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

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

"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).

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.

Date	Action
2010	Florida State Board of Education voted to adopt the Common Core State Standards (CCSS) with a four-phase implementation plan beginning in the 2011-12 school year with full implementation to occur during the 2014-15 school year.
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.
August 2013	Governor Rick Scott convened the state's top education leaders and bipartisan stakeholders to discuss the sustainability and transparency of the state's accountability system in a three-day accountability summit.
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.
October 2013	Invitation to Negotiate was posted for public review

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.

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.

> “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)

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

		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	✓			✓	✓	✓	✓	✓	✓
	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

Evaluation Target Areas	AERA et al. (2014) Source of Validity Evidence				
	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
Evaluation of Scoring, Scaling, and Equating	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

Evaluation Target Areas	AERA et al. (2014) Source of Validity Evidence				
	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' Ratings	FLDOE/AIR 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' Ratings	FLDOE/AIR 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' Ratings	FLDOE/AIR 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' Ratings	FLDOE/AIR 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' Ratings	FLDOE/AIR 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' Ratings	FLDOE/AIR 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.

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.

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).

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.

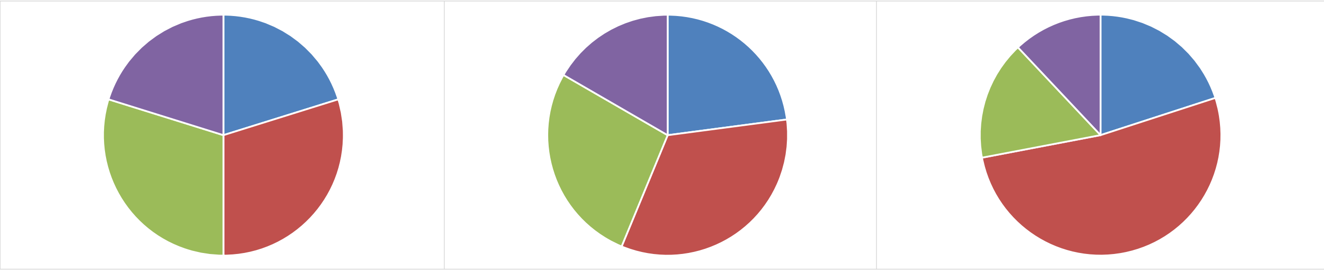
The results for ELA are generally positive, although a few of the categories were either under- or over-represented as indicated in the panelists' ratings. This result emerged even with the general agreement between the vendor ratings of the items and the panelist ratings described in Study 1. When there 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



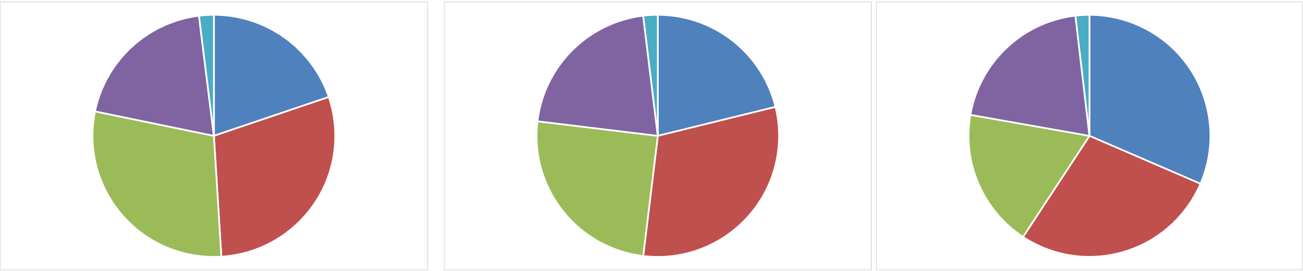
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
🔵 Key Ideas and Details	0.21	0.23	0.20
🔴 Craft and Structure	0.31	0.33	0.53
🟢 Integration	0.31	0.27	0.16
🟣 Language/Editing	0.21	0.17	0.12

