-random assignment.

In light of this problem, one typical approach with observational data is to use propensity score methods and estimate for each individual the probability of receiving the treatment, $\Pr(T = 1|\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ denotes some set of pretreatment characteristics and then subsequently use the probabilities as weights in ways similar to survey methods that weight observations to account for the unequal probability of selection into the sample. The limit of this approach is that the set of pretreatment characteristics, $\boldsymbol{\gamma}$, must be related to student assignment to the affected condition and in the current scenario no such covariates are known to have impacted whether a student experienced a TA issue or not.

An alternative approach when using observed data with non-random assignment to a treatment is to condition on a variable, such as a prior test score, x, to account for between student heterogeneity. In this way we can formalize our definition of treatment effects from data that are observed. Let $E(y_{it}|x = x_i)$ denote the conditional mean of students in the affected sample, and then $E(y_{ic}|x = x_i)$ denotes the conditional mean for students in the unaffected sample. We can then specify $E(y_i|x = x_i, z)$ where z is an indicator variable capturing the difference in means between the affected and unaffected groups. In this way, the outcome, y_i , is compared only between students sharing a common value of x_i. This framework situates the analysis within the context of a quasi-experiment (Campbell & Stanley, 1963; Cook & Campbell, 1979) using a pretreatment variable collected prior to FSA administration and offers a clear definition of the treatment effect under investigation.

DATA

There are four primary sources of data used to examine the research questions in this study. All data and program code used in this study has been made available to FLDOE to independently verify and replicate our results.

The first is the outcome test score of students from the spring 2015 FSA administration. FLDOE has yet to establish a linear transformation for this metric (as of writing this report) and for this reason, this metric is expressed on the item response theory (IRT) person ability metric. This is commonly referred to as a theta score and is typically distributed as approximately unit normal with mean zero.

For the IRT recalibrations, the data are the student-level item responses to each item. The same data files and input command files used to originally calibrate and score students are reused and modified only to remove affected students as described in this report.

Under a separate contract with FLDOE, AIR is also the provider of value-added modeling (VAM) services, a statistical model that uses longitudinal data to evaluate a teacher’s impact on student scores. For this reason, we also house the prior year Florida Comprehensive Assessment

Test (FCAT) scores and student identifiers that can be merged with the current year FSA outcome score and subsequently serve as the control variable. Some attrition naturally occurs when merging cross-year files to build longitudinal records for students. The data files used here reflect students that can be merged from 2014 and 2015 by combining all data from the current year with all students that were included in the 2014 value-added longitudinal database housed by AIR.

Finally, AIR was able to capture specific students affected by the test administration issues and has previously provided these lists to FLDOE and Alpine. For instance, our systems capture those students that completed the entire test in a single day or those who inadvertently entered into Session 2 items. These lists are the basis for grouping students into the affected and unaffected groups in the analyses that are reported in this study.

TECHNICAL METHODS

IRT Recalibration

All FSA student test scores are derived from the item parameters for the core (operational) items on the test. The item parameters are estimated using post-equating (Kolen & Brennan, 1995) on a subset of students in the population, referred to as the early processing sample (EPS). The EPS subset of students is a scientifically representative sample identified using a stratified random selection approach and contains approximately 15% of the tested population in each grade. The item parameters used for scoring are derived from the EPS for all tests except for grade 8 mathematics and each of the end-of-course tests, in which case entire populations are used.

To examine the degree to which the estimates of the item parameters are impacted by students experiencing TA issues, all students identified as affected by the TA issues are removed from the original data files used for calibrations. This modified data file is then used to estimate new item parameters. Our approach to compare and summarize the complete set of item parameters from the original and recalibration is to plot the test characteristic curves (TCC) for each test. If the item parameters from the original scoring set are the same under the recalibration, then the TCCs will be superimposed. If any systematic differences exist as a result of the TA issues, then the TCCs will show discrepancies at certain points along the score distribution between the original and recalibrated data and we can further explore any identified discrepancies.

The purpose of this analysis is to explore whether FSA test scores derived from the item parameters are impacted by the TA issues. If the TCCs overlap, then scores derived from the original calibrations remain unaffected. If the TCCs show some discrepancies, then the FSA scores may have been impacted by the TA issues.

Stability Analysis Correlations

To examine the trends and stability of the FSA test scores we estimate the observed correlation in test scores between 2013–2014 and 2014–2015 as students transition between grades g and $g + 1$. We then estimate the same correlation between 2012–2013 and 2013–2014. The observed correlation between 2012–2013 and 2013–2014 is referred to as the baseline correlation and is an indication of the typical trend observed in the state as these scores are collected prior to AIR delivering tests in the state. The observed correlation between 2013–2014 and 2014–2015 is the current year correlation and is used to infer how the current year trend compares to historical trends.

If test administration issues observed during the spring of 2015 impact test scores, then we would expect the current year correlation to differ markedly, and presumably be lower than, the baseline correlation. A lower current year correlation when compared to the baseline correlation would suggest anomalies in the 2015 FSA scores. On the other hand, if the baseline correlation and current year correlations are similar in magnitude, then we can assume that trends in scores observed during spring of 2015 are no different than trends previously observed in the state.

Linear Regression Models

To examine differences in the means between affected and unaffected groups we use a linear regression with a fixed effect for group membership. The general form of the linear model used here is

$$y_{ij} = \mu + x_{ij}\gamma + z_{ij}\beta + e_{ij}$$

where x_{ij} is the prior year FCAT 2.0 score for the *i*th student in the *j*th school to control for preexisting differences between students, $e_{ij} \sim N(0, \sigma^2)$, and z_{ij} is a binary variable denoting group membership such that

$$z_{ij} = \begin{cases} 1 \text{ if student } i \text{ is in affected group} \\ 0 \text{ otherwise} \end{cases}$$

While the model appears as a least squares regression, the variance/covariance matrix of the fixed effects is instead estimated with consideration of the complex sampling design. Least squares standard errors would underestimate the true sampling variance as students within common groups share a common group effect. This common group term induces a design effect larger than 1 as a result of the non-zero intra-class correlation and thus requires design-consistent standard errors. For this reason, the variance/covariance matrix estimated here yields design-based standard errors (Kish, 1965) where schools are treated as the cluster variable to explicitly account for common group membership.

The binary variable, z_{ij}, is used to estimate the coefficient, β which is an indication of the difference in means between unaffected and affected students. This model coefficient is the primary parameter of interest in this study and the null hypothesis of no difference can be supported when $\beta = 0$.

The parameterization of the regression model above implements a student-specific flag to account for TA issues at the student level. The Alpine report suggests there could be a larger "environmental" concern indirectly affecting all students within the same group even if only some students within the group were impacted. For instance, perhaps some specific students in a school inadvertently entered into Session 2 too early and other students in the same school did not. It may be plausible that the affected students shared information about the test items with other students in the same school, thus indirectly affecting all students in the school.s

For this reason, we extend the model above and estimate a regression model of the same form. But instead we code the binary variable as

$$z_{ij} = \begin{cases} 1 \text{ for all } i \ \in j \text{ if } i \text{ is in affected group} \\ \qquad 0 \text{ otherwise} \end{cases}$$

In this way, we can examine the aggregate, indirect effect of test administration issues on all students within a school even if onlsy a subset of students within the school experienced or reported a test administration problem.

Results

IRT Recalibration

Table 1 provides the number of students removed from the EPS used to recalibrate the item parameters in all ELA and mathematics tests. In this analysis, we remove students affected by each issue listed in the table and recalibrate the item parameters using the remaining set of students. The value "EPS N" is the total number of students used in the original calibration used for operational scoring and the value "recalibration N" is the number used for the recalibration after dropping students for the three reasons listed in the table, where applicable.

Table 1. Students Removed from the EPS Used to Recalibrate Item Parameters

Test	EPS N	Reported Writing Issue	Entire Test One Day	Preview Session 2	Recali-bration N
Grade 5 Math	26,156	0	42	45	26,069
Grade 6 Math	25,588	0	116	17	25,455
Grade 7 Math	23,519	0	115	72	23,332
Grade 8 Math	116,747	0	1278	395	115,074
Algebra 1	201,246	0	0	69	201,177
Algebra 2	155,465	0	1980	173	155,292
Geometry	191,801	0	2419	141	191,660
Grade 5 ELA	27,427	0	123	54	27,250
Grade 6 ELA	27,200	0	211	77	26,912
Grade 7 ELA	27,071	0	136	74	26,861
Grade 8 ELA	27,747	41	132	31	27,543
Grade 9 ELA	29,955	34	289	164	29,468
Grade 10 ELA	27,874	23	192	33	27,626

Item parameters for the same operational items in each grade were recalibrated using the same software and input command files in IRT PRO as used in the original calibration; the only difference here is the input data file, which removes students identified as affected by the TA issues listed in the table.

The TCCs for the recalibrated data compared to the original calibrations used to derive student ability estimates are provided in Figures 1 through 13. In all cases the TCCs are superimposed indicating that removing students from the affected condition has no impact on the estimates of the item parameters used for operational scoring. The TCCs, in fact, overlap to the degree that only one of the two curves is visible, showing no discrepancies at any score points along the continuum for any test.