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



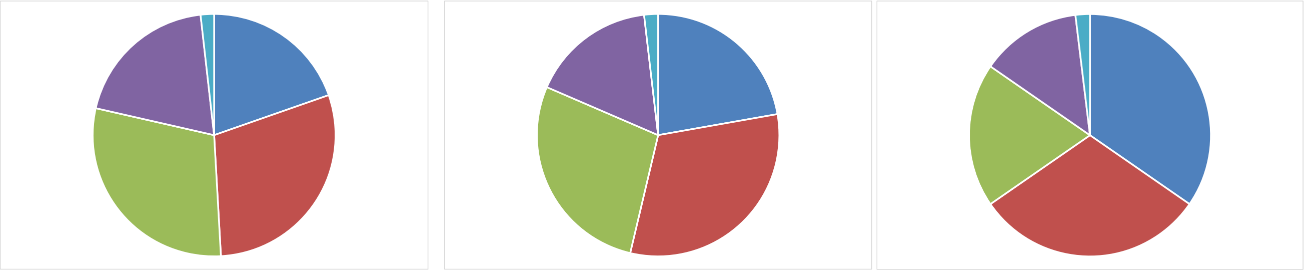
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Key Ideas and Details	0.20	0.21	0.31
● Craft and Structure	0.30	0.31	0.28
● Integration	0.30	0.25	0.19
● Language/Editing	0.20	0.21	0.20
● Text based writing	0.02	0.02	0.02

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



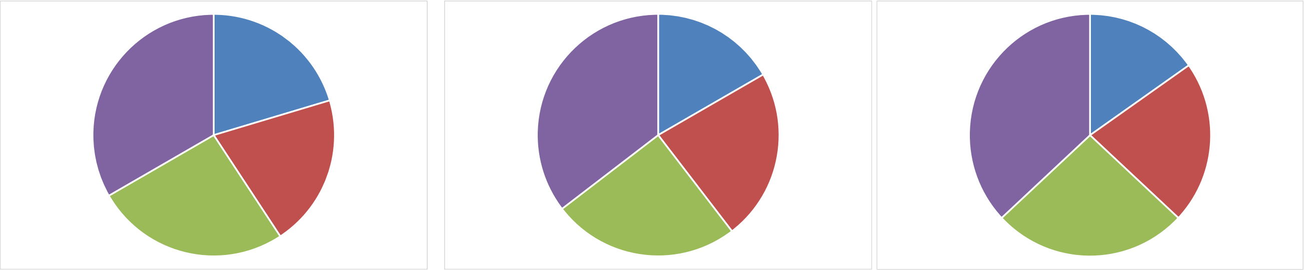
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Key Ideas and Details	0.20	0.22	0.35
● Craft and Structure	0.31	0.31	0.31
● Integration	0.31	0.28	0.19
● Language/Editing	0.20	0.17	0.13
● Text based writing	0.02	0.02	0.02

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



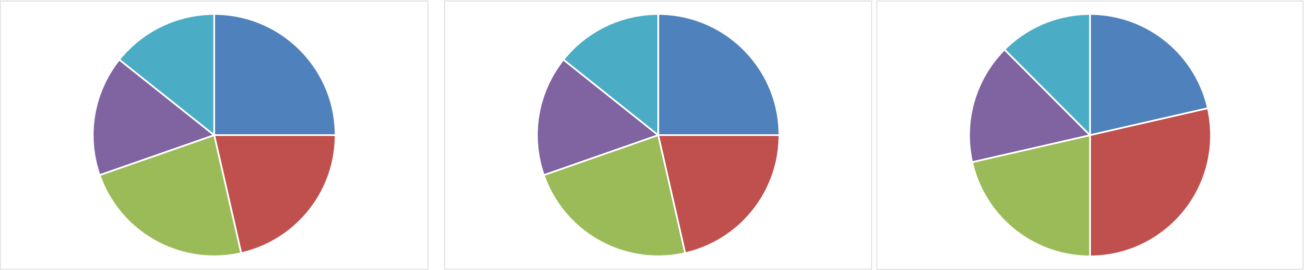
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
🔵 Operations and Algebraic Thinking	0.20	0.17	0.15
🔴 Numbers and Operations Base 10	0.20	0.23	0.22
🟢 Numbers-Operations – Fractions	0.26	0.25	0.26
🟣 Measurement, Data, Geometry	0.33	0.35	0.37

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



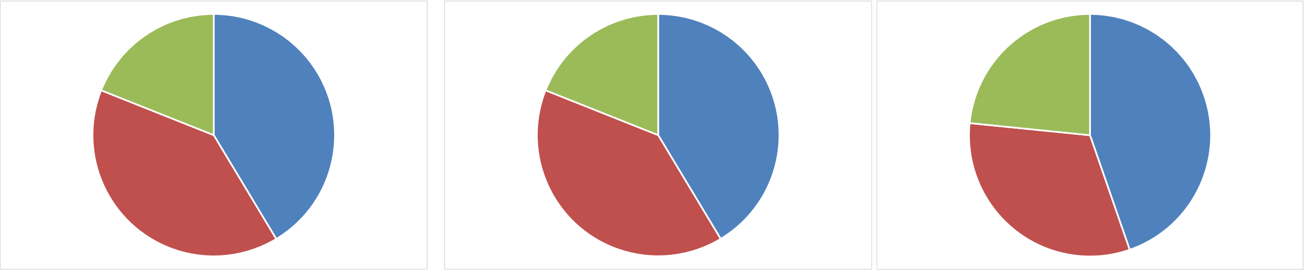
	Blueprint	Items as Rated by Vendor	Items as Rated by Panelists
● Ratio & Proportional Relationships	0.25	0.25	0.21
● Expressions & Equations	0.21	0.21	0.29
● Geometry	0.23	0.23	0.21
● Statistics & Probability	0.16	0.16	0.16
● The Number System	0.14	0.14	0.13

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)

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.

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)

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.

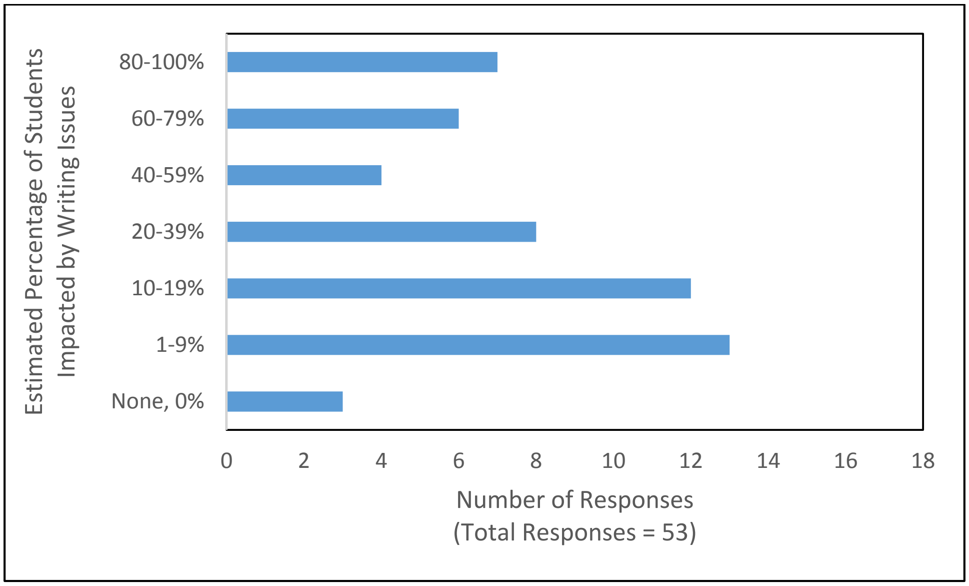


Figure 7: District Representatives' Estimated Percentage of Students Impacted by Writing Technology Issues

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 0.75 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.

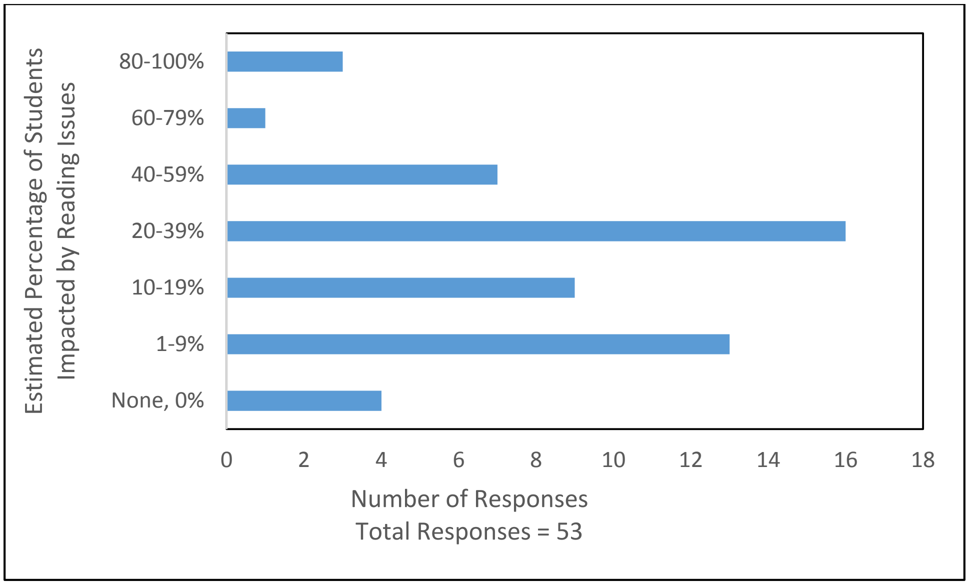


Figure 8: District Representatives' Estimated Percentage of Students Impacted by Reading Technology Issues

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.