Figure 1. G5M Test Characteristic Curves

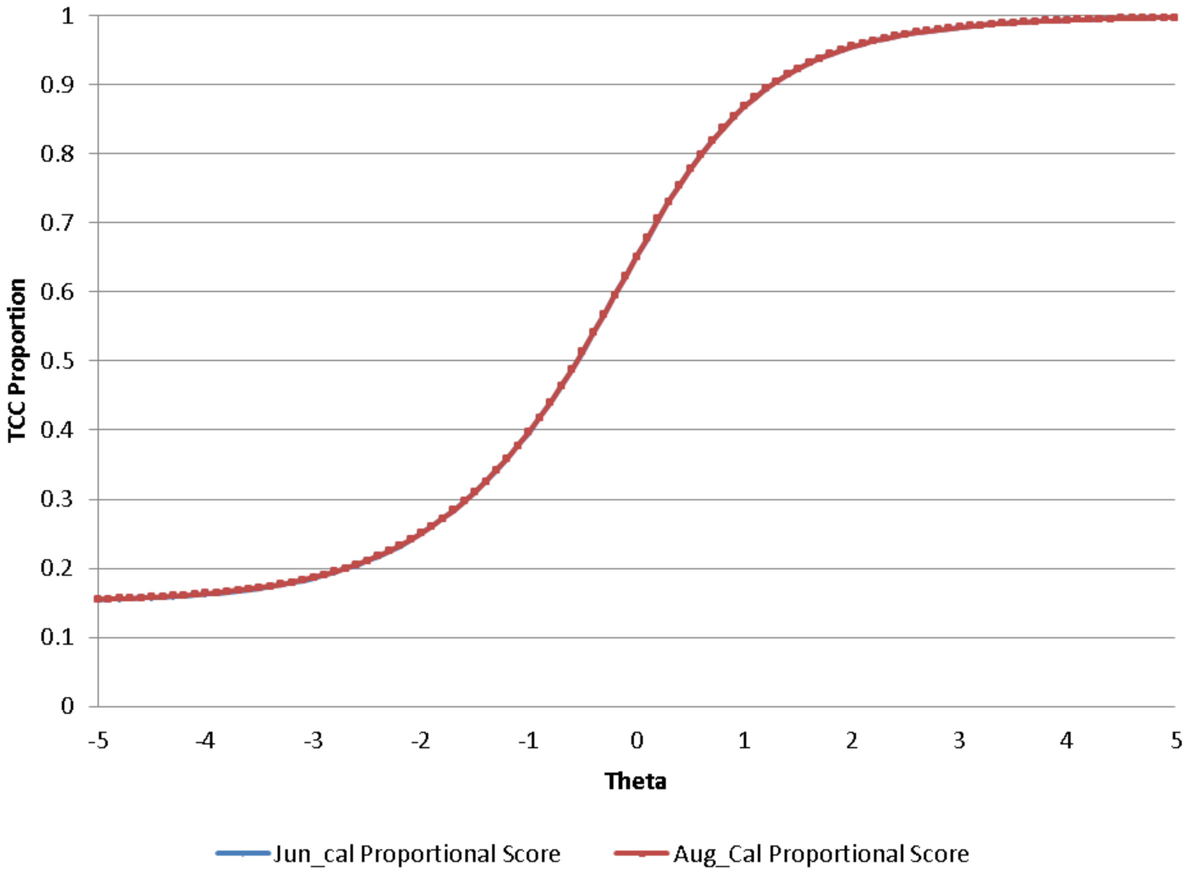


Figure 2. G6M Test Characteristic Curves

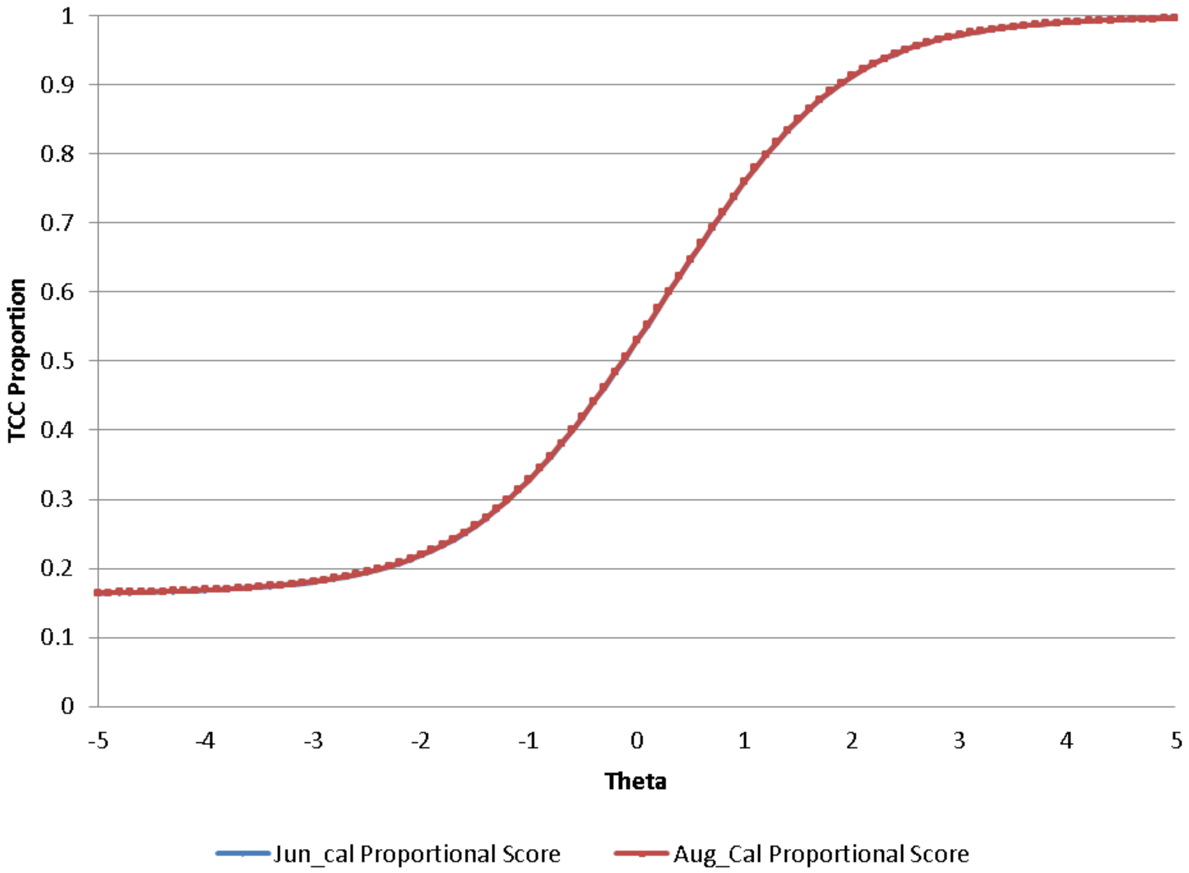


Figure 3. G7M Test Characteristic Curves

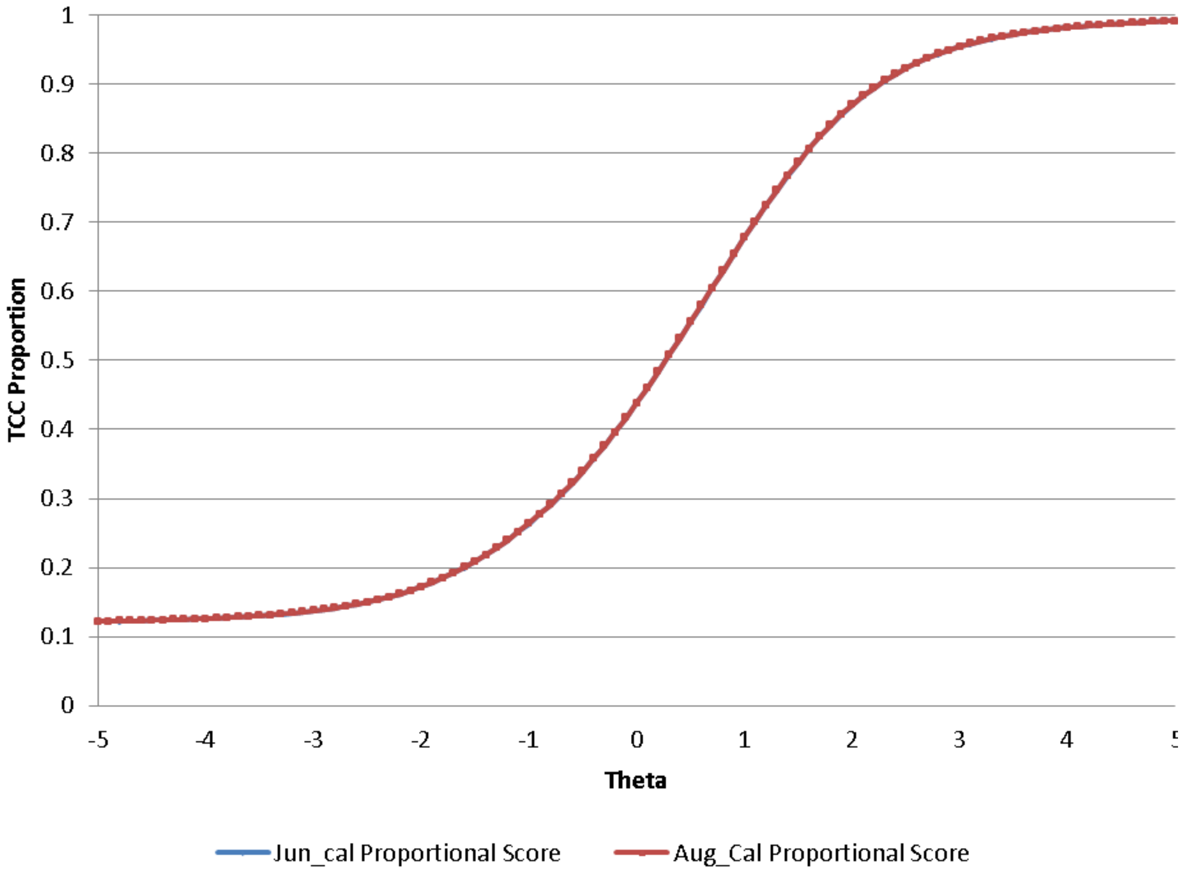


Figure 4. G8M Test Characteristic Curves

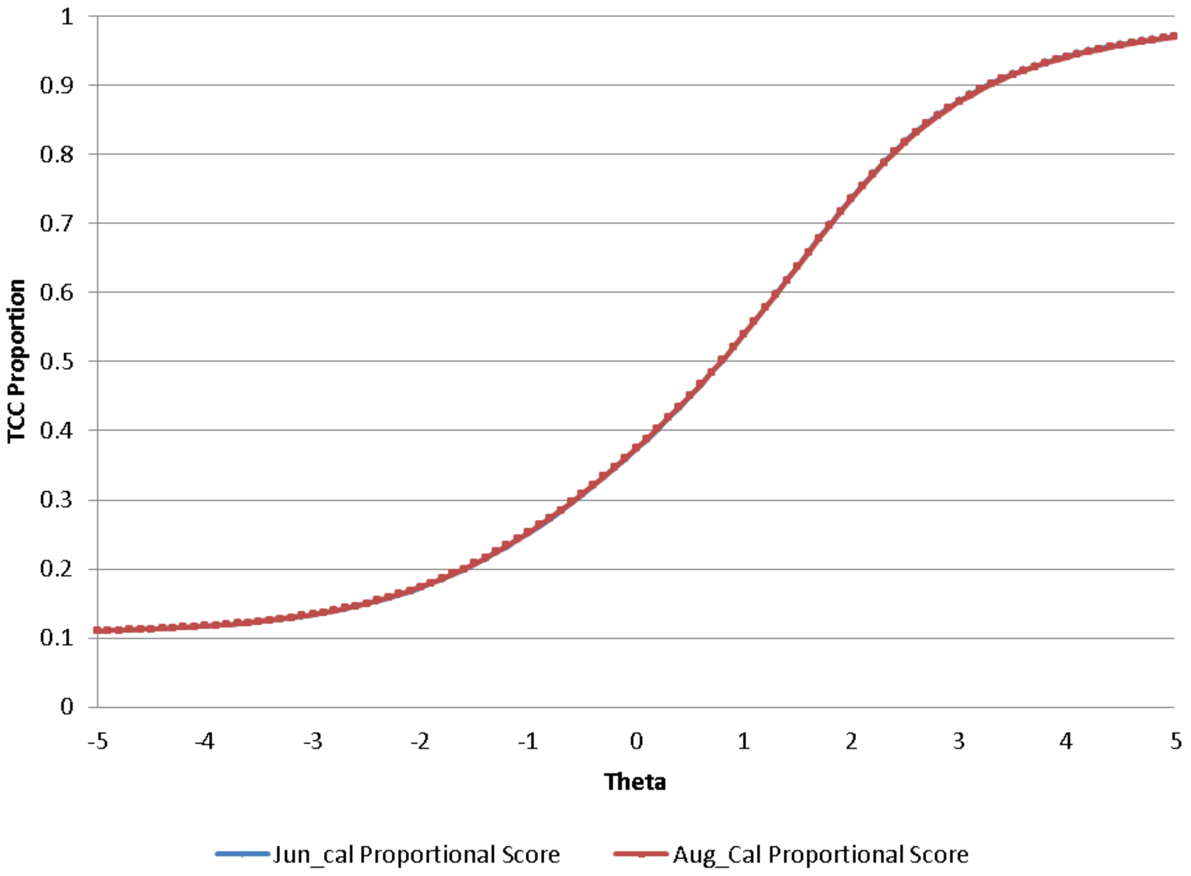


Figure 5. Alg1 Test Characteristic Curves

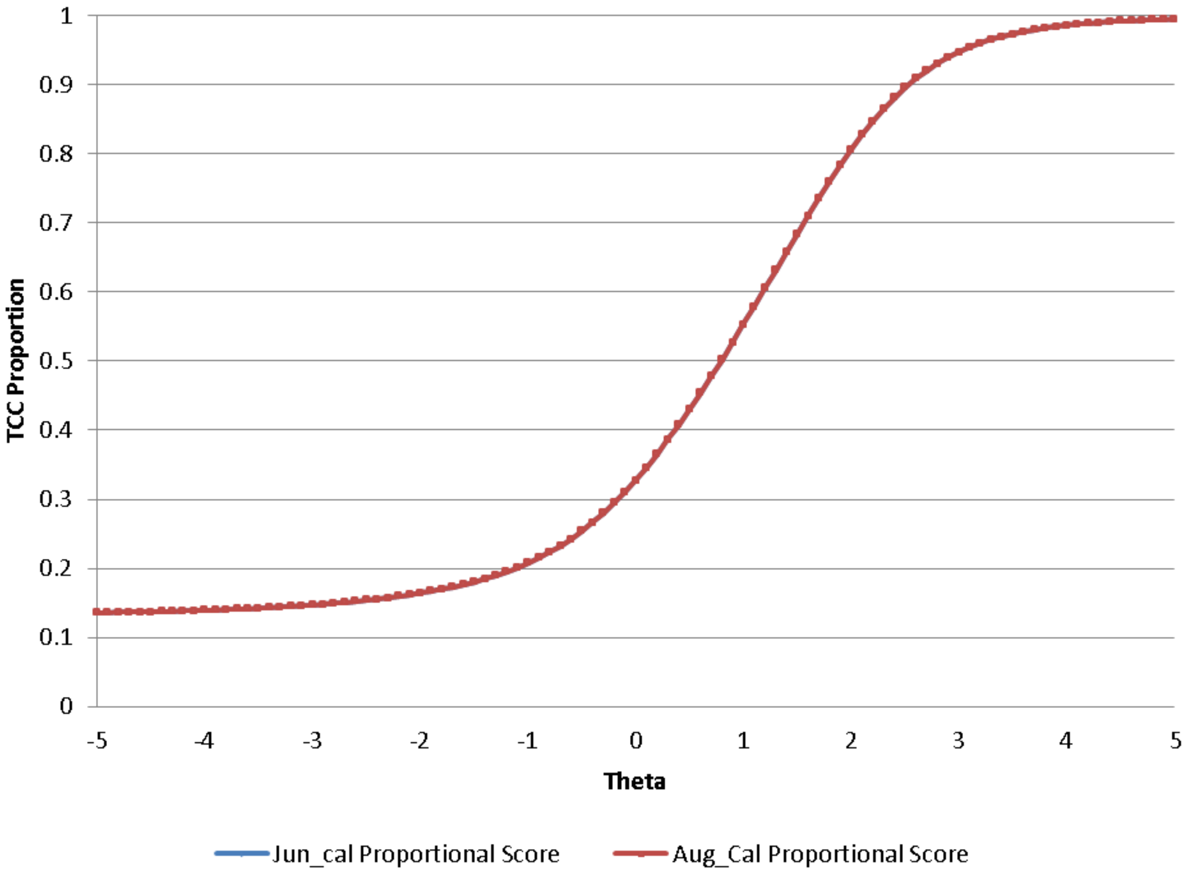


Figure 6. Alg2 Test Characteristic Curves

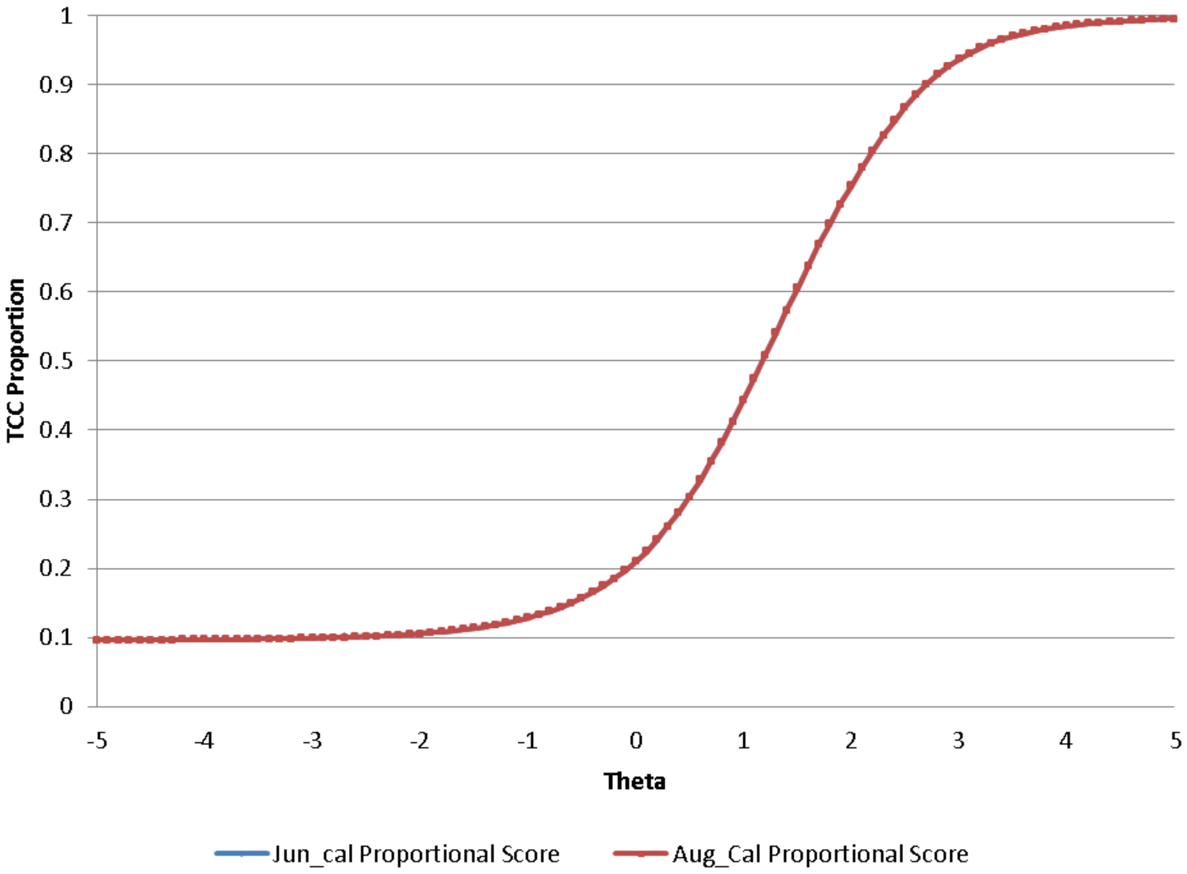


Figure 7. Geo Test Characteristic Curves

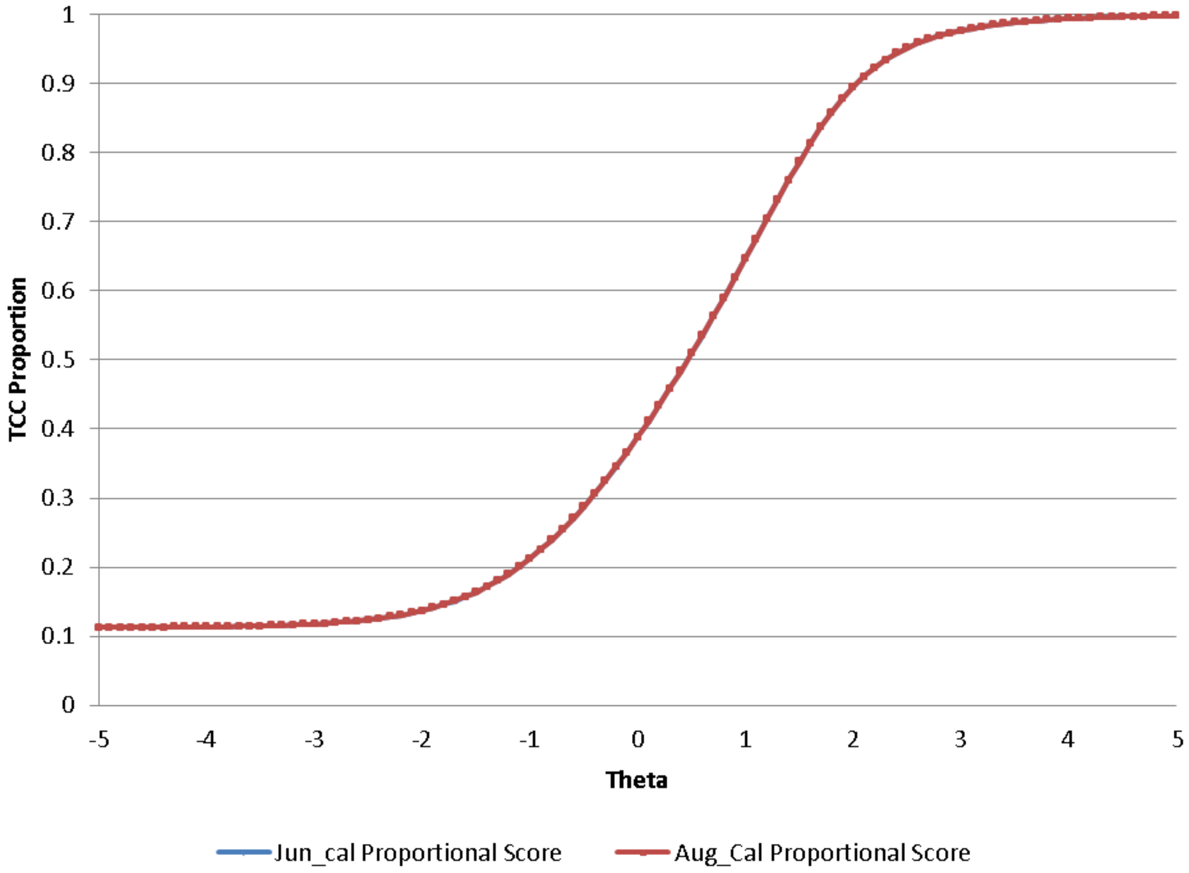


Figure 8. G5E Test Characteristic Curves

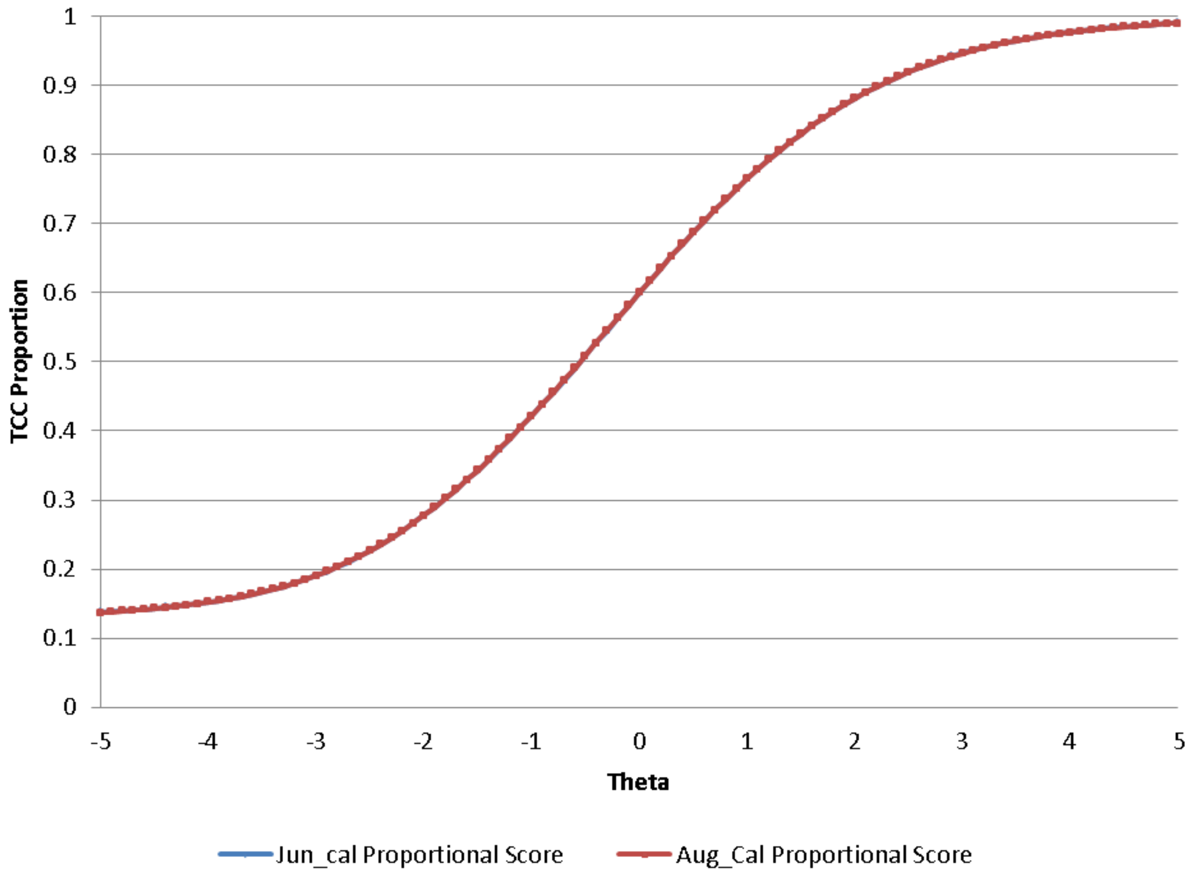


Figure 9. G6E Test Characteristic Curves

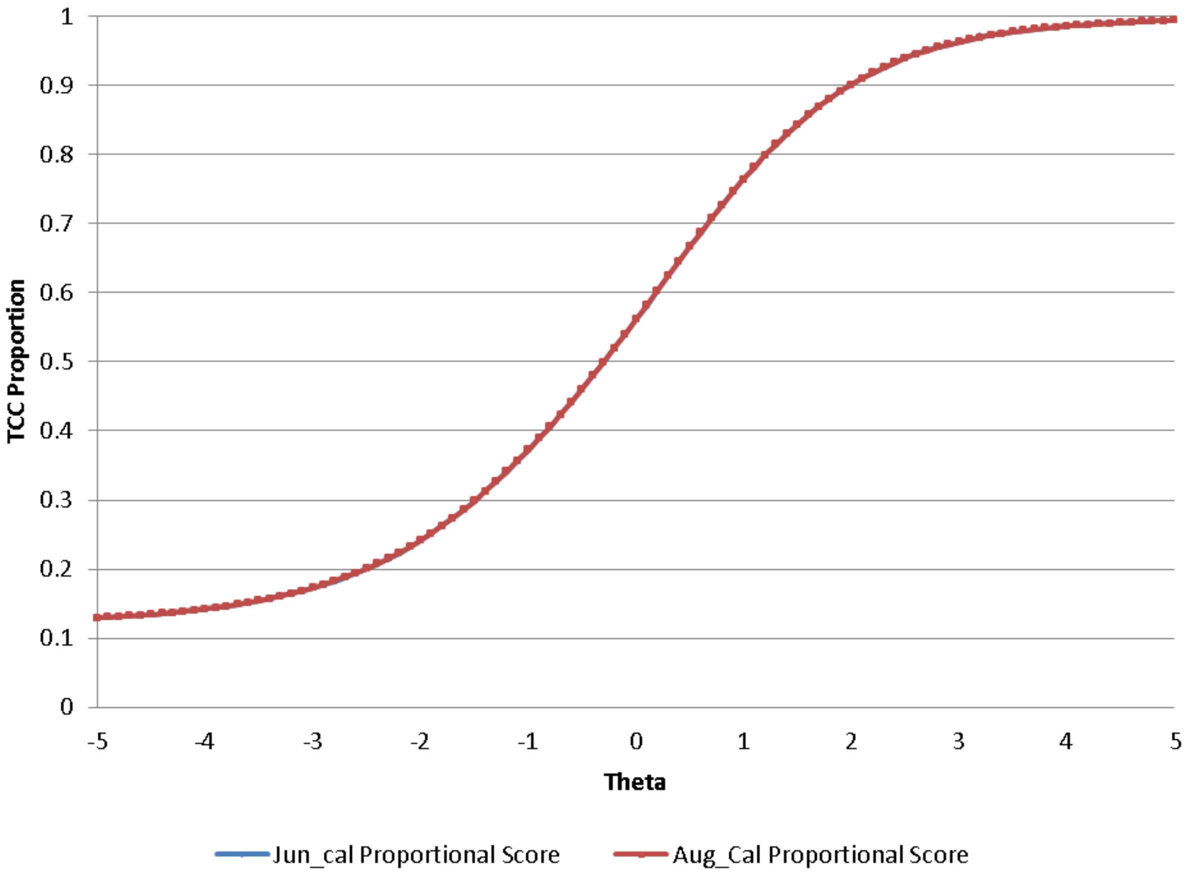


Figure 10. G7E Test Characteristic Curves

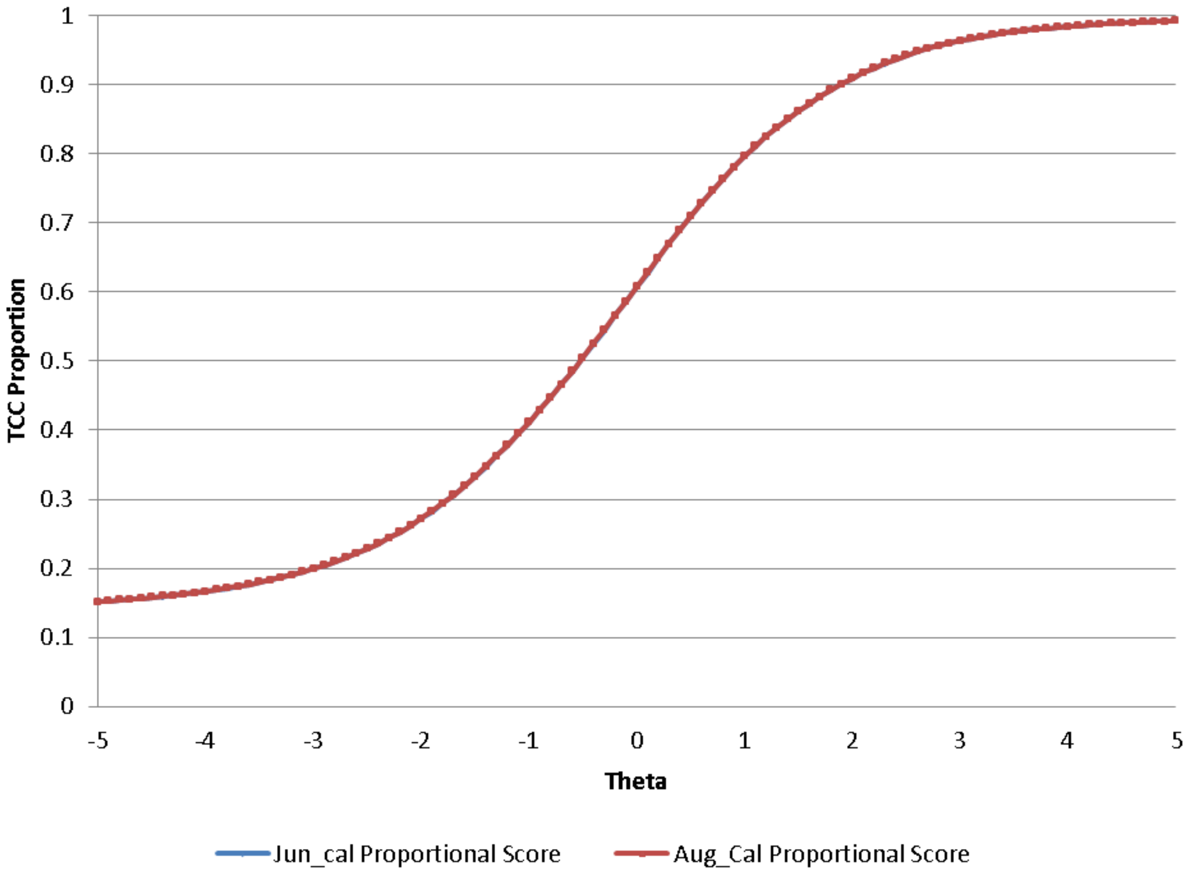


Figure 11. G8E Test Characteristic Scores

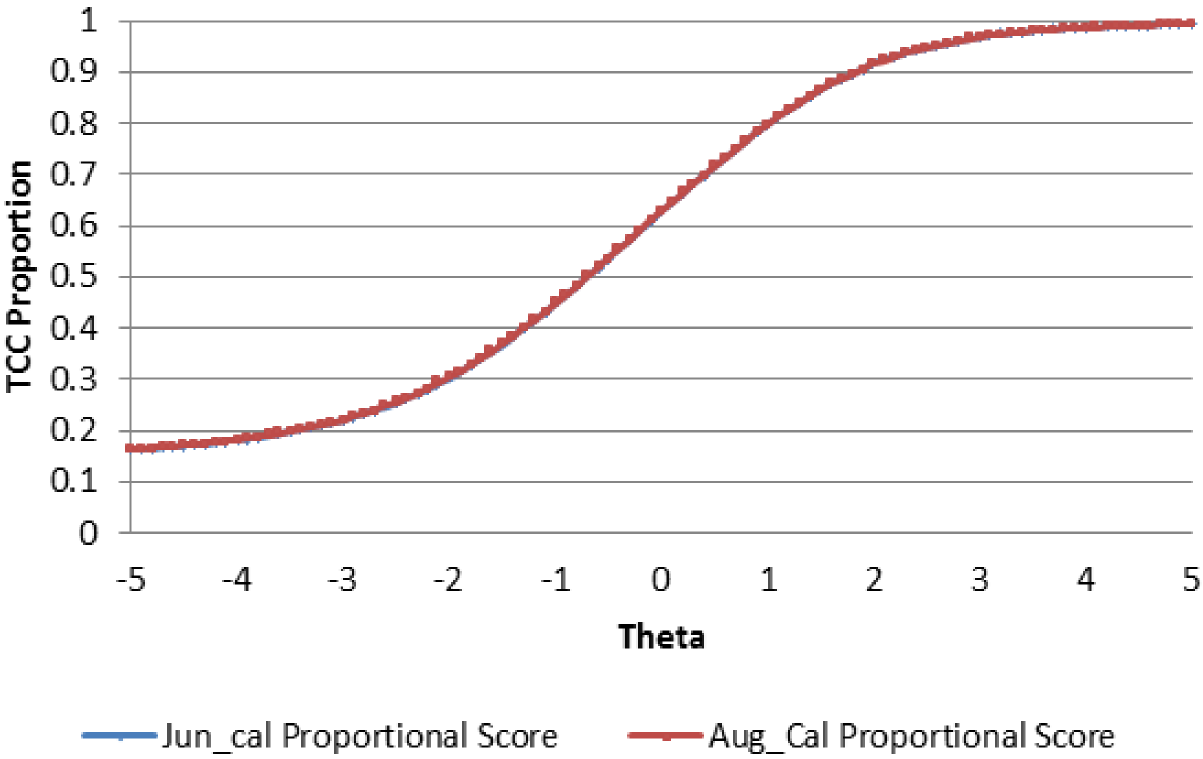


Figure 12. G9E Test Characteristic Curves

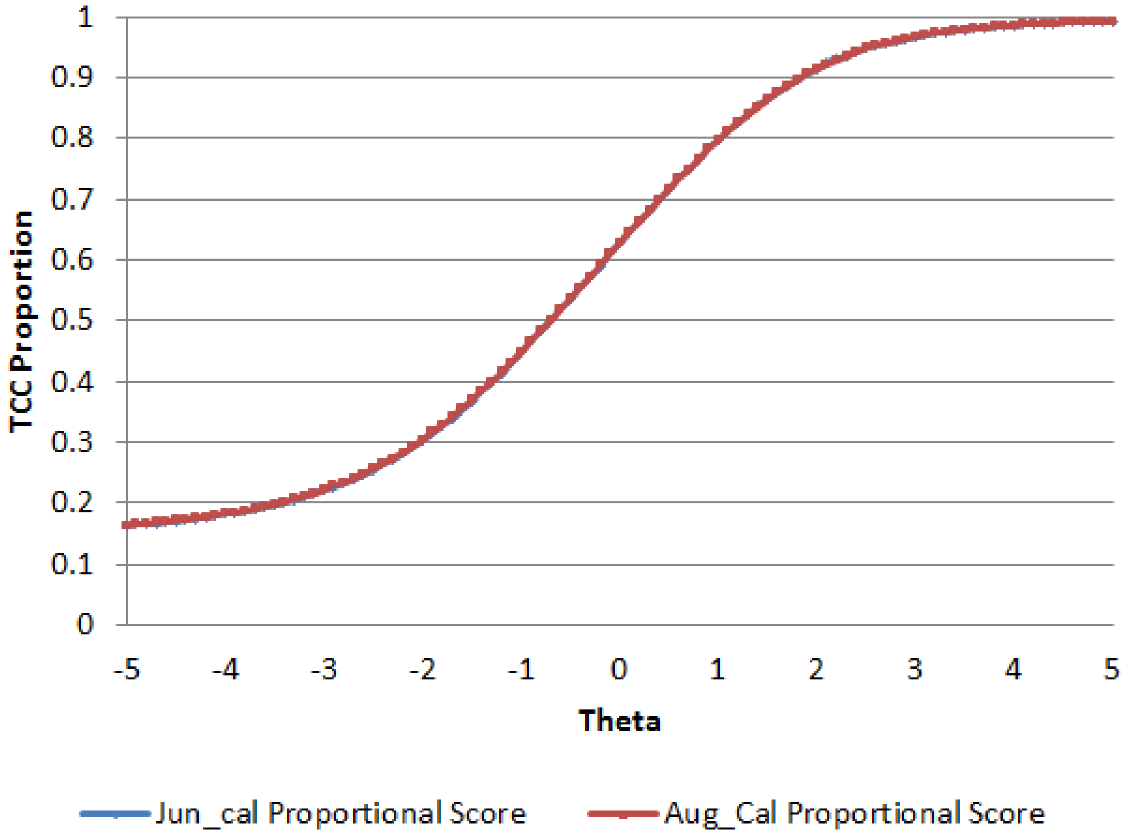
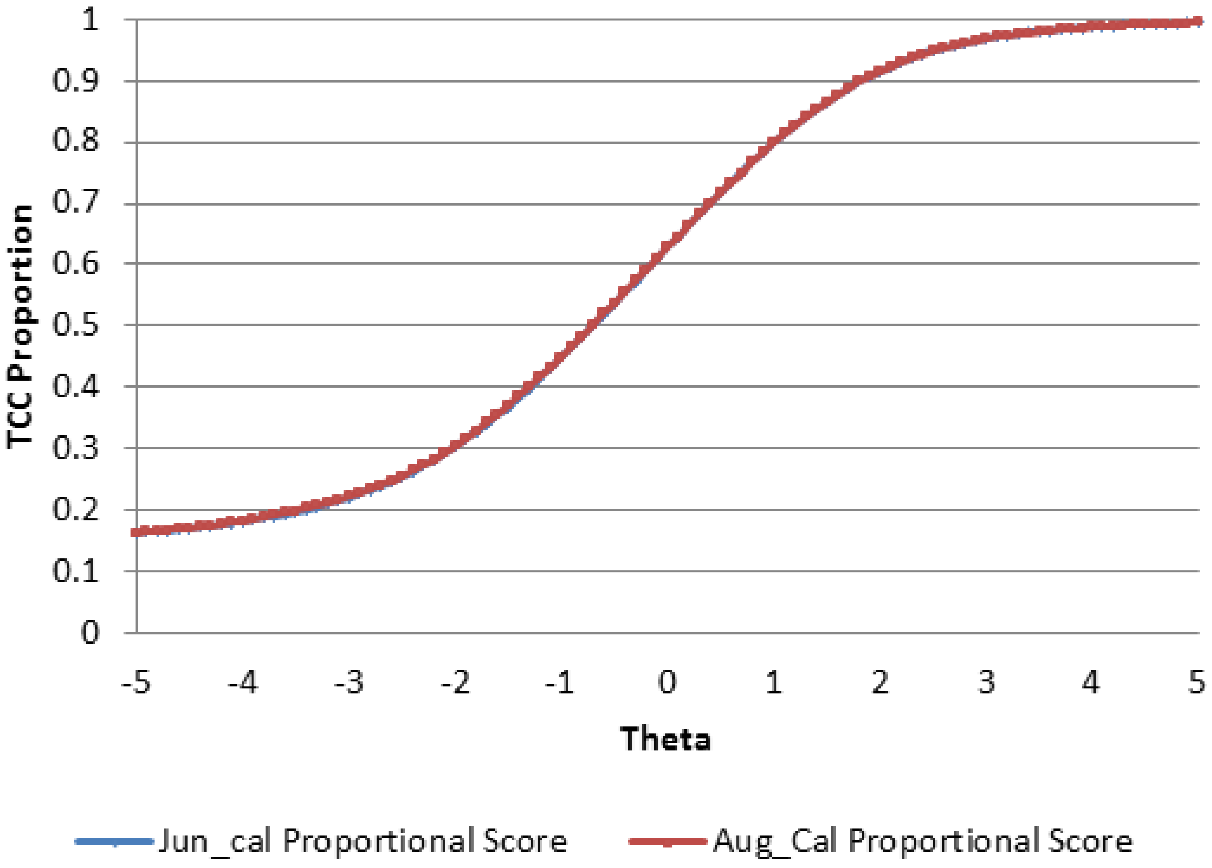


Figure 13. G10E Test Characteristic Curves



Score Stability Analysis

Tables 2 and 3 provide the baseline and current year correlations and the marginal test reliabilities between the ELA and mathematics test scores. These tables shed light on the degree to which scores in the current year align with trends previously observed across the state. In all cases, the current year trends are similar to the trends observed using the historical data.

The current year observed correlations are high and approach the theoretical upper limit for every test. Typically, observed correlations between different measures of a common trait can serve as validity coefficients, but their upper limit is the test reliability for a given test (ASA, 1999; Campbell and Fiske, 1959; Nunnally & Bernstein, 1994). This idea is derived from the fact that test reliability is how scores from the actual test form would correlate with scores that would arise from any other possible, parallel version of the same test. The principle then espoused is that a separate test cannot correlate more highly than a test can with itself.

Given this framework, we can use the baseline correlation as a lower bound estimate and we can then use the test reliability as an upper bound. In all cases but grade 8, we observe the current year correlations fall within the lower and proposed upper bounds.

Table 2. Stability Analysis (ELA)

Test	Baseline	Current	Upper Test Reliability
G4E to G5E	0.80	0.80	.91
G5E to G6E	0.82	0.82	.92
G6E to G7E	0.81	0.82	.92
G7E to G8E	0.82	0.82	.92
G8E to G9E	0.83	0.83	.93
G9E to G10E	0.82	0.82	.92

Table 3. Stability Analysis (Math)

Test	Baseline	Current	Upper Test Reliability
G4M to G5M	0.76	0.79	.93
G5M to G6M	0.79	0.82	.92
G6M to G7M	0.80	0.82	.92
G7M to G8M	0.74	0.71	.88

The grade 8 current year correlation is only marginally smaller than the baseline. However this can be explained via a real-world situation that is increasing in numbers over time. In Florida, students in grade 8 are enrolled in either the grade 8 general mathematics or the Algebra 1 mathematics course. In 2015, roughly one-half of the grade 8 population takes the Algebra 1 test in lieu of the grade 8 mathematics test. Further studies have shown that the students taking the grade 8 math test are lower performing relative to their grade 8 counterparts enrolled in the Algebra 1 course. Consequently, the correlation in this grade is affected more by the changing populations over time and perhaps less by the FSA outcome scores.

Number of Tests Completed Within a Single Day

The administration guidelines state that a student who begins a session must complete that session during the same day. In such case where students would complete multiple sessions during the same day, the guidelines require that the district ascertain from the students and parents whether they felt that the single-day administration impacted performance. If the districts suspect such impact, they may invalidate the test.

District personnel reported to Alpine that FLDOE shifted its enforcement of this rule toward the end of the testing window. If that were the case, we would see a decline in the number of students completing the test in a single day toward the end of the window. This pattern is not evident in the data.

Figures 4 and 5 show the number of tests in which both sessions were completed within a single day during the course of the entire test window. The *n*-sizes are 3,937 and 11,538 for mathematics and ELA, respectively over all tested grades. In reading, a spike appears in the first day with about 3,500 tests completed in a single day. In mathematics, a small upward trend toward the later part of the window is observed, but only for a small number of tests.

In totality, the data do not provide evidence of a shift in statewide policy toward the end of the window. In ELA, the first day of the window saw many students finish the test in a single day, but this pattern is not apparent in mathematics. To the extent that this pattern may be used to infer a policy shift, such a shift would have had to occur following the first day of testing, and somehow have been limited to ELA. What appears as a larger spike toward the end of the window in mathematics represents less than one quarter of 1% of the total tested population across the state. This pattern suggests, if anything, stronger enforcement of the rule toward the end of the window.

Figure 14. Number of Math Tests Submitted in One Day Aggregated Over All Grades

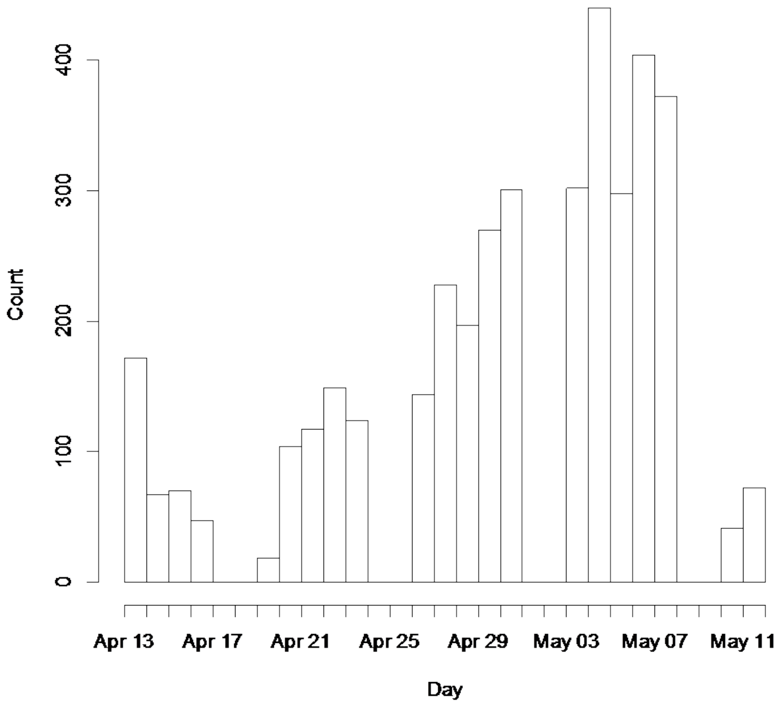
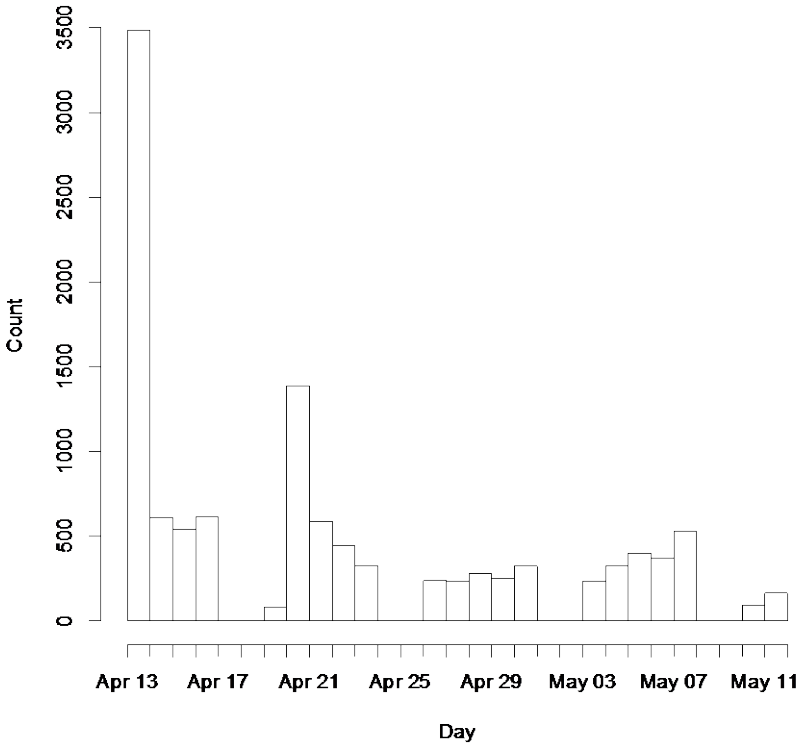


Figure 15. Number of Reading Tests Submitted in One Day Aggregated Over All Grades



Linear Regression Models

Impact of Entering Session 2 Too Early

Tables 4 and 5 provide the results of the linear regression models examining the impact on student test scores for students who inadvertently entered Session 2 of the test too early and previewed the items. By previewing the items early, we might expect for these affected students to perform better on the test than non-affected students, given that they had an advanced opportunity to consider correct answers to these items.

The tables provide the coefficients for all model parameters and their design-consistent standard errors; however, the *t*-statistic is provided only for the parameter of primary interest. The variable "student flag" denotes the binary variable in the regression showing the difference in means between the affected and unaffected student groups.