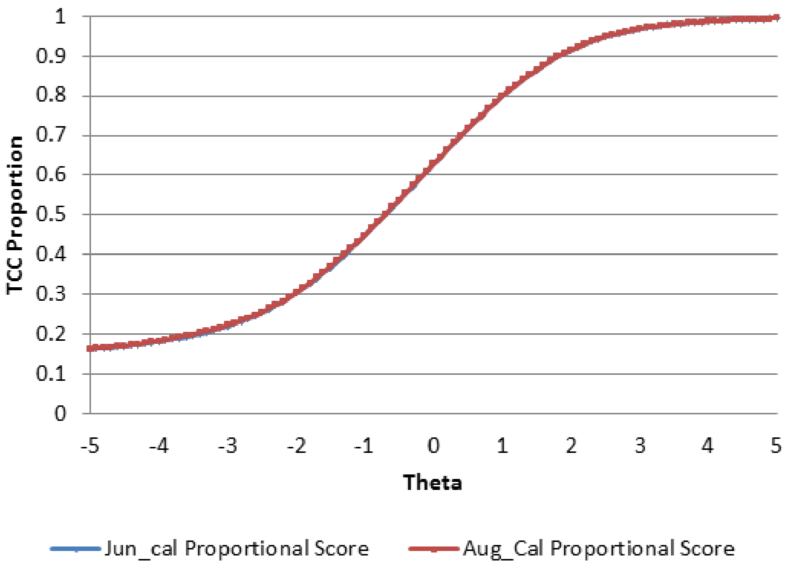


Figure 9. Test characteristic curve for Grade 10 ELA Florida Standards Assessments, with impacted students included and with impacted students removed.

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.

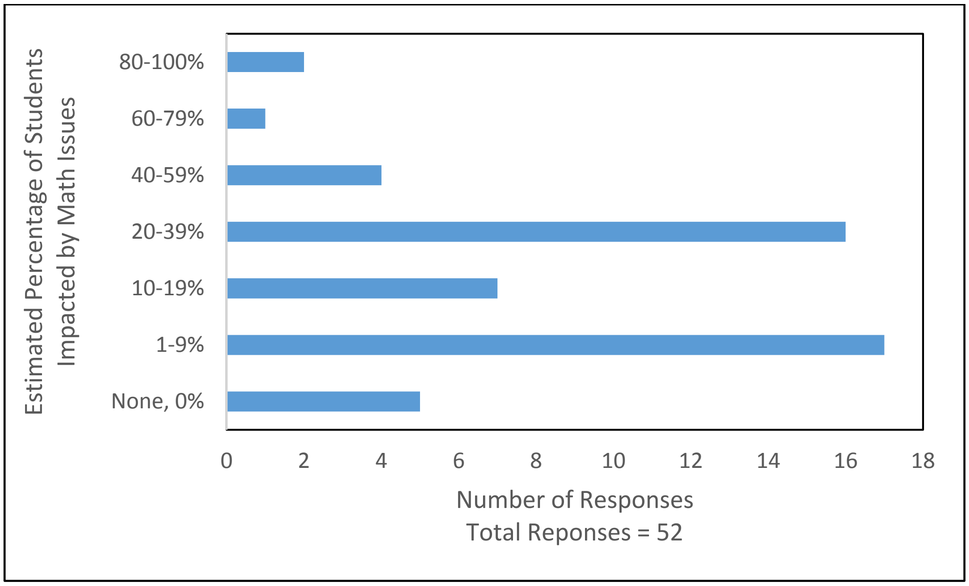


Figure 10: District Representatives' Estimated Percentage of Students Impacted by Math Technology Issues

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

Math	Total Students Tested (Statewide)*	Students in Session on Multiple Days	
		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

		Completed all sessions, 1 day	
Math	**Total Students Tested (Statewide)***	**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

Math	Total Schools Administered Assessment	Schools with Students Who Completed Math Session in One Day		
		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

Date	Time	Max Users
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.