Here we observe in all cases but one the null hypothesis of no difference is supported. The coefficient is significant in grade 10 ELA, but in the opposite direction of the hypothesis. Here it shows that students with an early preview in Session 2 performed lower than non-affected students. The school level analyses follow a similar trend with negative coefficients, only two of which are significant.

The student results in mathematics (Tables 6 and 7) follow a similar trend, and the school results for mathematics show small, non-significant positive effects. Taking all grades and tested subjects into consideration, there is no consistent, identifiable trend showing that an inadvertent preview of the item advantaged students.

Table 4. Students Who Previewed Items in Session 2 (ELA)

Test	Intercept	SE	Prior Score	SE	Student Flag	SE	t Statistic	N Flagged	N Total
G5E	-8.2	0.03	0.04	0.0001	-0.014	0.06	-0.25	164	106984
G6E	-8.92	0.04	0.04	0.0002	-0.039	0.06	-0.69	422	103863
G7E	-9.2	0.04	0.04	0.0002	-0.079	0.06	-1.33	227	99168
G8E	-8.7	0.04	0.04	0.0002	0.005	0.06	0.08	264	102841
G9E	-8.83	0.04	0.04	0.0002	-0.042	0.05	-0.92	442	101444
G10E	-9.65	0.04	0.04	0.0002	-0.12	0.06	-2.16	258	98503

Table 5. Schools With Students Who Previewed Items in Session 2 (ELA)

Test	Intercept	SE	Prior Score	SE	School Flag	SE	t Statistic	N Flagged	N Total
G5E	-8.2	0.03	0.04	0.0001	-0.0002	0.02	-0.01	6908	106984
G6E	-8.91	0.04	0.04	0.0002	-0.0597	0.02	-3.86	18942	103863
G7E	-9.19	0.04	0.04	0.0002	-0.0261	0.02	-1.54	17408	99168
G8E	-8.7	0.04	0.04	0.0002	-0.0375	0.02	-2.44	12516	102841
G9E	-8.83	0.04	0.04	0.0002	-0.022	0.02	-1.4	32416	101444
G10E	-9.65	0.04	0.04	0.0002	0.0039	0.02	0.24	25431	98503

Table 6. Students Who Previewed Items in Session 2 (Math)

Test	Intercept	SE	Prior Score	SE	Student Flag	SE	t Statistic	N Flagged	N Total
G5M	-8.49	0.04	0.04	0.0002	-0.02	0.09	-0.2	216	107823
G6M	-9.29	0.05	0.04	0.0002	-0.46	0.11	-4.35	177	100497
G7M	-9.57	0.06	0.04	0.0003	-0.06	0.06	-0.91	204	86898
G8M	-10.26	0.1	0.05	0.0005	0.15	0.29	0.51	273	58354

Table 7: Schools With Students Who Previewed Items in Session 2 (Math)

Test	Intercept	SE	Prior Score	SE	School Flag	SE	t Statistic	N Flagged	N Total
G5M	-8.49	0.04	0.04	0.0002	0.01	0.04	0.35	4264	107823
G6M	-9.28	0.05	0.04	0.0002	-0.08	0.03	-3.03	12971	100497
G7M	-9.57	0.06	0.04	0.0003	0.02	0.03	0.98	13445	86898
G8M	-10.25	0.11	0.05	0.0005	-0.1	0.05	-1.86	5552	58354

Impact of Students Completing Tests Within a Single Day

Tables 8 and 10 provide results of the regression models where the student flag denotes whether a student completed the entire test in a single day. The school flag (Tables 9 and 11) investigates the same issue, but evaluates all students within the school.

Here the results show that the mean for students in the affected group is almost always statistically significant and lower than the mean for students in the unaffected group. This difference in means is expected, as the test was designed to be administered over multiple days and instructions were provided by FLDOE to school districts to administer the test in such a manner.

As mentioned above, Department policy requires districts to ascertain whether the student experience was rushed, and to invalidate the test if it was.

Table 8. Students Who Completed the Exam in a Single Day (ELA)

Test	Intercept	SE	Prior Score	SE	Student Flag	SE	t Statistic	N Flagged	N Total
G5E	-8.2	0.03	0.04	0.0001	-0.25	0.04	-6.93	378	106984
G6E	-8.91	0.04	0.04	0.0002	-0.34	0.04	-9.09	809	103863
G7E	-9.19	0.04	0.04	0.0002	-0.32	0.04	-8.09	533	99168
G8E	-8.69	0.04	0.04	0.0002	-0.27	0.06	-4.46	691	102841
G9E	-8.83	0.04	0.04	0.0002	-0.24	0.05	-4.72	859	101444
G10E	-9.64	0.04	0.04	0.0002	-0.25	0.03	-9.49	1005	98503

Table 9. Schools With Students Who Completed the Exam in a Single Day (ELA)

Test	Intercept	SE	Prior Score	SE	School Flag	SE	t Statistic	N Flagged	N Total
G5E	-8.19	0.03	0.04	0.0001	-0.04	0.01	-3.14	11411	106984
G6E	-8.88	0.04	0.04	0.0002	-0.07	0.01	-4.96	36419	103863
G7E	-9.17	0.04	0.04	0.0002	-0.04	0.01	-3.38	30804	99168
G8E	-8.69	0.04	0.04	0.0002	-0.02	0.01	-1.73	33843	102841
G9E	-8.82	0.05	0.04	0.0002	-0.03	0.01	-2.12	48461	101444
G10E	-9.64	0.04	0.04	0.0002	-0.01	0.01	-0.48	55155	98503

Table 10. Students Who Completed the Exam in a Single Day (Math)

Test	Intercept	SE	Prior Score	SE	Student Flag	SE	t Statistic	N Flagged	N Total
G5M	-8.49	0.04	0.04	0.0002	-0.34	0.05	-6.88	216	107823
G6M	-9.29	0.05	0.04	0.0002	-0.41	0.05	-8.47	344	100497
G7M	-9.56	0.06	0.04	0.0003	-0.32	0.05	-6.66	402	86898
G8M	-10.25	0.1	0.05	0.0005	-0.42	0.06	-7.62	439	58354

Table 11. Schools With Students Who Completed the Exam in a Single Day (Math)

Test	Intercept	SE	Prior Score	SE	School Flag	SE	t Statistic	N Flagged	N Total
G5M	-8.48	0.04	0.04	0.0002	-0.03	0.02	-1.4	9855	107823
G6M	-9.27	0.05	0.04	0.0002	-0.07	0.02	-2.86	24469	100497
G7M	-9.55	0.07	0.04	0.0003	-0.04	0.02	-2.09	22779	86898
G8M	-10.27	0.1	0.05	0.0005	0.02	0.03	0.57	18514	58354

SUMMARY

Overall, the quantitative results explored in this study do not reveal any consistent, negative trends impacting test scores in ways that are not already mitigated through existing state policy. Students who completed the entire test designed for a 2-day administration in a single day tended to perform worse than would be expected if they had taken the test over two days. Evidence failed to support the hypothesis that the other test administration irregularities systematically influenced student performance. The quantitative results provided in this report suggest the following:

- Item parameter estimates are unaffected by any of the test administration issues.
- Test score patterns as evidenced in the year-to-year cross-grade correlations are aligned with the same trends historically observed across the state.
- The number of students completing the test within a single day did not appear to increase toward the end of the testing window for either mathematics or reading, and the number of students doing so are relatively small when compared to the total population.
- Test score means of students who inadvertently entered Session 2 are no different than the means of the unaffected group.
- Test score means for students who completed the test within a single day are lower than the scores of students completing the test on separate days.
- There is no evidence of an indirect impact on test scores for students within the same school as affected students when previewing items in Session 2.

The TCCs in the IRT recalibration section show that the item parameter estimates used for operational scoring are the same whether or not students in the affected condition are contained in the EPS sample. This finding suggests that the estimates of the item parameters used for operational scoring do not carry any negative impact of the TA issues and therefore any scores derived from these item parameters also remain unaffected.

The stability correlations provide a useful way in which we can judge the current year trend in scores and how it compares to what has been historically observed across the state. In all cases, the current-year correlations are just as high as prior year correlations, with the exception of grade 8 mathematics, which is largely impacted by changes in the tested population between Grades 7 and 8. This finding provides support for the notion that no major issues impacted FSA scores in ways that caused students to systematically perform better or worse than has been commonly observed for students in the State of Florida.

The regression models shed light on the student-level and environmental impact of the TA issues on student scores. The regression models examining the impact on scores for students who inadvertently entered Session 2 too early shows no identifiable trend across subjects or grades. These results suggest support for the notion that entering Session 2 too early did not advantage or disadvantage students in a systematic way with respect to their FSA test scores.

The second regression model examining the impact on students completing the entire test within a single day does show that these students have lower scores than students completing the test as directed in separate sessions. However, this result is aligned with expectations, as the test was structured with multiple sessions and was intended to be administered on separate days in order to minimize test fatigue. The FLDOE has a policy allowing for student scores from this condition to be invalidated if local education agencies believe students were impacted, thereby mitigating the potential negative effects arising from these student test scores.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Campbell, D.T, & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.) *Handbook of research on teaching.* Chicago: Rand McNally. (Reprinted as *Experimental and quasi-experimental designs for research*. Chicago: R. McNally, 1966).

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Boston, MA: Houghton Mifflin Company.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: Praeger.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Kolen, M., & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer-Verlag.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods, 9*(4), 403-425.

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association, 100*(469), 322-331.

Smarter Balanced Assessment Consortium:

Cognitive Laboratories Technical Report DRAFT

Developed by: The American Institutes for Research

September 5, 2013

Executive Summary

The Smarter Balanced Assessment Consortium conducted cognitive laboratories to better understand how students solve various types of items. A cognitive laboratory uses a think-aloud methodology in which students speak their thoughts while solving a test item. The interviewer follows a standardized protocol to elicit responses and record what a student says. While this one-on-one process is time consuming, the type of information elicited is often difficult to obtain by other means. This report presents the results of a series of cognitive laboratory observational studies. The studies were conducted with small numbers of students in order to gather in-depth qualitative data about how students react to different types of items, formats, etc. Due to the small number of subjects studied and the ad hoc nature of the achieved sample of participants, the findings should be used to point the way to more systematic studies, rather than be cited as an authoritative source of scientific findings.

This executive summary presents the major findings from various protocols. Most protocols were developed at multiple grade bands (e.g., 3, 6, and 11). A grade band is the level of content for which the protocol is targeted. Protocols were usually targeted to answer a specific question in one or more content areas (e.g., ELA, mathematics). Results are organized under topics or questions of interest.

Summary and Findings of Cognitive Lab Results by Research Question

Research Question 1: Do mathematics multi-part selected-response (MPSR) items provide similar information about the depth of understanding by the test taker as do traditional constructed-response (CR) items?

An MPSR item has students select several examples of a correct response rather than just one, as in the typical selected-response (SR) item. The intention of this research question was to see whether the MPSR items provided depth of understanding similar to that provided by CR items. If effective, an MPSR item would be a more efficient way to measure the content measured by CR items. Within a form, parallel items were constructed in both formats and presented to the same students. In the protocols the MPSR and CR items were presented in random order.

This research question sought to address two hypotheses. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis examined whether students who get full credit on MPSR items reveal depth of understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target. In most cases the depth of knowledge (DOK) demonstrated by the student either equaled or exceeded the DOK demonstrated for the CR items.

Students who got full credit on the MPSR items also revealed greater understanding of the material than those who did not obtain full credit. The percentage of students understanding the material is also quite similar for the MPSR and CR items. A typical interviewer comment was, “based on the accuracy of the student's responses to both types of items, it appears that item type is not a factor in determining how well the students respond[s]."

Research Question 2: Do TE item types and multi-part SR items approach the depth of knowledge of CRs?

The question is designed to assess whether different types of technology-enhanced (TE) items approach the DOK of CR items for specific content claim/targets and DOK levels. SR items were also included, where available, as a comparison item format. Comparisons were examined for specific TE item types at specific DOK levels for specific content claims/targets. CR and SR items were matched to specific content claims/targets and DOK 4 items in one of the three formats (SR, TE and CR) appeared in each form. Multiple forms were administered, each form to a different sample of students. It was hypothesized that students responding to items of a specific type would reveal that they are using thought processes consistent with a specific DOK level for items measuring a specific target. Different item types were administered to different students.

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. Two exceptions were two targets in the "select text" item type: "justifying interpretations" (grade band 6) and "analyzing the figurative" (grade band 11). A similar pattern was observed for the matched SR items versus the CR items. The same TE item types had higher percentages than did the CR items, with the exception of the "select text" items for the "writing or revising strategies" target (grade band 7) and the "citing to support inferences" target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following "select text" items:

- "select text" for justifying interpretations, claim 1, DOK 2 in grade band 6
- "select text" for citing to support inferences, claim 1, DOK 2 in grade band 11
- "select text" for analyzing the figurative, claim 1, DOK 2 in grade band 11

For mathematics, the pattern is less clear. The TE item types that showed a higher percentage of students demonstrating thought processes consistent with the DOK level included:

- "placing points" for fractions, claim 1, DOK 2 in grade band 3
- "single lines" for equations and inequalities, claim 1, DOK 2 in grade band 11
- "tiling" for fractions, claim 1, DOK 2 in grade band 3
- "tiling" for equations and inequalities, claim 1, DOK 2 in grade band 11 ("Student indicated use of multiple steps and solved correctly.")
- "vertex-base quadrilaterals" for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

The item types in which the CR items had a higher percentage of DOK-consistent thought processes included:

- "select and order" for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6

- “tiling” for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- “tiling” for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- “tiling” for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

The TE item types for which a higher percentage of students received full credit included only:

- “tiling” for equations and inequalities, claim 1, DOK 2 in grade band 11, and
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4.

In other cases the percentage of students receiving full credit was lower than for the comparable CR items. It should be noted that the percentage receiving full credit was generally low in mathematics. Even the matched SR items generally did not perform any better than either the CR or TE items.

Research Question 3: The Impact of Labeling on Mathematics Multi-Part Selected-Response (MPSR) Items: For multi-part selected response (MPSR) items where students may select more than one answer choice, which wording best indicates to the student that he or she is allowed to select more than one option? For multipart (e.g., YES/NO) dichotomous choice items, do students know that they need to answer each part?

Smarter Balanced sought to investigate whether students might become confused with MPSR items in mathematics and perhaps not complete the entire item. In order to investigate this, items were constructed with different amounts of labeling. *Labeling* is the identification of the parts of the problem with indicators such as “a,” “b,” “c” or “1,” “2,” “3.” A labeled and a non-labeled condition were investigated. An example of an item in the labeled and unlabeled format can be found in Exhibit 2.

This question is designed to assess whether labeling or not labeling an MPSR mathematics item produces a difference in performance. Results are reported in five grade bands. Each form contains one MPSR item followed by one CR item. The labeled and non-labeled items appeared in different forms of the test and thus were taken by different students.

Even though the labeling of MPSR items was intended to clarify the mathematic tasks for the students, in many cases it actually seemed to confuse the students. Little difference was observed between the labeled and non-labeled items in the lower grade bands (grade bands 3–6). However, students in grade band 7 tended to score higher with non-labeled items. Also, grade band 7 and 11 students tended to be confused by the labeling. In addition, the labeled items tended to receive more comments related to not understanding the instructions. The interviewer confirmed this, suggesting that the grade band 7 and 11 students better understood the instructions in the non-labeled condition than in the labeled condition.

Research Question 4: Does the ability to move one or more sentences to different positions provide evidence of students’ ability to revise text appropriately in the consideration of chronology, coherence, transitions, or the author’s craft?

Smarter Balanced is considering using items that have students reorder sentences to measure an editing/revising standard. Claim 2 of the standards states that students should be able to revise one or more paragraphs demonstrating specific narrative strategies (use of dialogue, sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to the purpose of the item (closure, detailing characters, plot, setting, or an event).

This question was designed to assess whether students' movement of one or more sentences to different positions provided evidence of students' ability to demonstrate consideration of chronology, coherence, transitions, or author's craft. Six ELA items were included in a test form.

Students who performed well on the items were more likely to consider the targeted writing skills (consider chronology, coherence, transitions, and author's craft) when answering the questions. Also, students who made appropriate sentence moves were more likely to consider the targeted writing skills than those who made inappropriate sentence moves. A high percentage of students considered chronology, coherence, and transitions; however, they were less likely to consider author's craft.

Research Question 5: Do Students Who Construct Text Reveal More Understanding of Targeted Writing Skills Than Students Who Manipulate Writing Through the Manipulation of Text (MT) Tasks?

Many believe that the best way to measure writing is to have students write. However, in a testing environment, it is often difficult to adequately sample the writing content domain with an assessment composed exclusively of CR items. An effort is ongoing to find items that are efficient but that can adequately measure the components of the writing domain, thus allowing a broader selection and greater number of items to be delivered. The question examines whether comparable understanding of the targeted writing skills can be achieved using a set of MT tasks in comparison to comparable CR tasks. Examples of the item types can be found in Exhibit 2.

Four pairs of ELA items were developed. Each pair contained one MT item and one CR version of the same item. Two forms were created, and each form contained a single version of an item. Each form contained two MT items and two CR items. The MT items were almost exclusively "select and order" items, though two items—one in grade band 3 and one in grade band 11—were "reorder text" items. All items assessed claim 1, target 1.

The results showed that the targeted writing skills are considered by students who manipulate text at a level comparable to (or greater than) that encountered when they are constructing text. The grade band 3 and 6 students showed comparable (or greater) levels of understanding when the items were in an MT format. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft. Score distributions were comparable for MT and CR item formats.

Research Question 6: Do different types of directions (minimal, concise or extensive) have an effect the performance of technology enhanced (TE) items in ELA and Mathematics?

The optimal amount of direction that should be given to a student working with TE items is unclear. With minimal directions students may not know how to approach an item; with extensive directions students may be distracted or slowed to a point where the item becomes inefficient. This may be particularly true with elementary school students, who may take longer to process text. This question examined this issue for ELA and mathematics items. Three types of directions were used (minimal, concise, and extensive).

In most cases in ELA the level of instruction did not make a difference. For most grade bands and item types, neither the level of instruction nor the item type showed a differential effect in ELA. Cases in which differences were observed included “select text” items when the directions were “concise.” With the “reorder text” items the grade band 3 students did less well with minimal directions. The grade band 11 students also had some difficulty with the “reorder text” items when the directions were “extensive.”

In mathematics, the level of instruction also did not make a difference for many item types and grade bands. “Select and order” items were difficult (grade bands 6 and 11) regardless of the direction type, however, no direction type proved better than another. High percentages of students received full credit on “select defined partition” and “straight lines” items; however, the direction type did not make a difference. Finally, “tiling” items were generally difficult, but no benefit was shown for different types of directions. Differences were observed in items including “placing points” items under the minimal and concise directions in grade band 11; however, under extensive directions all students received the maximum score. With “placing points and tiling” items a higher percentage of students received full credit with fewer instructions (grade band 6). Finally, “vertex-based quadrilateral” items seemed to benefit from minimal directions in grade band 11.

When asked if they had difficulty using the computer, ELA students, in grade band 3, under minimal directions, said they had trouble with both select text and reorder text items. The ELA grade band 11 students also seemed to have some difficulty with the “reorder text” items. Since these are related to specific item types it suggests that there was uncertainty about how to perform the task, rather than using the computer itself. Mathematics students did not seem to have any problems using the computer.

Research Question 7: Smarter currently intends to administer the passage first, and then administer the items one item at a time. Does this affect student performance?

Smarter Balanced is interested in the possibility of administering items adaptively within a passage. This would require administering items sequentially so that the ability estimate could be updated after each item. Presenting items one at a time may take longer, and students may object to not knowing what is coming next. This question is designed to assess whether administering an item set takes longer when the items are presented sequentially and whether there is a difference in confusion or frustration level when students are presented a passage and all the items together or are presented a passage with the items then being presented one at a time. The item sets were not administered adaptively.

Two sets of items were created for a given test form. Both sets contained passages of equivalent length and difficulty as well as items of equivalent difficulty.[1] The first set in a form presented the passage with all the items together. The second set presented the passage with the items presented one at a time.

The forms were administered, within grade band, to different samples of students. Each sample contained both a general education group (Gen Ed) and a group that received English language accommodations (ELL) students. One sample was timed without thinking aloud during the administration. Each item set in these forms was separately timed. This sample provided timing information only. The second sample involved thinking aloud while responding to the questions and was not timed.

The primary questions of interest were:

1. Does presenting the items individually after the passage appear to take longer (timed condition)?
2. Does presenting the items individually after the passage increase the student's negative emotional states (e.g., frustration, confusion; think-aloud condition)?
3. Do students prefer one approach or another (think-aloud condition)?

The time it took to complete the sets when all items were presented together or one at a time varied by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items were presented at once or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer for both the Gen Ed and ELL samples. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show the same pattern within grade band.

There appears to be slightly more *confusion* for both the Gen Ed and the ELL samples in grade band 3 when all the items are presented together. However, similar *frustration* levels were observed under the two formats for the grade band 3 students. Students working on the grade band 6 ELL sample showed similar patterns of frustration and confusion in both presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time.

[1] Comparable passage difficulty was achieved through the use of readability and lexile measures. Comparable item difficulty was achieved through DOK measures.

The grade band 6 students tended to score higher when the items were presented all at once (for both the Gen Ed students and the ELL students). The grade band 3 students showed similar results, regardless of sample or administration format. The grade band 11 Gen Ed students scored higher when the items were presented one at a time, while the grade band 11 ELL sample students scored higher when the items were presented altogether.

Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. Grade band 11 students had a slight bias toward having the items presented one at a time. Conversely, grade band 6 students preferred to have the items presented together.

Research Question 8: Smarter intends to present relatively long passages. Do longer passages reduce student engagement?

Smarter Balanced is interested in using passages that are longer than those presently used. The Smarter Balanced recommended passage lengths are: for grades 3–5: 450–562 words for short passages and 563–750 words for long passages; for grades 6–8: 650–712 words for short passages and 713–950 words for long passages; and for high school, 800–825 words for short passages and 826–1100 words for long passages. There is concern that the longer passages may tax the processing abilities of ELL students and students with disabilities (SWD).

This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Each form contained a standard-length passage and an extended-length passage. The first set contained a passage of standard length. The second set contained a passage that is longer than standard length (extended-length, the length equivalent to that intended for use by Smarter Balanced).

The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands: 3, 6, and 11. Twelve students took the forms. Of these, nine were grade band 3 Gen Ed students and one grade band 3 student was classified ELL/SWD. The single grade band 6 student was an ELL/SWD student. The two grade band 11 students were Gen Ed students.

All the ELL/SWD students were unaffected by the use of the longer passage. They were able to read the entire passage regardless of passage length and demonstrated that the longer passage was processed at a deep level. The ELL/SWD students also were not bored or distracted while reading either passage.

On the contrary, Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. About 75 percent of the grade band 3 students and all of the grade band 11 students were affected by the use of the longer passage. Only 43 percent of the grade band 3 Gen Ed students and 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing. Also, some percentage of the Gen Ed students were bored, regardless of the length of the passage

Research Question 9: How long does it take for students to read through complex texts, performance tasks, etc.? Is timing affected by the way students are presented the passage and items?

One way of making items more difficult is to increase their complexity. Complex items often take longer to solve or answer. In computer adaptive tests, added complexity may decrease the time a high ability student has to complete the test if the items are made more difficult through increased complexity. This potentially creates some fairness issues in an adaptive test if there is a time limit on the test. This question was designed to assess the time it takes for students to answer complex and simpler items. Complexity was defined as a function of the DOK demanded by the test question. It was hypothesized that more complex tasks would take more time.

Each ELA form had six items. These items varied in item complexity (simple or complex) and item format (SR, TE, or CR). The TE items were all “hot text” (HT) items. These items require the student to either highlight the text or drag the text to answer the item.

Forms were constructed in ELA at two grade bands: grade band 3–5 (referred to as grade band 3) and grade band 6 and 7 (referred to as grade band 6). Two forms were administered in grade band 3. One form was administered in grade band 6.

It was hypothesized that more complex items would take longer to complete than simpler items, but no evidence was found to support this hypothesis. SR items were answered in the shortest time. HT items took about one minute longer than SR items. CR items took the most time to answer, about 75 seconds longer then the hot text items.

Research Question 10: Working mathematics problems on computer: Communicating mathematics on computer—feasibility of measuring student understanding of items for Claims 2–4 on computer.

With paper tests some students write in their test books while working out mathematics problems. When mathematics items are presented on computer, scratch paper is often provided if students want to transfer the problem to paper and work it out there. Because scratch paper is often destroyed after an online testing session, the degree to which scratch paper is used is not known; neither is the importance of scratch paper in working out a problem (or potentially for use in scoring). This research question examines the need for paper when solving mathematics problems.

Each student was presented with three grade-appropriate items. The interviewer recorded whether the student made a comment, and the nature of the comment, while working the mathematics problems. The students first tried to work a problem without paper. Scratch paper was then offered to the student to rework the problem, if desired. The interviewer noted whether students chose to add anything additional and noted the nature of the addition (more text, equations, graphics). Note that there were only three comments for the third item in the lowest grade band, 3.

The general conclusion is that a subset of students benefit from being able to work mathematics problems on paper. This appears to be especially important when students are beginning to learn algebra concepts.

Grade band 3 students did not need paper to work the problems. However, in the grade band 6 and grade band 7 groups, 30–42 percent indicated they wanted to write an equation. In grade bands 6, 7, and 11, the additional information recorded on paper would have improved the response according to the rubric. Responses for specific items in grade bands 6 and 11 were improved by 15 percent of the students, and responses for all items in grade band 7 were improved when information on the scratch paper was taken into account. Improvement for this group ranged between 10 and 20 percent of the responses. ("Confused me, I didn't know how to write an equation." "Tried the keypad, but it wouldn't work." "It was much easier with paper.") This was supported by interviewer observations. About 5–10 percent of students in each grade band found the online system difficult to use, but few specifics were recorded.

Research Question 11: Usability of equation editor tool—can students use the tool the way it is meant to be used?

Although students begin to use technology at a very early age, it is prudent to verify that young students are able to use the assessment interface to be used during testing. This question sought to evaluate the ability of grades 3–5 students to use the equation editor tool to be included in the Smarter Balanced delivery system. Three mathematics items were presented to the students (*N*=33). The first item only required the student to copy his or her response. The second item was a simple mathematics item, and the third item was a more challenging mathematics item. The first item would demonstrate whether the student could use the equation editor tool. The second and third items would provide evidence of whether the ability to use the tool interacted with item difficulty.

Elementary students had some difficulty using the equation editor. Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. The examiner's assessment concurred that about 35 percent of students had difficulty using the equation editor and that about 50 percent of the students would get a given item correct.

Research Question 12: Can students compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication?

This question is designed to assess whether students with a strong understanding of fractions and the multiplication and division of fractions complete the items without performing the indicated multiplication. The task asked students to compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication. Also of interest was whether students who complete an item as intended (without using multiplication) spent less time on an item than those who did not. To investigate this question a single form was administered for grades 3–5.

There seemed to be little relationship between whether a student has a strong understanding of the multiplication and division of fractions and whether he or she used multiplication to solve the items. However, students who did not need to perform the multiplication completed the items in less time than students who had to perform the multiplication. While most students said they understood the questions, 70 percent had to use multiplication to solve them. Only about 40 percent of the students had a firm understanding of the multiplication/division of fractions, according to the interviewers.

Research Question 13: Contextual glossaries are item-specific glossaries that provide a definition of a word that is targeted to, and appropriate for, the context in which the word is used in the item. Are these a fair and appropriate way to support students who need language support?

This question addressed the efficacy of the use of contextual glossaries with non-native speakers when solving mathematics problems. Two sets of items were created that were parallel in difficulty. The first set of items contained no contextual glossaries with only single words translated. The second set of items contained contextual glossaries. The interviewer was asked to determine whether the student was having trouble understanding a word and whether the contextual glossary aided in the interpretation of the word or sentence.

Only three ELL students participated: one from grade 3 and two from grade 6.