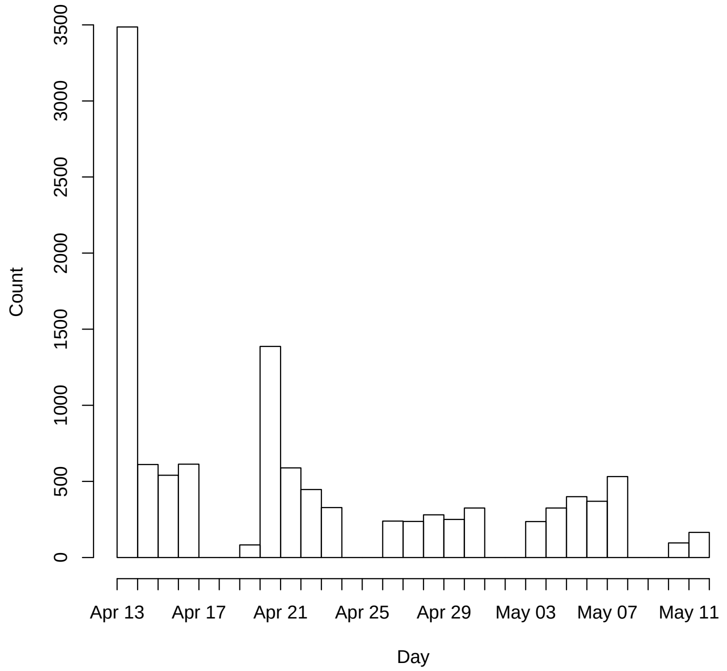


Figure 11: Number of students completing Reading in 1 day, by date.

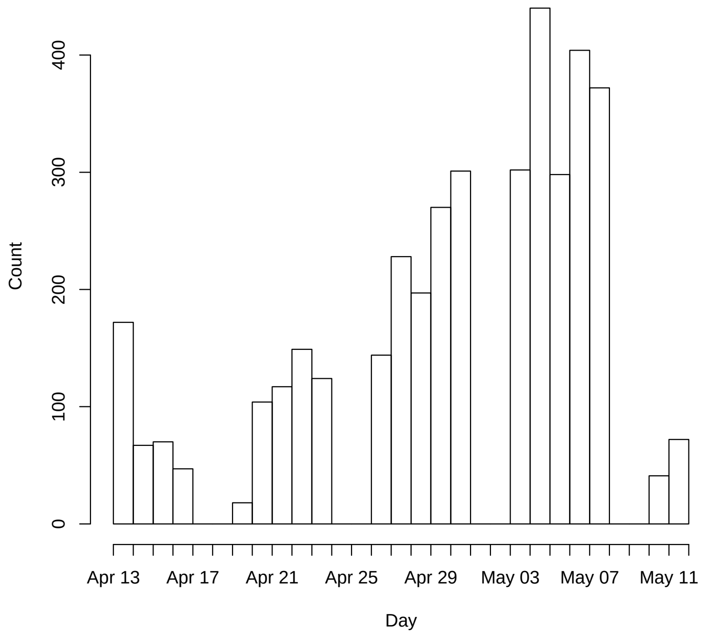


Figure 12: Number of students completing Mathematics in 1 day, by date.

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.

Appendix A: Glossary

Administration Accommodation—alterations to the administration procedures for students with disabilities or other limitations when such disabilities or limitations unfairly influence test performance. An example of an administration accommodation would be providing large print test materials for visually impaired test-takers.

AIR—American Institutes for Research, the primary testing vendor for the Florida Standards Assessments

Alignment—degree of overlap between (a) the knowledge, skills, and expertise measured by a test (as indicated by the test items), and (b) the knowledge and skills included within the test *content specifications.* Alignment can also refer to the degree of consistency between more than one set of content specifications or more than one assessment.

Alpine—Alpine Testing Solutions, a company selected by the legislatively created review panel, along with their partner edCount, LLC, to complete an independent verification of the psychometric validity of the Florida Standards Assessments

Bias—see *Item Bias*

Blueprint—an outline or framework of the specific knowledge or ability domains which will be assessed by the test and the number and types of items that will represent each test domain

Calibration The process of estimating item statistics, or parameters, that describe the characteristics of test items.

CBT—computer-based testing, the mode to administer some of the FSAs

Constructed Response Item—a test question which requires students to create (write) a response, versus selecting a response from among multiple alternatives.

DOK—Depth of Knowledge, a measure of the cognitive demand commonly applied to items

DIF—See *Differential Item Functioning*

Differential Item Functioning (DIF)—a difference in estimated difficulty of an item between two groups after controlling for any differences between the groups in subject-matter knowledge.

Distractor Analysis—consideration of the performance of the wrong options in multiple-choice items

DRC—Data Recognition Corporation, a testing vendor involved in the development and administration of the Florida Standards Assessments

edCount—edCount, LLC, a company selected by the legislatively created review panel, along with their partner Alpine Testing Solutions, to complete an independent verification of the psychometric validity of the Florida Standards Assessments

ELL—English language learner (see *Limited English Proficiency*)**Equating—**the practice of relating test scores from two or more test forms that are built to the same content to make the test scores comparable. A popular equating design utilizes information gathered from a set of common items (also referred to as anchor items or an anchor test) that are administered to all students in order to establish linkage between test scores.