The contextual glossaries appeared to be somewhat effective, but the impact was not always reflected in the score the student received for an item. The contextual glossaries appeared to be incomplete in that they did not include words the student needed. This limited the use of the glossaries in these situations. Interviewer's comments suggested that performance was improved when the students used the contextual glossaries.

Research Question 14: Under what conditions does the use of text-to-speech (TTS) help students with lower reading ability focus on content in ELA and mathematics?

TTS can provide access to an assessment for students with low reading ability. In order for this technology to be effective the language produced from the voice-pack must be clear enough to be understood. This is particularly true for non-native speakers of English.

Only students familiar with TTS were included in the study. Overall, 77 students used TTS at least once. Among them, 58 students were limited English proficient (LEP), 13 students had reading difficulties (IEP), and six were Gen Ed students.

In ELA four forms were administered with both high- and low-quality voice-packs. In mathematics, two forms were administered in grade bands 3 and 11. Only a single form was administered in grade band 6. The mathematics forms were only administered with high-quality voice-packs.

TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used. LEP students and students with reading difficulties tended to benefit more from the use of TTS. Using TTS with high-quality voice-packs improved focus on content in ELA. The use of TTS with low-quality voice-packs tended to distract students in ELA, whereas high-quality voice-packs did not. In mathematics, access was improved only for grade band 3 students. All Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting. This was in part a function of trying to describe a table verbally.

Introduction

Smarter Balanced has conducted cognitive laboratories to better understand how students solve items in different formats. A cognitive laboratory uses a think-aloud methodology in which students speak their thoughts while solving a test item. The interviewer follows a standardized protocol to elicit responses and record what a student says. While this one-on-one process is time consuming, the type of information elicited is often difficult to obtain by other means. Due to the nature of the process the sample sizes are often small; however, they are sufficient to detect large effects. In addition, because each student's comments are recorded, smaller, non-primary effects may be brought to light. Most protocols were developed at multiple grade bands (e.g., 3, 6, and 11). A grade band is the level of content for which the protocol is targeted.

What follows are in-depth analyses for each research question outlined in the executive summary. Because of the differences in the samples, study design, and questions asked, each research question result is presented separately. A summary of the findings for each research question is provided at the end of each research question section. Research questions have been organized into sections of similar content to improve integration of the material. Finally, a conclusions section appears at the end of the document. The overall demographics for the cognitive labs sample can be found in Appendix B.

Processing Selected-Response (SR), Technology-Enhanced (TE), and Constructed-Response (CR) Items

Research Question 1: Do mathematics multi-part selected-response (MPSR) items provide information about the depth of understanding of the test taker similar to traditional constructed-response Items?

An MPSR item has students select several examples of a correct response rather than just one, as in the typical SR item. The intention of this research question was to see whether the MPSR items provided depth of understanding similar to that of CR items. If effective, an MPSR item would be a more efficient way to measure the content measured by CR items. Also of interest was whether similar results would be obtained at different educational levels. To investigate these questions, forms were constructed at four grade bands: grades 3–4 (referred to as grade band 3), grades 6–7 (referred to as grade band 6), grades 7–8 (referred to as grade band 7), and grade 9–10 (referred to as grade band 11). Within a form, parallel items were constructed in both formats and presented to the same students. In the protocols the MPSR and CR items were presented in random order. In the tables below the SR and CR data for each item are presented adjacent to each other to facilitate comparisons between the two item formats.

Interviewers were asked to assess the highest level of DOK the student demonstrated during the think-aloud session. Table 1 (ELA) and Table 2 (mathematics) show the rubrics the interviewers used during this process.

Two hypotheses related to research question 1 were examined. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis examined whether students who get full credit on MPSR items reveal understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target.

Table 1. Depth of Knowledge Chart (ELA)

DOK Level	Definition	Types of statements
1	Recall and Reproduction	1. Recalls facts, details, and events 2. Uses word relationships (synonym/ antonym) to determine meaning 3. Recognizes or retrieves information from tables and charts
2	Basic Skills and Concepts	1. Summarizes information 2. Identifies central ideas 3. Uses context to determine word meanings 4. Analyzes text structure and organization 5. Compares literary elements, facts, terms, or events
3	Strategic Thinking and Reasoning	1. Uses supporting evidence to explain, generalize, or connect ideas 2. Analyzes or interprets author's craft (literary devices, viewpoint, potential bias) to critique a text 3. Develops a logical argument and cites evidence

Table 2. Depth of Knowledge Chart (Mathematics)

DOK Level	Definition	Types of statements
1	Recall and Reproduction	I remembered it. We learned the answer in class. I did what it said. I recognized it.
2	Basic Skills and Concepts	1. Any statement indicating putting two or more pieces of knowledge together 2. An statement indicating that they executed a sequence of steps that was not given to them 3. Any inference relating two different things 4. Expression of a hypothesis or guess about a relationship
3	Strategic Thinking and Reasoning	1. Any statement indicating that they are applying abstract concepts to concrete phenomenon, e.g., “Both patterns reflect exponential growth” 2. Statements indicating that the students evaluated several different approaches to solving the problem, accompanied by the ability to explain why they selected the solution path they chose 3. Explanations of their choices or decisions using data and information from multiple sources to construct a coherent and logical argument

Results

Twenty students were administered the grade band 3 form, 37 students were administered the grade band 6 form, 31 students were administered the grade band 7 form, and 19 students were administered the grade band 11 form.

Table 3 presents the average demonstrated DOK level by students who received full credit on an item for each grade band/target. Table 4 shows the correspondence between the target labels and the full target description. Blank cells are the result of incomplete data, either in the score or in the demonstrated DOK. In most cases the DOK the student demonstrated either equals to exceeds the DOK demonstrated for the CR items. Interviewers commonly commented that the student did equally well on both item formats.

Table 3. Average DOK Demonstrated by Students Who Received Full Credit for Paired MPSR and CR Items Measuring the Same Target

Grade Band	Target	Item Format	Avg. DOK
3	Geometric Measurement: Perimeters (J)	MPSR	2.00
		CR	1.50
3	Reason with Shapes (K)	MPSR	1.80
		CR	1.67
6	One Variable Equations (F)	MPSR	1.57
		CR	1.60
6	Analyze Proportional Relationships (A)	MPSR	
		CR	1.25
6	Generate Equivalent Expressions (C)	MPSR	2.00
		CR	2.00
6	Apply Arithmetic to Algebra (E)	MPSR	1.60
		CR	2.00
7	Analyze Proportional Relationships (A)	MPSR	
		CR	1.83
7	Generate Equivalent Expressions (C)	MPSR	1.77
		CR	2.00
7	Solve Linear Equations (D)	MPSR	2.00
		CR	1.80
11	Equivalent Problem Solving (E)	MPSR	2.33
		CR	1.75
11	Graph Equations and Inequalities (J)	MPSR	
		CR	1.70
11	Use of Functions (K)	MPSR	2.10
		CR	2.00

Table 4. Correspondence Between Target Label and the Full Target Description

Target Label	Full Target Description
Geometric measurement: Perimeters	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures
Reason with Shapes	Reason with shapes and their attributes
Place Value: Whole Numbers	Generalize place value understanding for multi-digit whole numbers
Converting Units of Measure	Solve problems involving measurement and conversion of measurements from a larger unit to a smaller unit
Geometric measurement : Perimeters	Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures
One Variable Equations	Reason about and solve one-variable equations and inequalities
Apply Arithmetic to Algebra	Apply and extend previous understandings of arithmetic to algebraic expressions
Generate Equivalent Expressions	Use properties of operations to generate equivalent expressions
Analyze Proportional Relationships	Analyze proportional relationships and use them to solve real-world and mathematical problems
Solve Linear Equations	Analyze and solve linear equations and pairs of simultaneous linear equations
Equivalent Problem Solving	Write expressions in equivalent forms to solve problems
Graph Equations and Inequalities	Represent and solve equations and inequalities graphically
Use of Functions	Understand the concept of a function and use function notation

The second hypothesis examined whether students who get full credit on the MPSR items reveal greater understanding of the material than those who do not obtain full credit. Table 5 presents these findings. In all cases those who receive full credit for an item showed greater understanding than those who did not receive full credit. The percentage understanding is also quite similar for the MPSR and CR items.

Table 5. Percentage of Students Who Appear to Understand the Material, by Item Type, Grade Band, and Whether Full Credit Was Received

	Grade Band							
	3		6		7		11	
Item	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
MPSR1	20	50	17	89	38	78	64	-
CR1	12	100	25	100	55	-	40	100
MPSR2	0	57	29	100	42	83	45	100
CR2	7	75	23	90	36	75	67	67
MPSR3	0	-	10	67	48	100	33	75
CR3	0	-	8	100	50	75	58	67

Summary

This research question sought to address two hypotheses. The first hypothesis examined whether students who get full credit on MPSR items reveal, through their think-aloud sessions, greater understanding than those students who do not achieve full credit. The second hypothesis examined whether students who get full credit on MPSR items reveal depth of understanding similar to that of students who get full credit on similarly challenging CR items measuring the same target. In most cases the DOK the student demonstrated either equaled or exceeded the DOK demonstrated for the CR items.

Students who got full credit on the MPSR items also revealed greater understanding of the material than those who did not obtain full credit. The percentage of students understanding the material is also quite similar for the MPSR and CR items. A typical interviewer comment was, “based on the accuracy of the student’s responses to both types of items, it appears that item type is not a factor in determining how well the students respond[s].”

Research Question 2: Under what conditions do specific types of TE items (and SR items) approach the depth of knowledge (DOK) of a written constructed response in ELA and mathematics?

The question is designed to assess whether different types of TE items approach the DOK of CR items for specific content claim/targets and DOK levels. SR items were also included, where available, as a comparison item format. Comparisons were examined for specific TE item types at specific DOK levels for specific content claims/targets (see Appendix A for a full description of the claims and targets). Where possible, parallel items were created in each item format at the same DOK level and content claim/target; however, some combinations were not available. In ELA, items in the different formats were administered for most item type/content target/DOK combinations. In mathematics, however, some item formats were not administered for all claim/target/DOK conditions and some data were incomplete. This limited the comparisons that could be made. Four items in one of the three formats (SR, TE, and CR) appeared in each form. Multiple forms were administered, each to a different sample of students. It was hypothesized that students responding to items of a specific TE type would reveal that they are using thought processes consistent with a specific DOK level for items measuring a specific target.

Forms were constructed in ELA at five grade bands: grade 3 (referred to as grade band 3), grades 4–5 (referred to as grade band 4), grades 6–7 (referred to as grade band 6), grades 7–8 (referred to as grade band 7), and grade 11 (referred to as grade band 11). In mathematics, forms were constructed at four grade bands: grades 3–4 (referred to as grade band 3), grades 4–5 (referred to as grade band 4), grades 6–7 (referred to as grade band 6), and grade 11 (referred to as grade band 11). Note that the grade band relates to the level of the material in the assessment and not necessarily the grade of the students to which the assessment is administered. A single form was administered in each grade band. This was a between-subjects design in which different item types were administered to different students. For this question, the comments presented are made by the interviewer, as opposed to the student, due to the nature of the information being captured (e.g., DOK level demonstrated).

Results

Table 6 shows the sample sizes within a grade band by item format across item types and content area. The ELA forms tended to have been administered to larger samples than were the mathematics forms.

Table 6. Sample Sizes Within Grade Band, by Content Area and Item Type

		Grade Band				
Content	Item Format	3	4	6	7	11
ELA	SR	18	16	13	8	6
ELA	TE	12	14	10	8	14
ELA	CR	14	13	13	15	10
Mathematics	SR	7	6	23	-	10
Mathematics	TE	7	4	13	-	3
Mathematics	CR	4	11	8	-	3

Tables 7a (ELA) and 7b (Mathematics) list the percentage of students whose thought processes were consistent with the DOK level of the items for the respective content areas. For each TE item type, the percentage of students who demonstrated thought processes consistent with the grade band/content claim and target/DOK was recorded. SR and CR items were matched to the same grade band/content claim and target/DOK levels. The primary comparison of interest is between the TE and CR formats.

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. ("Well thought out. Uses evidence she feels supports the main idea of the item.") Two exceptions were two targets in the "select text" item type: "justifying interpretations" (grade band 6) and "analyzing the figurative" (grade band 11). A pattern similar to that of the TE item types was observed for the matched SR items versus the CR items.

Table 7a. Percentage of Students Demonstrating That They Are Using Thought Processes at the Specified DOK level, by Item Type, Claim, Target, and DOK Level (ELA)

					% of Students With Consistent Thought Process		
TE Item Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Drag and Drop (Tiling)	6	Justifying interpretations (11)	1	3	63	78	40
Drag and Drop (Tiling)	7	Writing or revising strategies (6)	2	2	100	80	61
Reorder Text	3	Writing or revise strategies (3)	2	2	81	69	54
Reorder Text	6	Organizing ideas (3)	2	2	60		
Select Text	6	Justify interpretations (11)	1	2	33	50	60
Select Text	7	Identifying text to support inferences (1)	1	2	94	79	64
Select Text	7	Writing or revising strategies (6)	2	2	100	80	61
Select Text	11	Citing to support inferences (1)	1	2	72	82	69
Select Text	11	Analyzing the figurative (7)	1	2	33	50	55

For mathematics, the pattern is less clear. The TE item types that yielded a higher percentage of students demonstrating thought processes consistent with the DOK level included:

- “placing points” for fractions, claim 1, DOK 2 in grade band 3 (“This student had a thorough understanding of these fractions and how they related to the number line. He thoroughly and accurately placed each fraction and explained how/why using various steps.”)
- “single lines” for equations and inequalities, claim 1, DOK 2 in grade band 11
- “tiling” for fractions, claim 1, DOK 2 in grade band 3 (“This student clearly understood and explained how to solve this item using multiple methods. He used multiple steps to solve each item.”)
- “tiling” for equations and inequalities, claim 1, DOK 2 in grade band 11 (“Student indicated use of multiple steps and solved correctly.”)
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

Places where equal percentages were observed for the TE and CR formats included:

- “select and order” for fractions, claim 1, DOK 2 in grade band 3
- “select and order” for fractions, claim 1, DOK 2 in grade band 6
- “selecting points” for fractions, claim 1, DOK 2 in grade band 3
- “single lines” for everyday math problems, claim 2, DOK 2 in grade band 11

Item types for CR items yielding a higher percentage of students who demonstrate consistent thought processes included:

- “select and order” for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- “tiling” for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- “tiling” for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6 (The student was able to explain his answer in multiple steps and with a clear understanding of the distributive property.”)
- “tiling” for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4 (“This student understood right angles. She also understood that she had to name a similarity and a difference.”)

Table 7b. Percentage of Students Demonstrating That They Are Using Thought Processes at the Specified DOK Level, by Item Type, Claim, Target, and DOK Level (Mathematics)

					% of Students With Consistent Thought Process		
TE Item Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Placing Points	3	Fractions (F)	1	2	50	53	0
Select and Order	3	Fractions (F)	1	2	0	53	0
Select and Order	6	Apply arithmetic to algebraic expressions (E)	1	2	40	67	79
Select and Order	6	Everyday math problems (A)	2	3	50		
Selecting Points	3	Fractions (F)	1	2	0	53	0
Single Lines	11	Equations and inequalities (I)	1	2	50	82	42
Single Lines	11	Everyday math problems (A)	2	2	100		100
Tiling	3	Fractions as numbers (F)	1	2	71	53	0
Tiling	4	Everyday math problems (A)	4	3	0	33	52
Tiling	6	Apply arithmetic to algebraic expressions (E)	1	2	60	67	79
Tiling	11	Equations and inequalities (I)	1	2	50	82	42
Tiling	11	Everyday math problems (A)	2	3	0	39	50
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (A)	4	3	67	33	52
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (L)	1	2	33	83	72

Also of interest was how students performed on these item types. Since not all items are 1-point items, the percentage obtaining the maximum score was used. Table 8a presents this information for ELA; Table 8b presents this information for mathematics. In ELA, the pattern is very similar to the consistency of thought process table. The same TE item types had higher percentages than the CR items, with the exception of the “select text” items for the “writing or revising strategies” target (grade band 7), and the “citing to support inferences” target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following “select text” items:

- “select text” for justifying interpretations, claim 1, DOK 2 in grade band 6
- “select text” for citing to support inferences, claim 1, DOK 2 in grade band 11
- “select text” for analyzing the figurative, claim 1, DOK 2 in grade band 11

Table 8a. Percentage of Students Receiving Full Credit for an Item (ELA)

					% of Students Who Answered Correctly		
TE Type	**Grade Band**	**Target**	**Claim**	**DOK**	**TE**	**SR**	**CR**
Drag and Drop (Tiling)	6	Justifying interpretations (11)	1	3	80	67	18
Drag and Drop (Tiling)	7	Writing or revising strategies (6)	2	2	67	22	47
Reorder Text	3	Writing or revise strategies (3)	2	2	64	0	44
Reorder Text	6	Organizing ideas (3)	2	2	12		
Select Text	6	Justifying interpretations (11)	1	2	10	70	25
Select Text	7	Identifying text to support inferences (1)	1	2	77	19	41
Select Text	7	Writing or revising strategies (6)	2	2	0	22	47
Select Text	11	Citing to support inferences (1)	1	2	22	67	40
Select Text	11	Analyzing the figurative (7)	1	2	8	46	31

In mathematics, the TE items in which a higher percentage of students received the maximum possible score included only:

- “single lines” for equations and inequalities, claim 1, DOK 2 in grade band 11
- “tiling” for equations and inequalities, claim 1, DOK 2 in grade band 11
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

For the SR items in mathematics, the percentage receiving the maximum score was higher than both the CR and TE formats for the following items:

- “placing points” for fractions, claim 1, DOK 2 in grade band 3
- “select and order” for fractions, claim 1, DOK 2 in grade band 3
- “selecting points” for fractions, claim 1, DOK 2 in grade band 3
- “tiling” for fractions, claim 1, DOK 2 in grade band 3
- “tiling” for everyday math problems, claim 4, DOK 3 in grade band 4

- “vertex-base quadrilaterals” for fraction equivalence and ordering, claim 1, DOK 2 in grade band 3

In other cases the percentage receiving the maximum score was lower than for the comparable CR items. It should be noted that the percentage receiving the maximum scores was generally low in mathematics.

Table 8b. Percentage of Students Receiving Full Credit for an Item (Mathematics)

					% of Students Who Answered Correctly		
TE Type	Grade Band	Target	Claim	DOK	TE	SR	CR
Placing Points	3	Fractions (F)	1	2	17	35	25
Select and Order	3	Fractions (F)	1	2	14	35	25
Select and Order	6	Everyday math problems (A)	2	3	0		
Select and Order	6	Apply arithmetic to algebraic expressions (E)	1	2	0	21	44
Selecting Points	3	Fractions (F)	1	2	14	35	25
Single Lines	11	Everyday math problems (A)	2	2	0		80
Single Lines	11	Equations and inequalities (I)	1	2	33		27
Tiling	3	Fractions (F)	1	2	33	35	25
Tiling	4	Everyday math problems (A)	4	3	0	25	5
Tiling	6	Apply arithmetic to algebraic expressions (E)	1	2	31	21	44
Tiling	11	Everyday math problems (A)	2	3	33	16	40
Tiling	11	Equations and inequalities (I)	1	2	33	0	27
Vertex-Based Quadrilaterals	3	Fraction equivalence and ordering (F)	1	2	21	35	25
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (A)	4	3	0	25	5
Vertex-Based Quadrilaterals	4	Lines, angles, and shapes (L)	1	2	67	50	0

Summary

For ELA, students demonstrated a higher DOK level for most of the TE item types than for the matched CR items. Two exceptions were two targets in the “select text” item type, “justifying interpretations” (grade band 6) and “analyzing the figurative” (grade band 11). A similar pattern was observed for the matched SR items versus the CR items. In ELA, the pattern is very similar to the consistency of thought process table. The same TE item types had higher percentages than did the

CR items, with the exception of the “select text” items for the “writing or revising strategies” target (grade band 7) and the “citing to support inferences” target (grade band 11).

For the SR items in ELA, the percentage receiving the maximum score was higher than both the CR and TE formats for the following “select text” items:

- “select text” for justifying interpretations, claim 1, DOK 2 in grade band 6
- “select text” for citing to support inferences, claim 1, DOK 2 in grade band 11
- “select text” for analyzing the figurative, claim 1, DOK 2 in grade band 11

For mathematics, the pattern is less clear. The TE item types that showed a higher percentage of students demonstrating consistent thought process with the DOK level included:

- “placing points” for fractions, claim 1, DOK 2 in grade band 3
- “single lines” for equations and inequalities, claim 1, DOK 2 in grade band 11
- “tiling” for fractions, claim 1, DOK 2 in grade band 3
- “tiling” for equations and inequalities, claim 1, DOK 2 in grade band 11 (“Student indicated use of multiple steps and solved correctly.”)
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 4, DOK 3 in grade band 4

Places where equal percentages were observed for the TE and CR formats included:

- “select and order” for fractions, claim 1, DOK 2 in grade band 3
- “select and order” for fractions, claim 1, DOK 2 in grade band 6
- “selecting points” for fractions, claim 1, DOK 2 in grade band 3
- “single lines” for everyday math problems, claim 2, DOK 2 in grade band 11

Item types where the CR items had a higher percentage of consistent thought processes included:

- “select and order” for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- “tiling” for everyday mathematic problems, claim 4, DOK 3 in grade band 4
- “tiling” for apply arithmetic to algebraic expressions, claim 1, DOK 2 in grade band 6
- “tiling” for everyday mathematic problems, claim 2, DOK 3 in grade band 11
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

The TE item types where a higher percentage of students received full credit included only:

- “tiling” for equations and inequalities, claim 1, DOK 2 in grade band 11
- “vertex-base quadrilaterals” for lines, angles, and shapes, claim 1, DOK 2 in grade band 4

For the SR items in mathematics, the percentage receiving the maximum score was higher than both the CR and TE formats for the following items:

- “placing points” for fractions, claim 1, DOK 2 in grade band 3
- “select and order” for fractions, claim 1, DOK 2 in grade band 3
- “selecting points” for fractions, claim 1, DOK 2 in grade band 3

- “tiling” for fractions, claim 1, DOK 2 in grade band 3
- “tiling” for everyday math problems, claim 4, DOK 3 in grade band 4
- “vertex-base quadrilaterals” for fraction equivalence and ordering, claim 1, DOK 2 in grade band 3

In other cases the percentage receiving full credit was lower than for the comparable CR items. It should be noted that the percentage receiving full credit was generally low in mathematics.

Research Question 3: For multi-part selected response (MPSR) items where students may select more than one answer choice, which wording best indicates to the student that he or she is allowed to select more than one option? For multipart (e.g., YES/NO) dichotomous choice items, do students know that they need to answer each part?

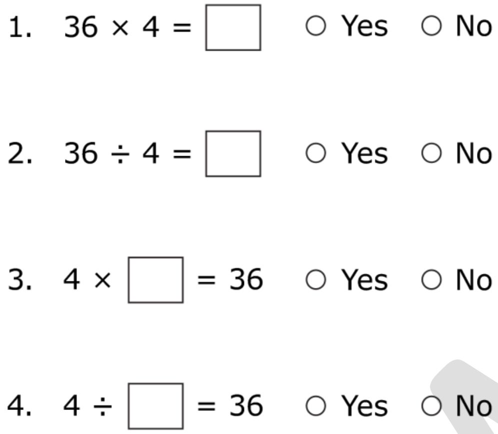
Smarter Balanced sought to investigate whether students might become confused by MPSR items in mathematics and perhaps not complete the entire item. In order to investigate this, items were constructed with different amounts of labeling. Labeling is the identification of the parts of the problem with indicators such as “a,” “b,” “c” or “1,” “2,” “3.” A “labeled” and a non-labeled” condition were investigated. An example of items in the labeled and unlabeled format is presented below (Exhibit 1).

This question is designed to assess whether labeling or not labeling an MPSR mathematics item produces a difference in performance. Results are reported in five grade bands. The five grade bands are designated as grade band 3 (which includes form difficulty levels 3 and 4), grade band 4 (which includes form difficulty levels 4 and 5), grade band 6 (which includes form difficulty levels 6 and 7), grade band 7 (which includes form difficulty levels 7 and 8), and grade band 11 (which includes form difficulty level 11). Each form contains one MPSR item followed by one CR item. The labeled and non-labeled items appeared in different forms of the test and thus were taken by different students.

Exhibit 1. Example of a Labeled Item

Marcus has 36 marbles. He is putting an equal number of marbles into 4 bags.

Indicate whether each equation could be used to find the number of marbles Marcus puts in each bag.



1. 36 × 4 = □ ○ Yes ○ No

2. 36 ÷ 4 = □ ○ Yes ○ No

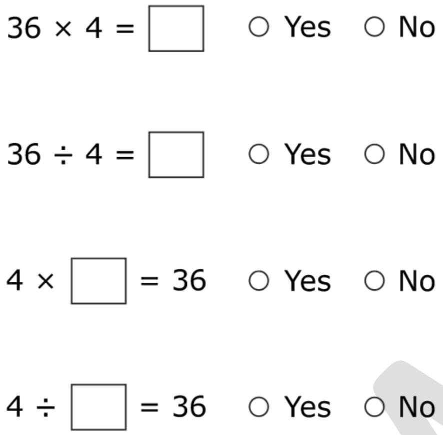
3. 4 × □ = 36 ○ Yes ○ No

4. 4 ÷ □ = 36 ○ Yes ○ No

Example of an Unlabeled Item

Marcus has 36 marbles. He is putting an equal number of marbles into 4 bags.

Indicate whether each equation could be used to find the number of marbles Marcus puts in each bag.



$36 \times 4 = \square$ ○ Yes ○ No

$36 \div 4 = \square$ ○ Yes ○ No

$4 \times \square = 36$ ○ Yes ○ No

$4 \div \square = 36$ ○ Yes ○ No

Results

Ninety-six students were administered the grade band 3 forms, 66 students were administered the grade band 4 forms, 133 students were administered the grade band 6 forms, 33 students were administered the grade band 7 forms, and 85 students were administered the grade band 11 forms.

Table 9 shows the percentage of students receiving full credit on the items by grade band and labeling condition. For grade bands 3, 4, 6, and 11 little difference between the labeled and non-labeled conditions is observed. However, in grade band 7 a higher percentage of students received full credit in the non-labeled format.

Table 9. Percentage of Students Receiving Full Credit, by Grade Band and Labeling Condition.

	Grade Band				
Condition	3	4	6	7	11
Non-Labeled	32	32	20	62	16
Labeled	29	31	18	34	9

Table 10 shows whether the students understood the instructions under the different item labeling conditions. Up through grade band 6 the type of instructions received seemed to have little impact on the understanding of the instructions. However, in grade bands 7 and 11 a higher percentage of students tended not to understand the instructions when the items were labeled. The interviewers commented that “Student did not have a complete understanding of instructions” and “He said he understood, however, he only selected one bubble.”

Table 10. Percentage Understanding the Instructions, by Grade Band and Labeling Condition

	Grade Band				
Condition	3	4	6	7	11
Non-Labeled	63	83	93	97	84
Labeled	78	93	93	69	61

Table 11 shows the percentage of students who made comments about not understanding the instructions. Grade bands 3 and 11 had more comments about not understanding the instructions than the other grade bands, but the pattern was similar for labeled and non-labeled items. However, in grade band 7, non-labeled items generally received no comment, with labeled items receiving more comments. This is consistent with a lower percentage of grade band 7 students understanding the instructions in the “labeled” condition.

Table 11. Did the Student Make Comments About not Understanding the Instructions (Percentage Making Comments)?

	Grade Band				
Condition	3	4	6	7	11
Non-Labeled	34	17	15	3	33
Labeled	32	26	8	38	41

Summary

Even though the labeling of MPSR items was intended to clarify the mathematic tasks for the students, in many cases it actually seemed to confuse the students. Little difference was observed between the labeled and non-labeled items in the lower grade bands (grade bands 3–6). However, students in grade band 7 tended to score higher with non-labeled items. Also, grade band 7 and 11 students tended to be confused by the labeling. In addition, the labeled items tended to receive more comments related to not understanding the instructions. The interviewer confirmed this, suggesting that the grade band 7 and 11 students better understood the instructions in the non-labeled condition than in the labeled condition.

Research Question 4: Does the ability to move one or more sentences to different positions provide evidence of students' ability to revise text appropriately in the consideration of chronology, coherence, transitions, or the author's craft?

Smarter Balanced is considering using items that have students reorder sentences to measure an editing/revising standard. Claim 2 of the standards states that students should be able to revise one or more paragraphs demonstrating specific narrative strategies (use of dialogue, sensory or concrete details, description), chronology, appropriate transitional strategies for coherence, or authors' craft appropriate to purpose (closure, detailing characters, plot, setting, or an event).

This question was designed to assess whether students' movement of one or more sentences to different positions provides evidence of students' ability to demonstrate consideration of chronology, coherence, transitions, or author's craft. Six ELA items were included in a test form. The forms were administered to five students: two in grade 5, two in grade 6, and one in grade 10. Because there is little difference in the pattern of responses and because the sample sizes are small, the results will be reported for the sample as a whole.

Results

It was hypothesized that students who do well on these items would recognize the need to revise for chronology, coherence, transitions, or author's craft. Table 12 shows the percentage of students who recognize the need to revise for chronology, coherence, transitions, or author's craft for students who performed well on the items and those who performed poorly. The results show that students who performed well are more likely to consider chronology, coherence, transitions, or author's craft in their revisions than students who do not. Among the four writing skills examined, author's craft was considered less often than the other three writing skills.

Table 12. Percentage of Students Considering Targeted Writing Skills When Revising, by Those Students Who Performed Well and Those Who Performed Poorly

Characteristic	Students Who Perform Well	Students Who Perform Poorly
Chronology	100%	33%
Coherence	100%	33%
Transitions	100%	33%
Author's Craft	50%	0

Also of interest was whether students referenced organization, coherence, transitions, or author's craft when moving sentences. Table 13 shows the percentage of students who considered each of the targeted writing skills relative to the number of appropriate and inappropriate sentence moves. The results suggest that students who make more appropriate sentence moves (and fewer inappropriate sentence moves) are more likely to consider the writing skills of chronology, coherence, and transitions; however, the pattern is less clear for consideration of author's craft.

Table 13. Percentage of Students Who Considered Chronology, Coherence, Transitions, and Author's Craft at Each Number of Appropriate and Inappropriate Sentence Moves

% Students Who Recognized Need For	*N* Appropriate Sentences Moved								*N* Inappropriate Sentences Moved							
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
Chronology	38	50	50	67	75		100	100	100	100	40	50	33	0		0
Coherence	38	43	67	67	75		100	100	100	100	40	50	33	0		0
Transitions	38	50	50	67	75		100	100	100	100	40	50	33	0		0
Anthor's craft	13	13	0	33	25		67	0	29	17	20	50	33	0		0

Table 14 shows the percentage of students who considered chronology, coherence, transitions, and author's craft when answering the items as observed by the interviewers. Students did express consideration of chronology ("I moved the first sentence because it goes at the top," "This seems to be in order," "This should be the second to last sentence"); coherence ("This seems like something you'd say," "I don't need to take out more phrases, it sounds OK," "I removed the two sentences because they did not make sense and were irrelevant"); and transitions ("This would sound better here") when answering the questions; however, fewer took author's craft ("I think there is a flow to the story," "Some sentences are awkward and need to be moved") into account when answering these questions.

Table 14. Percentage of Students Who Considered Chronology, Coherence, Transitions, and Author's Craft When Answering, Across Items

Writing Skills	Chronology	Coherence	Transitions	Author's Craft
Percentage	68	68	57	18

Summary

Students who performed well on the items were more likely to consider the targeted writing skills (chronology, coherence, transitions, and author's craft) when answering the questions. Also, students who made appropriate sentence moves were more likely to consider the targeted writing skills than those who made inappropriate sentence moves. A high percentage of students considered chronology, coherence, and transitions; however, they were less likely to consider author's craft.

Research Question 5: Do Students Who Construct Text Reveal More Understanding of Targeted Writing Skills Than Students Who Manipulate Writing Through the Manipulation of Text (MT) Tasks?

Many believe that the best way to measure writing is to have students write. However, in a testing environment, it is often difficult to adequately sample the writing content domain with an assessment composed exclusively of CR items. There is an ongoing effort to find items that are efficient, but that can adequately measure the components of the writing domain, thus allowing for a broader selection and greater number of items to be delivered. Examples of the types of questions used can be seen in Exhibit 2. The question examines whether comparable understanding of the targeted writing skills (see Table 15) can be achieved using a set of MT tasks in comparison to comparable CR tasks.

Four pairs of ELA items were developed. Each pair contained one MT item and one CR version of the same item. Two forms were created, with each form containing a single version of an item. Each form contained two MT items and two CR items. The MT items were almost exclusively “select and order” items, though two of the items, one in grade band 3 and one in grade band 11, were “reorder text” items.

Forms were constructed in ELA at three grade bands: grades 3–5 (referred to as grade band 3), grades 6 and 7 (referred to as grade band 6), and grades 10 and 11 (referred to as grade band 11). In grade band 3 two forms were administered; in grade bands 6 and 11, only a single form was administered. All forms assessed claim 1, target 1.

The sample consisted of seven students in grade band 3, two students in grade band 6, and one student in grade band 11.

Exhibit 2. Sample Items Used in this Research Question

<u>Stem</u>

A student wrote the first draft of a story about a girl who eats nine berries for an afternoon snack every day. Read the story. Then complete the task that follows.

Every day after school, Kim eats nine red, juicy raspberries. One day, Kim sits down at the big kitchen table and has a surprise. She notices that one of her berries is missing! “[ ],” she says.

“I counted nine just a minute ago,” Dad says.

“[ ],” Kim says. “[ ].”

Kim begins her search in the garage. “[ ]?” Kim asks.

Dialogue

Oh no! There are only eight raspberries in my bowl

I wonder what happened to the ninth berry

Grandma, why are your mouth and lips red

It looks like I have a mystery to solve

Revise the story to include dialogue that introduces the plot. Place each piece of dialogue in the correct place in the story.

The dialogue will go in the brackets.

CR Prompt

A student wrote the first draft of a story about a girl who eats nine berries for an afternoon snack every day. Read the story. Then complete the task that follows.

> Every day after school, Kim eats nine red, juicy raspberries. One day, Kim sits down at the big kitchen table and has a surprise. She notices that one of her berries is missing!
>
> Her dad had counted nine just a few minutes ago.
>
> Kim knew she had a mystery to solve.
>
> Kim began her search in the garage. She found her grandmother in the garage with bright red lips.

Revise the story to include dialogue. Use dialogue to introduce the plot.
Type your response in the space provided.

Table 15. Targeted Writing Skills with Examples of Representative Statements

Target	Types of Statements
Chronology	- I knew it was telling a story, so looked for the beginning then moved the rest around to make sense. - I knew what the end was, so worked backwards from there. - I knew the youngest son went last, so put him at the end, then put the two older ones before him. Then picked the beginning and put it first. - Some spots didn't sound quite right, so added the sentences in. - Read the sentences, then looked for related sentences in the passage that they'd go with. - I used transitions to cue position of sentences. - I need to revise the order of the sentences so that they more clearly support the main idea of the article. I do not need to move the first or last sentence.
Coherence	- Sentence is like a preview of the rest of the essay, so it should go first. - This sentence sounds professional and it also connects to the facts that follow. This is the best thesis statement. - This sentence wraps up the author's argument/point of view and finishes the essay by restating the main point. - The conclusion often just rephrases the thesis, which this sentence does, but it also talks about other things from the passage, so it should be the conclusion. - I have to choose the two sentences that shouldn't be part of the paragraph. - I have to take the sentence at the top and drag it to best spot in the paragraph below.
Transitions	- The word "next" tells him it comes after something else. - The word "first" is a clue that it goes at the beginning. - "Finally" usually tells you you're at the end. - A transition like "therefore" at the start of a sentence connects it to the sentence before. They have the same topic but this one comes second. - I have to use transitions words to make the paragraph clearer. - I looked at the transition words to see what should come before them, then put in a sentence if needed.

Author's Craft	- I found the parts that didn't give me a really clear picture in her mind and changed them. - I looked for the parts that weren't as descriptive as the rest and made them more descriptive. - I looked for the parts that sounded a little boring and made them more exciting. - I read the topic sentences and looked for the sentence that didn't go with it. - If a sentence makes the argument weaker, then it should be taken out, so these two need to be removed.

Results

It was hypothesized that student think-alouds on MT items would reference the appropriate writing skills reflected in the assessment target at a level comparable with CR items. Table 16 shows the percentage of students who referenced the targeted writing skills, by item format and grade band. In grade band 3, chronology was more likely to be considered during revision when the item format was MT ("First, next, last order of events") than when the item format was CR ("Historically probably comes first, having trouble ending story"). Similar patterns, but less pronounced, were seen with coherence, transitions ("This is a cause…as a result (an effect) should be here"), and author's craft. Grade band 3 students only considered author's craft during revision for about one-third of the items regardless of item format. Grade band 6 students always considered chronology and coherence during revision, but transitions and author's craft were only considered about half the time. In grade band 11 chronology, coherence, and transitions were always considered in both formats. Author's craft was only considered about half the time in the CR format and not mentioned at all in the MT format. One interviewer commented, "Student made no comment about author's craft."

Table 16. Percentage of Items in Which Students Considered Target Characteristics When Responding to the Item, by Item Format

		Grade Band		
Target Characteristics	Item Format	3	6	11
Chronology	CR	31	100	100
	MT	94	100	100
Coherence	CR	63	100	100
	MT	75	100	100
Transitions	CR	44	50	100
	MT	69	50	100
Author's Craft	CR	31	50	50
	MT	43	100	0

Table 17 shows the counts for item scores received for the two item formats, by grade band. Comparable scores were achieved for the two item formats.

Table 17. What Score (Across Items) Would the Student Receive on this Type of Item?

		Grade Band								
		3				6		11		
	Item Format	0	1	2	0	1	2	0	1	2
	CR	8	5	3	0	0	2	0	0	2
	MT	7	6	2	1	0	1	0	2	0

Table 18 provides information about whether the students who construct text through writing reveal comparable or greater understanding of targeted writing skills than students who manipulate text. The grade band 3 and grade band 6 students were either more effective in applying the targeted writing skills when the items were in a MT format or no differences were observed in effectiveness between item formats. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft.

Table 18. Effectiveness of Applying Targeted Writing Skills by Item Format (Percentage of Students as Assessed by Interviewer)

	Grade Band								
	3			6			11		
Target Characteristics	MT More Effective	No Difference	CR More Effective	MT More Effective	No Difference	CR More Effective	MT More Effective	No Difference	CR More Effective
Chronology	38	63	0	0	100	0	0	100	0
Coherence	38	63	0	0	100	0	100	0	0
Transitions	25	75	0	50	50	0	0	0	100
Author's Craft	38	63	0	50	50	0	0	0	100

Summary

The results showed that the targeted writing skills are considered by students who manipulate text at a level comparable to (or greater than) that encountered when they are constructing text. The grade band 3 and 6 students showed comparable (or greater) levels of understanding when the items were in an MT format. For the grade band 11 students the results were mixed, but students tended to be more effective in applying the targeted writing skills in the CR format, particularly for transitions and author's craft. Score distributions were comparable for MT and CR item formats.

Research Question 6: Do different types of directions (minimal, concise or extensive) have an effect on the performance of different item types in ELA and Mathematics?

The optimal amount of direction that should be given to students for some item types is unclear. With minimal directions students may not know how to approach the item; with extensive directions students may be distracted or slowed to a point where the item becomes inefficient. This may be particularly true with elementary school students, who may take longer to process text. This question examined these issues for ELA and mathematics items. Three types of directions (minimal, concise, and extensive) were examined for different item types.

Forms were constructed in ELA at five grade bands: grade 3 (referred to as grade band 3), grades 4 and 5 (referred to as grade band 4), grades 6 and 7 (referred to as grade band 6), grades 7 and 8 (referred to as grade band 7), and grade 11 (referred to as grade band 11) with a single form administered in each grade band. In mathematics, forms were constructed at four grade bands: grades 3 and 4 (referred to as grade band 3), grades 4 and 5 (referred to as grade band 4), grades 6 and 7 (referred to as grade band 6), and grade 11 (referred to as grade band 11).

Parallel items were created with minimal, concise, or extensive directions in ELA and for most item types in mathematics. However, not all direction types appeared with all item types in all grades in mathematics. Four items in one of the three formats (SR, TE, and CR) appeared in each mathematics form. Two items in one of the three formats appeared in each ELA form. Multiple forms were administered, each one to a different sample of students. An example of the different direction types for an ELA item and a mathematics item is presented in Exhibit 3.

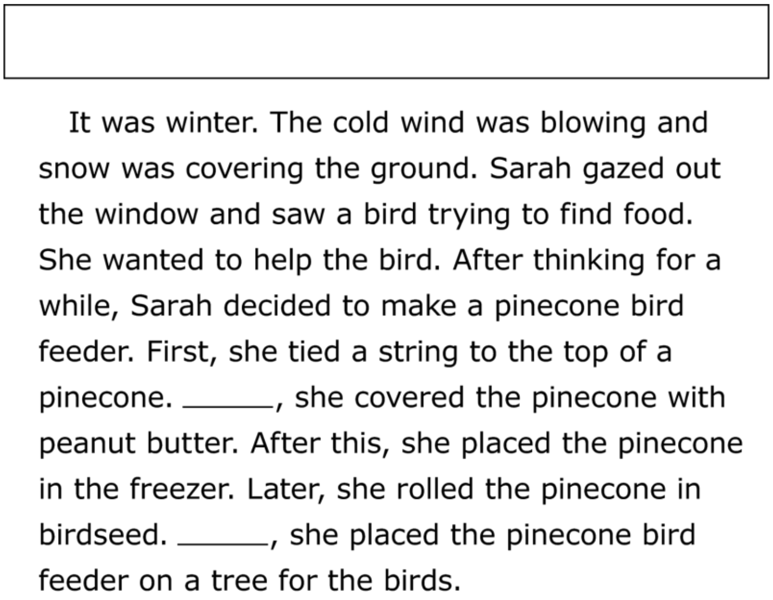
Exhibit 3. Example of the Types of Instructions Under the Minimal, Concise, and Extensive Instruction Condition for the Item That Follows

ELA Example

Minimal Directions
Drag the **best** transition word to each blank in the paragraph.

Concise Directions
Complete the paragraph by selecting the **best** transition word that fits in each blank. Drag each transition word you selected to the correct blank in the paragraph.

Extensive Directions
There are six transition words in the text box. Complete the paragraph correctly by choosing a transition word that **best** fits each blank. Drag the transition word you selected from the text box to the correct blank in the paragraph.



It was winter. The cold wind was blowing and snow was covering the ground. Sarah gazed out the window and saw a bird trying to find food. She wanted to help the bird. After thinking for a while, Sarah decided to make a pinecone bird feeder. First, she tied a string to the top of a pinecone. ______, she covered the pinecone with peanut butter. After this, she placed the pinecone in the freezer. Later, she rolled the pinecone in birdseed. ______, she placed the pinecone bird feeder on a tree for the birds.

Mathematics Example

Minimal Directions

Drag numbers to make the equations true.