Field Testing—part of the test construction process whereby the assessment is administered to a sample of examinees, prior to the operational administration, to assess the psychometric quality of test items. The results of field tests are used to develop the final test form.

FLDOE—Florida Department of Education

FL Standards—The content standards that the Florida Standards Assessments are intended to measure.

FSA—Florida Standards Assessments, the statewide student tests used beginning in 2014-15

HumRRO—Human Resources Research Organization, the vendor who provides an independent audit of the Florida Standards Assessment

IEP—Individualized education program—these programs are created for students with disabilities and are reviewed to determine if a student qualifies for an *accommodation.*

Inter-rater Agreement Reliability—the consistency (agreement) of scores or ratings given by two or more raters for the same set of responses.

IRT—See *Item Response Theory*

Item—a question included on the assessment which may be designed to collect demographic information (see *Background Variables*) or assess the knowledge, skills, or abilities of examinees.

Item Bias—item or test bias occurs when one group is unfairly disadvantaged based on a background or environmental characteristic that is unique to their group.

Item Difficulty—A statistic used to measure how difficult an item is for students to answer correctly. The value used most frequently for this statistic is *p value*, which represents the proportion of students who answered the item correctly. The p value can range from 0 (no students getting the item correct) to 1 (all students getting the item correct).

Item Discrimination—A statistic used to measure how well an item distinguishes between a high performing and a low performing student. It is calculated by comparing students' performance on each item to their performance on the exam as a whole. One way to calculate the item discrimination statistic is using a biserial or a polyserial statistic.

Item Pool—the group of test questions created for a testing program from which a test publisher/administrator will create a test form.

Item Response Theory (IRT)—a measurement model that mathematically defines the relationships between observed item responses (that examinees provide when taking a test) and one or multiple latent (i.e., not directly observable) traits (e.g., mathematics ability, U.S. history knowledge).

LEP—limited English proficiency (also known as English language learners [ELL])

Linking—the practice of relating scores from two different tests. *Equating* is a more stringent type of Linking.

Operational Scoring—scoring of actual examinee item responses using scoring procedures determined during the test development process.

Parameter Estimate—a statistical quantity which is derived from a sample and is used to make an inference about a population.

PARCC—Partnership for Assessment of Readiness for College and Careers, a group of states, of which Florida was a member, that are working together to develop student assessments

Performance Levels/Standards—also referred to as achievement levels, these represent the expected performance of examinees on a measure to be classified within specific achievement levels.

PP—Paper-and-pencil testing, the mode used to administered some of FSAs

Psychometrics—the theory and techniques of educational and psychological testing. Psychometrics involves construction of appropriate assessments with the goal of providing valid and fair test score interpretations.

Reliability—the consistency of measurement. In educational assessment, reliability typically refers to internal consistency (consistency of items within an assessment) or test-retest reliability (consistency of test scores across repeated measurements). See also *Inter-Rater Agreement Reliability.*

Sample/Sampling—A sample is a subset of the target population (e.g., districts, schools, students). Sampling is the process of selecting members of the population to be included in a sample.

SAGE—Student Assessment of Growth and Excellence, the Utah assessments for which the FSA items were originally developed

Scale Score—A value representing an estimate of an examinee’s ability on some type of reporting scale.

Scaling—the process of converting raw scores into equivalent values on an established reporting scale.

Scoring Rubrics—guidelines used to evaluate student responses to a constructed-response item by specifying criteria for scoring that distinguish between possible score points (e.g., a one-point response versus a two-point response)

SEM—see *Standard Error of Measurement*

Standard Deviation—a statistical value that describes the variance or dispersion of data points around a group average. Higher values indicate more variance in a dataset.

Standard Error of Measurement (SEM)—the degree of error associated with observed test scores. SEM is inversely related to test score reliability.

Standard-Setting—the process used to establish cut score for an assessment. A cut score is chosen to distinguish between adjacent achievement levels. Methods of standard setting include, but are not limited to, the Angoff, Bookmark, and Contrasting Groups methods.

SWD—students with disabilities

TAC—Technical Advisory Committee, a group of testing experts and stakeholders who provide consultation and input for a testing program. For the FSA, the TAC includes the following members:

Richard G. Baum
Betsy Becker
Allan Cohen
Melissa Fincher
Claudia Flowers
Richard Itzen
Peggy Jones
Akihito Kamata
Mark Reckase
Charlene Rivera
Craig Wells
Sam Whitten

TDC—Test Development Center, an organization affiliated with the FLDOE that provides assistance and content expertise in the development of the Florida Standards Assessments

Technology Enhanced Item—computer-delivered item that includes specialized interactions, beyond those that are typical with multiple choice or constructed response items, to collect response data

Test Segment—See *Test Session*

Test Session—a group of FSA items, one or more of which make up the FSAs, intended to be administered together

Test Specifications—See *Content Specifications*

Test Standards—*Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014)

Validity—the degree to which a test is measuring what it is intended to measure. Validity evidence can be gathered through appropriate processes or through research studies, and supports the meaningfulness of the test scores for the intended purpose(s) of the test.

Vertical Scale—score scale that includes multiple tests that differ in difficulty but are intended to measure similar constructs.

Appendix B: Sources of Evidence

Over the course of the audit, the evaluation team received nearly 700 individual files that documented the Utah SAGE development process, the FSA development process, the FSA administration process, the scoring and data analyses process, feedback, and data files. The List of Evidence provided in this appendix is a summary of the major topics and grade levels (as appropriate) of the documentation received from FLDOE and its partner vendors. It should be noted that some items within this list include several individual documents.

List of Evidence	
1	AIR Production Load Test Report Attachment 2-- December 2014
2	Blueline Proofs - ELA (Grades 3-10)
3	Data: 2014 CBT Tests Completed In One Day
4	2014 Fall Bias & Sensitivity Review Meeting Comments
5	2014 TAC II Meeting Agenda and Meeting Minutes PEARSON
6	2014 TAC II Meeting Agenda and Meeting Minutes AIR
7	2014-15 Test Administration and Security Agreement
8	2014-15 Test Administrator Prohibited Activities Agreement
9	2014-15 Assessment Accommodations FAQ
10	2015 FSA Reading and Mathematics Test Construction Specs v0.2 DRAFT
11	2015 HumRRo QC Results Summary (Math, EOC, ELA)
12	2015 Rubrics
13	2015 M-DCPS Survey of Test Chairpersons
14	2015 District Rank By Size
15	AIR Automated Scoring Engine FAQ
16	AIR Accessibility Statement v3
17	AIRs White Paper: Recommendations for ELA Scaled Scores
18	2015 Item specifications (Algebra 1, Algebra 2)
19	Alignment Study Participant Information
20	Analysis Plan
21	FSA Assessments Portal Google Analytics Report
22	Utah Annual Technical Report Volume 1, Appendix A - Field Test Items: Classical Item Statistics
23	Utah Annual Technical Report Volume 1, Appendix B - Field Test Items: Item Parameters
24	Utah Annual Technical Report Volume 1, Appendix C - Field Test Items: Differential Item Functioning Classifications
25	Utah Annual Technical Report Volume 1, Appendix D - Percentage of Students in Performance Levels for Overall and by Subgroup
26	Utah Annual Technical Report Volume 1, Appendix E - Student Accommodations, Test Settings and Special Codes
27	Utah Annual Technical Report Volume 1, Appendix F - DIF Flag Results

List of Evidence	
28	2014 Florida Statutes s.1008.22
29	SY2013-14 Online Reporting - Attachment 1 Student Data
30	2014-15 Statewide, Standardized Testing Time and Testing Windows PPT
31	AIR TDS Architecture
32	Attribute Mapping - Utah to Florida
33	AIR Audit Trail Column and Value Descriptions
34	FSA Validity Study: Description of the Blueprint Process
35	Bookmaps (Grade 3-8 ELA and Math; Grade 9-10 ELA; and Algebra 1, Algebra 2 and Geometry EOC)
36	ELA Score Flag and Report Status - 04.24.2015
37	Calculator Policy and Supporting Documents
38	Calibration Report - ELA Grade 8
39	Calibration specifications
40	02132015 Letter from Faulk
41	02182015 Letter to Stewart from Runcie
42	2014-15 Certification Process Diagram and Memo
43	Common FSA System Message IDs and Descriptions
44	Configuration Files
45	Patent - Constructed Response Scoring
46	Content Committee and Bias and Sensitivity Report for Utah SAGE
47	FSA Contingency Plan
48	Recruiting Contact Information (Math, ELA)
49	CPALMS Content Complexity Florida Standards
50	Cumulative District ELA Writing Completion Rates (March 2-4, 2015)
51	Q&A Session with District Assessment Coordinators (Aug 27-28, 2014)
52	Best practices to mitigate DDoS attacks (Network World Article)
53	DDoS I-net Trends, Security, Analysis & Data Report (Arbor ATLAS Initiative)
54	DDoS Appendix B: Distributed Denial of Service Attack Log (AIR)
55	Patent - Steptoe & Johnson LLP Patent Application & Correspondence
56	Investigating the Effects of Dictionary Access on Item Performance (Case Study)
57	District Assessment Coordinators Contact Information 2014-15
58	Brief Overview of Problems Reported (April 13-16) - from C. Sozio
59	Domain Subscore Information (Email from FLDOE to Alpine)
60	Florida Scoring Engine Specifications
61	ELA Blueprints Grades 3-11
62	ELA Summary Table (Item Descriptions & IDs, Standards, Keys & Rationale)