Concise Directions

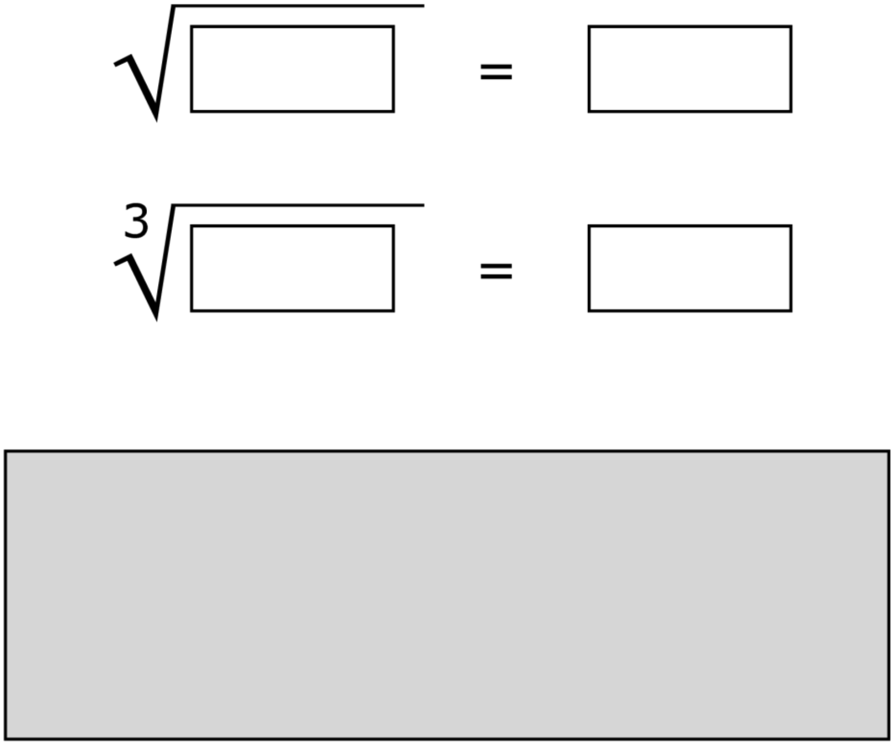
Move numbers to make the equations true.

Drag the numbers to the answer space.

Extensive Directions

Drag numbers to make the equations true.

Each number can be used only once. To use a number, drag it to the appropriate box in an equation.



Results

Table 19 provides a count of the students in a grade band, by content area and direction type.

Table 19. Sample Sizes by Content Area, Direction Type, and Grade Band

Content	Direction Type	Grade 3	Grade 4	Grade 6	Grade 7	Grade 11
ELA	Minimal	14	12	14	14	10
ELA	Concise	12	15	12	12	14
ELA	Extensive	18	17	15	7	6
Mathematics	Minimal	4	11	8	-	18
Mathematics	Concise	20	4	27	-	27
Mathematics	Extensive	19	4	27	-	16

Table 20a shows the percentage of students receiving full credit for the ELA items by direction type, item type, and grade band. In grade band 3, “select text” items were more challenging than “reorder text” items. This was especially true when the directions were “concise.” With the “reorder text” items the grade band 3 students did less well with minimal directions. The grade band 11 students also had some difficulty with the “reorder text” items when the directions were “extensive.” For the other grade bands, neither the level of instruction nor the item type showed a differential effect.

Table 20a. Percentage of Students Who Received Full Credit on ELA Items by Direction Type and Grade Band

ELA		Grade Band				
Direction Type	Item Type	3	4	6	7	11
Minimal	Reorder Text	40				71
Concise	Reorder Text	100				59
Extensive	Reorder Text	67				33
Minimal	Select and Order		69			
Concise	Select and Order		75			
Extensive	Select and Order		53			
Minimal	Select Text	33		100	41	
Concise	Select Text	0		100	60	
Extensive	Select Text	38		100	50	

In mathematics, a low percentage of students received full credit for “placing points” under the minimal and concise directions in grade band 11 (Table 20b). However, under extensive directions all students received full credit. With “placing points and tiling” items a higher percentage of students received full credit as the amount of instructions were reduced (grade band 6). “Select and order” items were difficult (grade bands 6 and 11) regardless of the direction type; however, no direction type proved better than another. The “select defined partition” items and the “straight lines” items showed high percentages of students receiving the maximum score, but the direction type did not make a difference. “Vertex-based quadrilateral” items seemed to benefit from minimal directions in grade band 11. Finally, “tiling” items were generally difficult, but no benefit was shown for different types of directions. The incompleteness of the data limits other comparisons.

Table 20b. Percentage of Students Who Received Full Credit on Different Types of Mathematics Items, by Direction Type and Grade Band

Direction	Template	Grade Band			
		3	4	6	11
Minimal	Placing Points				21
Concise	Placing Points				21
Extensive	Placing Points				100
Minimal	Placing Points and Tiling			67	
Concise	Placing Points and Tiling			57	
Extensive	Placing Points and Tiling			38	
Minimal	Select and Order				44
Concise	Select and Order			32	43
Extensive	Select and Order			33	0
Minimal	Select Defined Partitions	100	70		
Concise	Select Defined Partitions	76	100		
Extensive	Select Defined Partitions	71	83		
Extensive	Single Ray			15	
Minimal	Straight Lines		100		100
Concise	Straight Lines	100	100		100
Extensive	Straight Lines	100			
Extensive	Straight Line and Tiling			29	
Concise	Tiling	19			
Extensive	Tiling	20	20		
Minimal	Vertex-Based Quadrilaterals				69
Concise	Vertex-Based Quadrilaterals			64	88
Extensive	Vertex-Based Quadrilaterals	30			

Understanding instructions

In ELA (Table 21a), for most item type/direction type/grade band combinations few students had difficulty understanding instructions. Cases in which difficulties were mentioned included about 50 percent of the students in grade band 4 with both minimal and extensive instructions for the “select and order” items. This was also true in grade band 3 for the “reorder text” items with extensive instructions and for the “select test” items with concise and extensive instructions. Finally, in grade band 11 the “reorder text” items with minimal and concise instructions elicited more comments.

In mathematics (Table 21b), the cases in which more comments were made about the instructions included “placing points” with minimal and concise instructions (grade band 11), “single ray” items with extensive instructions (grade band 6), “straight lines” items with extensive instructions, and “vertex-based quadrilateral” items with extensive instructions (grade band 3). The single ray item with extensive instructions in grade band 6 stood out as an item in which instructions were not well understood. (“Weren’t totally sure how instructions were to be completed.”) The percentage of students getting the maximum score on this item type was also low.

Table 21a . Percentage of Students Who Express the Difficulties in Understanding Each Type of Instruction for Each TE Type in Their Think-Alouds (ELA)

ELA		Grade Band									
		3		4		6		7		11	
Direction Type	Item Type	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
Minimal	Reorder Text	0	0							0	20
Concise	Reorder Text		25							0	20
Extensive	Reorder Text	33	12							0	0
Minimal	Select and Order			50	11						
Concise	Select and Order			0	6						
Extensive	Select and Order			44	0						
Minimal	Select Text	0	0				0	6	0		
Concise	Select Text	33					14	0	22		
Extensive	Select Text	25	40				19	10	10		

Table 21b. Percentage of Students Who Express the Difficulties in Understanding Each Type of Instructions for Each TE Type in Their Think-Alouds (Mathematics)

Math		Grade Band							
		3		4		6		11	
Direction Type	TE type	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit	Non-Full Credit	Full Credit
Minimal	Placing Points							55	33
Concise	Placing Points							67	0
Extensive	Placing Points								20
Minimal	Placing Points and Tiling					0	0		
Concise	Placing Points and Tiling					33	25		
Extensive	Placing Points and Tiling					7	33		
Minimal	Select and Order							14	9
Concise	Select and Order					13	8	4	10
Extensive	Select and Order					12	8	12	
Minimal	Select Defined Partitions		0	0	0				
Concise	Select Defined Partitions	14	9		0				
Extensive	Select Defined Partitions	25	13	0	20				
Extensive	Single Ray					82	33		
Minimal	Straight Lines				25				
Concise	Straight Lines				0			100	0
Extensive	Straight Lines		50						
Extensive	Straight Lines and Tiling					0	0		
Concise	Tiling	15	0						
Extensive	Tiling	6	0	0	0				
Minimal	Vertex-Based Quadrilaterals							25	0
Concise	Vertex-Based Quadrilaterals	30	0			67	12	33	14
Extensive	Vertex-Based Quadrilaterals	43	17						

Difficulty Using the Computer

The results for ELA related to difficulty using the computer were mixed (Table 22). In grade band 3 under minimal directions for both “select text” and “reorder text” items, the students seemed to have difficulty using the computer. The grade band 11 students seemed to have some difficulty with the “reorder text” items.

Table 22. Percentage of Students Who Said They Had Trouble Using the Computer (ELA)

		Grade				
Direction Type	Item Characteristic	3	4	6	7	11
Minimal	Select Text	43		4	0	
Concise	Select Text	25		0	4	
Extensive	Select Text	19		8	0	
Minimal	Select and Order		22			
Concise	Select and Order		25			
Extensive	Select and Order		16			
Minimal	Reorder Text	31				25
Concise	Reorder Text	11				48
Extensive	Reorder Text	24				30

Most students in mathematics had little trouble using the computer with mathematics items.

Summary

In most cases in ELA the level of instruction did not have an influence. For most grade bands and item types, neither the level of instruction nor the item type had a differential effect in ELA. Cases in which differences were observed included “select text” items when the directions were “concise” (grade band 3). With the reorder text items the grade band 3 students did less well with minimal directions. The grade band 11 students also have some difficulty with the “reorder text” items when the directions were “extensive.”

In mathematics, the level of instruction also did not make a difference for many of the item types and grade bands. “Select and order” items were difficult (grade bands 6 and 11) regardless of the direction type; however, no direction type proved better than another. High percentages of students received full credit on the “select defined partition” items and the “straight lines” items; however, the direction type did not make a difference. Finally, “tiling” items were generally difficult, but no benefit was shown for different types of directions. Places where differences were observed included “placing points” under the minimal and concise directions in grade band 11; however, under extensive directions all students received the maximum score. In working with “placing points and tiling” items, a higher percentage of students received full credit with fewer instructions (grade band 6). Finally, “vertex-based quadrilateral” items seemed to benefit from minimal directions in grade band 11.

The results for ELA related to trouble using the computer were mixed. In grade band 3 under minimal directions with both select text and reorder text items the students seemed to have difficulty using the computer. The grade band 11 students seemed to have some difficulty with the “reorder text” items. Mathematics students did not seem to have any problems using the computer.

ELA Questions, Passage Processing

Research Question 7: Smarter currently intends to administer the passage first, and then administer the items one item at a time. Does this affect student performance?

Smarter Balanced is interested in the possibility of administering items adaptively within a passage. This would require administering items sequentially so that the ability estimate could be updated after each item. Presenting items one at a time may take longer, and students may object to not knowing what is coming next. This question is designed to assess whether administering an item set takes longer when the items are presented sequentially and whether there is a difference in confusion or frustration level when students are presented a passage and all the items together or are presented a passage with the items then being presented one at a time. The item sets were not administered adaptively.

Two sets of items were created for a given test form. Both sets contained passages of equivalent length and difficulty as well as items of equivalent difficulty.[2] The first set in a form presented the passage with all the items together. The second set presented the passage with the items presented one at a time.

The forms were administered, within grade band, to different samples of students. Each sample contained both a general education group (Gen Ed) and a group that received ELL students. One sample was timed without thinking aloud during the administration. Each item set in these forms was separately timed. This sample provided timing information only. The second sample involved thinking aloud while responding to the questions and was not timed. Forms were constructed in ELA at three grade bands: grades 3–5 (referred to as grade band 3), grades 6–8 (referred to as grade band 6), and grades 10 and 11 (referred to as grade band 11).

The primary questions of interest were:

1. Does presenting the items individually after the passage appear to take longer (timed condition)?
2. Does presenting the items individually after the passage increase the student’s negative emotional states (e.g., frustration, confusion; think-aloud condition)?
3. Do students prefer one approach or another (think-aloud condition)?

[2] Comparable passage difficulty was achieved through the use of readability and lexile measures. Comparable item difficulty was achieved through depth of knowledge (DOK) measures.

Results

Table 23 shows the sample sizes taking each form of the tests, by grade band, for the ELL and Gen Ed samples. Sample sizes are smaller for the ELL sample in grade band 11.

Table 23. Student Counts by Grade Band, Testing Population, and Testing Condition

		Grade Band		
		3	6	11
Timed	Gen Ed	9	6	8
	ELL	8	4	1
Think-Aloud	Gen Ed	6	6	7
	ELL	8	7	2

Table 24 shows the time (in seconds) it took to complete the item sets when all items were presented together or items were presented one at a time, by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items were presented in one block or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show a similar pattern. Note that the grade band 11 ELL sample was a single student and is not presented to avoid misleading results.

Table 24. Average Time to Complete the Passage and Items, by Administration Format, Grade Band, and Sample

Grade Band	Sample	*N*	Passage + All Items	Passage + One Item at a Time	Difference (All One at a Time)
3	Gen Ed	9	250	239	11
3	ELL	8	263	239	24
6	Gen Ed	6	401	462	–61
6	ELL	5	336	465	–129
11	Gen Ed	8	270	285	–15

Tables 25 and 26 show whether the ELL or Gen Ed sample students expressed confusion (Table 25) or frustration (Table 26) with the passages or items. There appears to be slightly more confusion for both the Gen Ed and the ELL sample students in grade band 3 when all the items are presented together. However, similar frustration levels were observed under the two formats for the grade band 3 students. The grade band 6 ELL sample, showed similar patterns of frustration and confusion for the two presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time. The grade band 11 Gen Ed students showed similar levels of confusion and frustration under both administrative formats. The grade band 11 ELL sample included only two students and is not reported.

Table 25. Percentage of Students Expressing Confusion with the Different Components of the Test by Administration Format, Grade Band, and Sample

		All Items			One at a Time		
		Grade Band			Grade Band		
Sample	**Test Component**	**3**	**6**	**11**	**3**	**6**	**11**
Gen Ed	Passage	33	29	17	0	43	14
	Items	25	30	17	9	36	18
ELL	Passage	50	50		25	50	
	Items	32	50		16	50	

Table 26. Percentage of the Students Expressing Frustration with the Different Components of the Test, by Administration Format, Grade Band, and Sample

		All Items			One at a Time		
		Grade Band			Grade Band		
Sample	**Test Component**	**3**	**6**	**11**	**3**	**6**	**11**
Gen Ed	Passage	0	29	17	0	29	14
	Items	13	18	17	13	11	14
ELL	Passage	13	38		13	38	
	Items	3	41		3	50	

Table 27 presents the average score students obtained for the think-aloud protocols. The grade band 6 students tended to score higher when the items were presented all at one time (for both the Gen Ed students and the ELL students). The grade band 3 students scored higher when the items were presented one at a time, regardless of sample or testing condition. The grade band 11, Gen Ed students scored higher when the items were presented one at a time, while the grade band 11, ELL sample students scored higher when the items were presented all at one time, though the latter sample size is small.

Table 27. Average Score, by Administration Format, Grade Band, and Sample

	All Items at Once			One Item at a Time		
	Grade Band			Grade Band		
Gen Ed	2.2	3.0	1.8	2.5	2.3	2.5
ELL	2.4	2.9	2.0	2.5	1.7	1.5

Table 28 shows the preference for a presentation format. Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. ("I preferred one at a time—less confusing than seeing too many questions," "One at a time made me less nervous about how many more there were," "I liked one at a time because it did not seem overwhelming.") Grade band 11 students (Gen Ed and ELL) had a slight bias toward having the items presented one at a time ("Let's me focus on that one question"). Conversely, grade band 6 Gen Ed students preferred to have the items presented together ("I liked them altogether," "This way I know I was on the same passage," "All together, you can refer to the questions while you read the passage," "I liked everything on one page because it was more easy," "With all together, I was able to refer back and I could see where I was going," "I liked altogether, though it was more confusing and distracting.") The grade band 6 ELL students were equally divided between the two formats.

Table 28. We Presented the Questions to You in Two Different Ways. Which Way Did You Prefer: All Together or One at a Time (Percent Responding)?

	Grade Band								
	3			6			11		
Sample	All Together	No Preference	One at a Time	All together	No Preference	One at a Time	All Together	No Preference	One at a Time
Gen Ed	33		67	57	29	14	29	14	57
ELL	14		86	43	14	43		50	50

Summary

We were interested in assessing whether there is a difference in timing and increased negative emotional states (confusion, frustration) when students are presented a passage with all the items or are presented a passage with the items presented one at a time. Forms were administered to two groups of students: a group that received English language accommodations and a Gen Ed group.

The time it took to complete the sets when all items were presented together or one at a time varied by grade band and sample. For the grade band 3 and grade band 11 samples, timing differed little whether the items are presented in one block or one at a time. However, for grade band 6, presenting the items one at a time took substantially longer for both the Gen Ed and ELL samples. While there is some variability between the ELL and the Gen Ed samples, the differences are not large and show the same pattern within grade band.

There appeared to be slightly more *confusion* for both the Gen Ed and the ELL samples in grade band 3 when all the items were presented together. However, similar *frustration* levels were observed under the two formats for the grade band 3 students. The grade band 6 ELL sample students showed similar patterns of frustration and confusion for the two presentation formats. However, the Gen Ed grade band 6 students showed slightly more confusion when the items were presented one at a time.

The grade band 6 students tended to score higher when the items were presented all at one time (for both the Gen Ed students and the ELL students). The grade band 3 students showed similar results, regardless of sample or administration format. The grade band 11, Gen Ed students scored higher when the items were presented one at a time, while the grade band 11 ELL sample students scored higher when the items were presented altogether.

Both the ELL and Gen Ed grade band 3 students preferred to have the items presented one at a time. Grade band 11 students had a slight bias toward having the items presented one at a time. Conversely, grade band 6 students preferred to have the items presented together.

Research Question 8: Smarter intends to present relatively long passages. Do longer passages reduce student engagement?

Smarter Balanced is interested in using passages that are longer than those presently used. The Smarter Balanced recommended passage lengths are: for grades 3–5: 450–562 words for short passages and 563–750 words for long passages; for grades 6–8: 650–712 words for short passages and 713–950 words for long passages; and for high school, 800–825 words for short passages and 826–1100 words for long passages. There is concern that the longer passages may tax the processing abilities of ELL and SWD students.

This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Each form contained a standard-length and an extended-length passage. The first set contained a passage of standard length. The second set contained a passage that is longer than standard length (extended-length, the length equivalent to that intended for use by Smarter Balanced).

Forms were constructed in ELA at three grade bands: grade band 3–5 (referred to as grade band 3), grade band 6–8 (referred to as grade band 6), and grade band 10 and 11 (referred to as grade band 11). The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands (3, 6, and 11). Thirteen students took the forms. Of these, nine were grade band 3 Gen Ed students. One grade band 3 student was classified ELL/SWD. The single grade band 6 student was an ELL/SWD student. The two grade band 11 students were Gen Ed students.

Results

Table 29 shows the percentage of students whose engagement was improved or unaffected by the longer passage, by subgroup. All the ELL/SWD students were unaffected by the use of the longer passage. Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. All the ELL/SWD students were able to read the entire passage regardless of passage length. Only about 25 percent of the grade band 3 Gen Ed students and none of the grade band 11 Gen Ed students were unaffected by the use of the longer passage (see Table 29; “I have to read the whole passage?”). The ELL/SWD students all demonstrated that the longer passage was processed at a deep level (“It was a good story”). However, only 43 percent of the Grade band 3, Gen Ed, students demonstrated a level of deep processing (“I learned many new things”) and only 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing (Table 31). The ELL/SWD students were not bored or distracted while reading either passage; however, some percentage of the Gen Ed students were bored regardless of the length of the passage.

Table 29. Percentage of Students Whose Engagement Is Improved or not Affected by the Longer Passage

	Grade Band		
Subgroup	**3**	**6**	**11**
GE	25		0
ELL/SWD	100	100	

Table 30. Percentage of Students Who Appear to Read the Entire Passage

Standard Length	Grade Band		
Subgroup	**3**	**6**	**11**
GE	88		100
ELL/SWD	100	100	
Extended Length	Grade Band		
Subgroup	**3**	**6**	**11**
GE	88		50
ELL/SWD	100	100	

Table 31. Percentage of Students Whose Think-Aloud Demonstrate Deep Processing as Assessed by the Interviewer

Standard Length	Grade Band		
Subgroup	**3**	**6**	**11**
GE	43		100
ELL/SWD	100	100	
Extended Length	Grade Band		
Subgroup	**3**	**6**	**11**
GE	43		50
ELL/SWD	100	100	

Table 32. Percentage of Students Who do not Appear Bored or Distracted

Standard Length	Grade Band		
Subgroup	**3**	**6**	**11**
GE	63		100
ELL/SWD	100	100	
Extended Length	**Grade Band**		
Subgroup	**3**	**6**	**11**
GE	88		50
ELL/SWD	100	100	

Summary

Smarter Balanced is interested in using passages that are longer than those presently used. There is concern that the longer passages may tax the processing abilities of ELL and SWD) students. This question is designed to assess whether longer passages reduce student engagement, hamper the completion of the longer passages, or affect the depth of processing of the passage. The design was intended to compare the performance of two groups of students—ELL/SWD and Gen Ed students—across three grade bands (3, 6, and 11). Two sets of items were created. Both sets contained passages of equivalent difficulty with four items of equivalent difficulty attached to each passage. Both sets present the passage and all the items together. Both the standard-length and the extended-length passage were included in a given form and administered to the same student.

All the ELL/SWD students were unaffected by the use of the longer passage. They were able to read the entire passage regardless of passage length and demonstrated that the longer passage was processed at a deep level. The ELL/SWD students also were not bored or distracted while reading either passage.

On the contrary, Gen Ed students did appear to be affected by the longer passage in grade bands 3 and 11. About 75 percent of the grade band 3 students and all of the grade band 11 students were affected by the use of the longer passage. Only 43 percent of the Grade band 3 Gen Ed students demonstrated a level of deep processing and only 50 percent of the grade band 11 Gen Ed students demonstrated a level of deep processing. Also, some percentage of the Gen Ed students were bored, regardless of the length of the passage

Research Question 9: How long does it take for students to read through complex texts, performance tasks, etc.? Is timing affected by the way students are presented the passage and items?

One way of making items more difficult is to increase their complexity. Complex items often take longer to solve or answer. In computer adaptive tests, added complexity may decrease the time a high ability student has to complete the test if the items are made more difficult through increased complexity. This potentially creates some fairness issues in an adaptive test if there is a time limit on the test. This question was designed to assess the time it takes for students to answer complex and simpler items. Complexity was defined as a function of the DOK demanded by the test question. It was hypothesized that more complex tasks would take more time.

Each ELA form had six items. These items varied in item complexity (simple or complex) and item format (SR, TE, or CR). The TE items were all “hot text” items. These items require the student to either highlight the text or drag the text to answer the item.

Forms were constructed in ELA at two grade bands: grade band 3–5 (referred to as grade band 3) and grade band 6 and 7 (referred to as grade band 6). Two forms were administered in grade band 3. One form was administered in grade band 6.

Results

Eight students took the grade band 3 forms with four students taking each form, and two students took the grade band 6 form.

Table 33 presents the average time (in seconds) a student took to answer an item. SR items were answered in the shortest time. HT items took about one minute longer than the SR items. CR items took the most time to answer, about 75 seconds longer than the “hot text” items. With the exception of the complex CR item administered to grade band 6 students, item complexity did not seem to have an impact on item performance. (An interviewer commented, “Student took about the same time for complex and easy items.”)

Table 33. Average Time (in seconds) to Answer an Item by Grade Band, Item Type, and Item Complexity

			Grade Band	
			3	6
Item Format	Difficulty	Item	Avg. Time	Avg. Time
SR	Simple	1	49	52
SR	Complex	2	29	59
TE (HT)	Simple	3	83	126
TE (HT)	Complex	4	96	123
CR	Simple	5	182	168
CR	Complex	6	158	185

Table 34 presents a summary of the average time students took to complete complex and simple items across item types by grade band. Complex items seemed to have more impact in grade band 6, but there is no evidence that complex items, as defined here, take longer than simpler items.

Table 34. Interviewer's Summary of Item Timing by Grade Band and Item Difficulty

	Grade Band	
	3	6
Difficulty	Avg. Time	Avg. Time
Simple	104	115
Complex	94	126

Summary

It was hypothesized that more complex items would take longer to complete than simpler items. No evidence was found to support this hypothesis. In terms of the time spent on an item, SR items were answered in the shortest time. "Hot text" items took about one minute longer than SR items. CR items took the most time to answer, about 75 seconds longer then the "hot text" items.

Effective Communication of Mathematics

Research Question 10: Working mathematics problems on computer: Communicating mathematics on computer—feasibility of measuring student understanding of items for Claims 2–4 on computer.

With paper tests some students write in their test books while working out mathematics problems. When mathematics items are presented on computer, scratch paper is often provided if students want to transfer the problem to paper and work it out there. Because scratch paper is often destroyed after an online testing session, the degree to which scratch paper is used is not known; neither is the importance of scratch paper in working out a problem (or potentially for use in scoring). This research question examines the need for paper when solving mathematics problems. Forms were constructed at four grade bands: grade band 3 and 4 (referred to as grade band 3), grade band 6 and 7 (referred to as grade band 6), grade band 7 and 8 (referred to as grade band 7), and grade band 11 (referred to as grade band 11) to investigate whether the scratch paper usage was uniform or varied by educational level.

Each student was presented with three grade-appropriate items. The interviewer recorded whether the student made a comment, and the nature of the comment, while working the mathematics problems. The students first tried to work the problem without paper. Scratch paper was then offered to the student to rework the problem, if desired. The interviewer noted whether students chose to add anything additional and noted the nature of the addition (more text, equations, graphics). Note that there were only three comments for the third item in the lowest grade band, 3.

Results

Twenty students were administered the grade band 3 form, 37 students were administered the grade band 6 form, 21 students were administered the grade band 7 form, and 19 students were administered the grade band 11 form.

Table 35 shows the percentage of comments made for an item and the type of comment made. Two types of comments were of interest: did the students who wanted paper draw a picture or write an equation or did they find the online system difficult to use. The lowest grade band students (grade band 3) did not need paper to solve any of the problems (Table 635. Some students in the highest grade band (grade band 11) commented that they would like to draw a picture for the items they were administered (15–30 percent). ("I wanted to graph the area.") There was also one item (Item 2) for which about 15 percent of students wanted paper to write equations. About 5–10 percent of students in each grade band found the online system difficult to use. ("Confused me, I didn't know how to write an equation," "Tried the keypad, but it wouldn't work," "It was much easier with paper.") The strongest result came from the grade band 6 and grade band 7 groups, where 30 to 42 percent of the sample, respectively, indicated that they wanted to write an equation. Between 3 and 23 percent of the grade band 6 and 7 groups also indicated that they wanted to draw a picture. This may be a function of newly introduced algebra concepts for this group.

Table 35. Percentage of Comments for an Item, by Question Type and Grade Band

		Grade Band			
Question	**Item**	**3**	**6**	**7**	**11**
Picture	**1**	5	0	23	32
	2	15	12	3	16
	3	0	4	6	16
System Difficulty	**1**	5	9	10	5
	2	11	3	10	5
	3	0	4	7	5
Equation	**1**	0	31	45	6
	2	0	32	34	16
	3	0	29	43	6

Table 36 shows the nature of the student comments made on paper and whether the additional information recorded on the paper improved the response according to the rubric. For all grade bands the additional information recorded on the paper included a graphic. In grade bands 6, 7, and 11, the additional information recorded on paper included an equation. The grade band 6, 7, and 11 groups provided additional information on paper that improved the response according to the rubric. For example, one administrator noted, “When given paper, she was able to do the proper equation and solve for x. She was more confident with paper and pencil.” The number of cases in which improvement was observed varied by item. For grade band 6, item 2, about 11 percent of the responses were improved when scratch paper information was taken into account during scoring. For grade band 11, item 3, about 16 percent of the responses were improved when scratch paper information was taken into account during scoring. Responses to all items in grade band 7 were improved when scratch paper information was taken into account. The improvement for this group ranged between 10 and 20 percent across items.

Table 36. Percentage of Changes Made When Paper Was Introduced

		Grade Band			
Nature of Students' Changes	Item	3	6	7	11
No Additions Made	1	80	57	71	53
	2	60	65	67	63
	3	10	32	52	58
Addition Included Graphic	1	5	3	33	37
	2	15	5		11
	3			10	32
Addition Included Equation	1		22	19	5
	2	20	16	38	16
	3		19	38	11
Addition Improved Response According to Rubric	1		11	14	
	2		3	29	
	3			24	16

The interviewer's comments suggested that most students in grade band 3 (75 percent) and grade band 11 (63 percent) were able to accurately respond to the mathematics items they saw only using the online text editor. However, fewer than half of the students in grade band 6 (45 percent) could accurately respond to questions using only the text editor and only 13 percent of the students in grade band 7 were observed to be able to accurately respond to questions using only the text editor. One student commented, "It's much easier with paper."

Summary

The general conclusion is that a subset of students benefit from being able to work mathematics problems on paper. It appears to be especially important when students are beginning to learn algebra concepts.

Grade band 3 students did not need paper to work the problems. However, in the grade band 6 and grade band 7 groups, 30–42 percent of students indicated that they wanted to write an equation. In grade bands 6, 7, and 11, the additional information recorded on paper would have improved the response according to the rubric. Responses for specific items in grade bands 6 and 11 were improved by 15 percent of the students and responses for all items in grade band 7 were improved when information on the scratch paper was taken into account. Improvement for this group ranged between 10 and 20 percent of the responses. This was supported by interviewer observations. About 5–10 percent in each grade band found the online system difficult to use, but few specifics were recorded.

Research Question 11: Usability of equation editor tool—can students use the tool the way it is meant to be used?

Although students begin to use technology at a very early age, it is prudent to verify that young students are able to use the assessment interface to be used during testing. This question sought to evaluate the ability of grade 3–5 students to use the equation editor tool to be included in the Smarter Balanced delivery system. Three mathematics items were presented to the students (*N*=33). The first item only required the student to copy his or her response. The second item was a simple mathematics item and the third item was a more challenging mathematics item. The first item would demonstrate whether the student could use the equation editor tool. The second and third items would provide evidence of whether the ability to use the tool interacted with item difficulty.

Results

Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. About 30 percent had trouble just copying the answer, as required by item 1. The examiners assessed that 35 percent had difficulty using the equation editor and that only 40–57 percent of the students would get a given item correct. Students had more difficulty with the more challenging items. A summary of representative comments made by students about the equation editor during the administration of the think-aloud protocol is presented below:

1. Clicked on the + sign, but it didn't work, twice.
2. How do I choose the numbers?
3. I needed paper to make a picture.
4. How do I use the number pad?
5. I tried to use the numbers on the keyboard, but wouldn't work.
6. Some symbols didn't respond to first click.
7. I had trouble getting bottom half of fraction to record.
8. Unclear what possible value meant.
9. I didn't see decimal point down there [due to scrolling].
10. Couldn't find x symbol.
11. Unclear whether to click and drag or type.
12. Would rather type than use a mouse.
13. Difficult to use fraction tool.

Summary

Elementary students had some difficulty using the equation editor. Between 15 and 30 percent of the students indicated that they had difficulty using the equation editor. The examiner's assessment concurred that about 35 percent had difficulty using the equation editor and that about 50 percent of the students would get a given item correct.

Research Question 12: Intuitive understanding of the relationships in multiplying fractions.

This question is designed to assess whether students with a strong understanding of fractions and the multiplication and division of fractions complete the items without performing the indicated multiplication. The task asked students to compare the size of a product to the size of one factor, on the basis of the size of the other factor, without performing the indicated multiplication. Also of interest was whether students who complete an item as intended (without using multiplication) spent less time on an item than those who did not. To investigate this question a single form was administered for grades 3–5.

Results

The form was administered to 33 students at the elementary level. Table 37 compares those with a strong understanding of fractions with those who do not have a strong understanding of fractions and whether they completed the task with or without using multiplication. There does not appear to be a relationship between strength of understanding of fractions (multiplication and division) and whether they used multiplication to solve the problems.

Table 37. Strength of Understanding of Fractions and Whether Multiplication was Performed

	Not Strong Understanding of Fractions		Strong Understanding of Fractions	
Item Number	Performed Multiplication	Did not Perform Multiplication	Performed Multiplication	Did not Perform Multiplication
1	9	7	8	1
2	9	8	9	1
3	10	6	6	1
4	6	7	10	1
5	7	7	9	1
6	4	6	15	0

Table 38 presents descriptive statistics for the timing of each item (in seconds). In addition to means, medians are reported because timing distributions tend to be highly skewed. On average, those who did not have to perform the multiplication completed the items in less time. The results for item 6 were comparable for the two groups.

Table 38. Comparison of the Time to Complete the Item for Those Who Did not Use Multiplication to Solve the Item and Those Who Did

	Performed Multiplication				Did not Perform Multiplication			
Item Number	Mean	Std Dev	Median	Range	Mean	Std Dev	Median	Range
1	210	136	179	59–543	136	90	114	53–360
2	145	119	106	36–420	126	110	89	30–336
3	75	104	42	10–480	34	28	25	3–90
1	123	111	70	21–480	88	69	57	25–195
2	133	130	95	28–480	79	67	68	9–185
3	69	118	32	4–540	65	63	51	3–170

Table 39 shows the percentage of students answering the item correctly. The students tested generally found the items to be difficult. ("Multiplying fractions was hard.") Some students did not understand the inequality signs, while others did not understand improper fractions or how to make a whole number into a fraction. One interviewer commented that the "student had little or no understanding of fractions."

Table 39. Percentage of Students Answering an Item Correctly.

Item Number	Percent
1	17
2	20
3	28
4	42
5	26
6	33

About 69 percent of the students used multiplication to solve the problems (Table 40). Student comments support this. "I multiplied… each box and put them in the correct boxes (columns)." "I timesed [sic] the numbers." "I looked at each number expression and multiplied it in my head and moved it to where I thought it was right." "Some numbers on the bottom depends on the top number which is bigger or smaller." Only about 40 percent of the students understood fractions or at least the multiplication of fractions. The examiner's comments (Table 41) concur with this conclusion.

Table 40. Percentage of Students Using Multiplication to Solve the Items

Item Number	Yes
1	68
2	70
3	75
4	72
5	73
6	81

Table 41. Interviewer's Assessment of: (1) Whether the Student Used Multiplication and (2) Whether the Student Had a Strong Understanding of Fractions

Summary	Percent
Did student use multiplication?	72
Did student have a strong understanding of fractions (multiplication/division)?	40

Summary

There seemed to be little relationship between whether a student has a strong understanding of the multiplication and division of fractions and whether he or she used multiplication to solve the items. However, students who did not have to perform the multiplication completed the items in less time than students who had to perform the multiplication. While most students said they understood the questions, 70 percent had to use multiplication to solve them. Only about 40 percent of the students had a firm understanding of the multiplication/division of fractions, according to the interviewers.

Special Populations

Research Question 13: Contextual glossaries are item-specific glossaries that provide a definition of a word that is targeted to, and appropriate for, the context in which the word is used in the item. Are these a fair and appropriate way to support students who need language support?

This question addressed the efficacy of the use of contextual glossaries with non-native (Spanish) speakers (see Exhibit 4 for an example of a contextual glossary item) when solving mathematics problems. A contextual glossary item contains highlighted words when presented online. Clicking any of these highlighted items produces a list of all highlighted words in the item with Spanish definitions for each. Two sets of items were created that were parallel in difficulty. The first set of items contained no contextual glossaries with only single words translated. The second set of items contained contextual glossaries. The interviewer was asked to determine whether the student was having trouble understanding a word and whether the contextual glossary aided in the interpretation of the word or sentence.

Only three ELL students participated: one from grade 3 and two from grade 6.

Exhibit 4. Example of a Contextual Glossary Item

1. A roller coaster has a large rise and drop followed by a complete circle. The following diagram shows measurements for the track. An extra 20 feet are needed for cutting and welding. How many feet of track should be ordered? (Use π = 3.14)

A. 280 feet

B. 407 feet

C. 415.6 feet

D. 1,537.4 feet

Glossary Window

Roller coaster

montaña rusa

Rise

subida

Drop

bajada
caída

Complete

completo
entero

Diagram

diagrama
quema
gráfico

Track

vía
riel

Cutting

cortar

Welding

cortar

Results

The grade 3 student had trouble understanding a few items, but had few word confusions. For the second set of items, this student used the contextual glossaries for one item but not for the other items. The student said that there was not a problem understanding the items because the student used “sentence context” to answer them, or the words the student didn't know weren't in the glossary so the student stopped using it. In terms of scoring, this student answered two of the three “translated” items correctly, but did not answer any of the “contextual glossary” items correctly, so the results are difficult to interpret as to whether the use of contextual glossaries aided the students’ performance.

The two grade 6 students (one ELA form and one Math form) both had difficulty with the “translated” items in the first set with six or more word confusions each for most items. Both students found the contextual glossary useful to some degree, though not for all items. (“The words I don’t know aren’t in the glossary.”) However, the interviewers suggested that the use of the contextual glossary improved the performance for both grade 6 students. Though the ELA student got all questions incorrect, the interviewer believed that this was mainly due to careless mistakes and that the student used the glossary to help make sense of the key components of the questions and understood the procedures for answering the questions. The math student got two-thirds of the items correct when the items were translated, and one-third of the items correct when the contextual glossary was used. The student had difficulty understanding an essential word in one of the incorrect items. However, the interviewer commented that once he understood the words, he could confidently work on the problem and he knew how to proceed.

Summary

In summary, contextual glossaries appeared to be somewhat effective when they were used, but the impact was not always reflected in the score the student received for an item. The contextual glossaries appeared to be incomplete in that they did not include words that the students needed. This limited the use of the glossaries in these situations. Interviewer’s comments suggested that performance was improved when the students used the contextual glossaries.

Research Question 14: Under what conditions do students with lower reading ability use text-to-speech (TTS) to help focus on content in ELA and mathematics? Is this affected by the quality of the voice-pack?

TTS is a technology that can give students with low reading ability access to an assessment. For this technology to be effective the language produced from the voice-pack must be clear so that it can be understood. This is particularly true for non-native speakers of English.

This question is designed to assess whether students with lower reading ability and non-native speakers of English use TTS to help focus on content in ELA and mathematics. Only students familiar with TTS were included in the study. Overall, 77 students used TTS at least once. Among them, 58 students are LEP students, 13 students had reading difficulties (IEP), and six students were Gen Ed students.

Forms were constructed at three grade bands: grade band 3 (referred to as grade band 3), grade band 6 and 7 (referred to as grade band 6), and grade 11 (referred to as grade band 11). In ELA, four forms were administered with both high- and low-quality voice-packs. In mathematics, two forms were administered in grade bands 3 and 11. Only a single form was administered in grade band 6. For all mathematics forms only high-quality voice-packs were administered. In Tables 42–45, yellow shading denotes the use of high-quality voice-packs while a white background denotes the use of a low-quality voice-pack.

Results

For ELA (Table 42), for all groups and grade bands, a high percentage of students tended to make comments indicating an improved focus on the content when the voice-pack was of high quality. About one-third of the students (except the Gen Ed grade band students) indicated that TTS kept their focus on content even when low-quality voice-packs were used. For ELA, students in all groups tended to make greater use of TTS when the voice-pack was of high quality.

About 50 percent of the LEP students in mathematics in grade bands 3 and 11 made comments indicating that TTS helped them focus on content. All of the LEP grade band 6 group and the IEP students in grade band 3 found that TTS helped them focus on content. (“It made me think about the question.”) The Gen Ed students in grade band 3 found that TTS helped them focus on content; however, the Gen Ed grade band 6 students did not find TTS useful.

Table 42. Percentage of TTS Students Who Made Any Comment Indicating That He/She Is Mainly Focused on the Content of the Item, by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	**Voice Pack Quality**	**Grade Band**	**3**	**6**	**11**	**3**	**6**	**11**	**3**	**6**	**11**
ELA	**Low**			32	39	35				0	
	High		36	67	100	100					100
Mathematics	**Low**										
	High		50	100	60	100			100	0	

Table 43 shows the percentage of students who answered the items correctly, averaged across items. In ELA, the grade band 6 and 11 LEP students and the grade band 3 IEP students found the items more difficult using a low-quality voice-pack. The Gen Ed grade band 6 ELA students were not administered a high quality voice-pack. In the LEP grade band 6 group, about half the students answered an item correctly using the high-quality voice-pack. The percentage answering an item correctly was close to 75 percent for the other LEP grade bands and the grade band 3 low-level reading students when the high-quality voice-pack was used.

In mathematics, in grade band 3, about 40 percent of the LEP students answered an item correctly. For the other grade bands, for the LEP and IEP samples, no items were answered correctly, even with the high-quality voice-packs. This was also true for the Gen Ed grade band 3 students. However, the general education students in grade band 6 answered all the items correctly.

Table 43. Percentage of TTS Students Who Answered the Items Correctly by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			14	0	50				75	
	High		77	50	80	75					0
Mathematics	Low										
	High		40	0	0	0			0	100	

Tables 44 and 45 summarize the interviewer's assessment for ELA and mathematics related to whether TTS improved access to the content or was a distraction. TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used in ELA except in grade band 11. This is probably an artifact of the very small sample size. The low-quality voice-pack appeared less effective at providing access and was distracting in ELA, where the high-quality voice-pack was not distracting at all. One student said, "[I] didn't like using TTS … the sound was robotic and would break my concentration."

In mathematics, TTS helped to improve access for some grade band 3 LEP students, but not for middle- and upper-level LEP students or the IEP or Gen Ed grade band 3 students. All the Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting in mathematics. This was in part a function of trying to describe a table verbally. ("When TTS read the chart aloud, I got lost in the numbers and couldn't figure out what the question was asking.")

Table 44. Assessment by the Interviewer of the Percentage of TTS Students Whose Access to Content Was Improved by the Use of TTS by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	**Voice Pack Quality**	**Grade Band**	**3**	**6**	**11**	**3**	**6**	**11**	**3**	**6**	**11**
ELA	**Low**			57	75	79				100	
	High		76	100	33	100					100
Mathematics	**Low**										
	High		43	0	0	0			0	0	

Table 45. Assessment by the Interviewer of the Percentage of TTS Students Who Were Distracted by TTS, by Content, Voice-Pack Quality, Sample, and Grade Band

			LEP			IEP			Gen Ed		
Content	Voice Pack Quality	Grade Band	3	6	11	3	6	11	3	6	11
ELA	Low			12	20	33				0	
	High		0	0	0	0					0
Mathematics	Low										
	High		44	100	40	0			100	100	

Summary

TTS improved access in ELA regardless of the quality of the voice-pack. Greater access was achieved when high-quality voice-packs were used. LEP students and students with reading difficulties tended to benefit more from the use of TTS. Using TTS with high-quality voice-packs improved focus on content in ELA. The use of TTS with low-quality voice-packs tended to distract students in ELA, whereas high-quality voice-packs did not. In mathematics, access was improved only for grade band 3 students. All the Gen Ed, IEP, and grade band 6 LEP students found the high-quality voice-pack distracting. This was in part a function of trying to describe a table verbally.

Final Summary

Smarter Balanced is moving toward an assessment model that is largely scored automatically and delivered adaptively on computer. The Smarter Balanced cognitive laboratories were conducted to investigate questions that arise from such an automated design. While think-aloud protocols are time consuming, they have the potential to provide a level of information not easily accessed through large-scale studies. However, the sample sizes are small. Therefore, should a more rigorous investigation of any of the research questions be of interest, specifically designed studies with large samples will be needed.

This report presents the results from 14 small think-aloud studies that addressed topics that pertain to an automated test delivery system.

1. Can non-constructed-response item formats assess components that have historically been believed to be measured only with CR items?
2. What is the optimal amount of direction to provide for TE items? Does this vary with grade level?
3. What is the appropriate degree of labeling to provide for MPSR items so that students know to complete all parts?
4. Does it matter whether items associated with a passage are presented in a single block or presented one item at a time? Are ELL students impacted by these different arrangements?
5. Do the longer passages favored by Smarter Balanced reduce student engagement?
6. How much time do items in different formats take to answer? Are ELL students affected more than general education students?
7. In mathematics, could information captured on scratch paper facilitate the working of a problem and benefit the performance and scoring of a student?
8. Do contextual glossaries help improve the performance of students with language disabilities?
9. Does TTS help focus students of low reading ability on the content of an item?
10. Can younger students effectively use the equation editor?
11. Mathematics intuition: Can students compare the size of a product to the size of one factor, on the basis of the other factor without multiplying?

On the whole, the cognitive laboratories were successful in providing answers to most of these questions. They provide a glimpse of issues that may exist and need to be investigated further. To investigate these issues more completely, larger-scale studies should be conducted.