List of Evidence	
63	ELA Text-Based Writing Development Process Summary
64	ELA Writing Test Issues & Concerns - Letter from FATA President
65	ELA Writing Test Timing Decision & Summary
66	FSA Item Specifications - ELA (Grades 3-10)
67	Test Design Summary & Blueprint - ELA (Grades 3-11)
68	Email Exchanges related to Calibrations on Mock Data
69	Email Exchanges related to Equipercentile Linking on Mock Data (Dryrun)
70	Equation Response Editor Tool & Item Tutorial (Practice Site)
71	Math Content Rubric & Standards (Expressions and Equations, 6-8)
72	Field Testing Study (Standards Review, Method and Sources)
73	Data: Students Active in Both Sessions of Reading on the Same Day (By School & District)
74	Students Active in a Single Session on Multiple Days (By School & District)
75	Data: Students Who Completed Math EOC in a Single Day
76	AIR DRC Contact List (05272015)
77	Student Response History
78	Server Data with Error Logs
79	AIR Systems Presentation for District Assessment Coordinator Meeting
80	Extracted Invalidations with Student Data
81	Extracted Invalidation Codes Meta Data
82	Data: Tests Completed in Appropriate Number of Sessions or Less
83	FSA Online Reporting System (ORS) Screenshots (04102015)
84	AIR's Testing Quality Control Process
85	Data: Testing Completion Rates (By Date, Test Name, District & School)
86	Scoring & Reporting Specifications 2014-2015, V1.0
87	Data: Tests Completed in Appropriate Time Limit Conditions
88	UAT Log File Showing Errors & Other User Reported Issues
89	FSA Test Administrator User Guide 2014-2015
90	AIR Secure Browser Installation Manual 2014-2015
91	AIR Technical Specifications Manual for Technical Coordinators 2014-2015
92	AIR Organization Chart
93	FSA Score Status Flag Rules v7
94	Florida's Transition to Computer-Based Testing
95	Florida Writing Scorer Remediation
96	2015 M-DCPS Survey of Test Chairpersons Summary
97	FSA & FCAT 2015 Test Chairpersons Survey Open Ended Responses

List of Evidence	
98	FSA Assessments Approval Log
99	FSA Calibration Decisions Overview
100	FSA Computer-Based Testing Issues Guide
101	FSA ELA Writing Component Make-Up Windows Email
102	FSA Restoring Saved Student Responses FAQs
103	FSA Lost Progress Master File
104	FSA ELA, Mathematics and EOC Quick Guide Spring 2015
105	FSA Scoring Engine Specifications
106	FSA Session Re-open Scenarios & Help Tips
107	FSA Time Limits
108	FSA Log Summaries
109	FSA FCAT 2.0 Resource Distribution Timelines
110	FSA Item Review Log File
111	FSA ELA Student Report Mockups Latest Modifications
112	FSA System Requirements for Online Testing 2014-2015
113	FSA Calculator and Reference Sheet Policy
114	FSA ELA Reading Instructions for Oral Presentation Accommodations
115	FSA Mathematics Reference Sheets Packet 1
116	FSA Paper-Based Materials Return Instructions
117	FSA Mathematics Functions & Content Rubric 8
118	Behind the Scenes M-DCPS Safeguards Against Cyber Attacks Email with Meeting Information
119	032015 Update Emails from Stewart
120	FLDOE Description of Special School Types
121	Vertical Linking (Math, Grades 3, 4, 5)
122	2015 Form Builder Notes (Core and Anchor)- Grades 4, Geometry, Algebra 1
123	Math Item Card with Stats (Grades 7-8)
124	Fit Plot Graphing
125	ELA Accommodated Form (Grades 6-9, 11)
126	FSA Mathematics Functions & Content Standards (Geometry)
127	Blueline Proofs - Math (Grades 3-8, Algebra 1, Algebra 2, Geometry)
128	FSA Test Item Specifications (Geometry EOC)
129	Good Cause Exemptions for Grade 3 Promotion Email
130	Reading Core Form (Grades 6-9, 11)
131	FSA Test Item Specifications (ELA Grades 3-10)
132	FSA Test Item Specifications (Math Grades 3-8)

List of Evidence	
133	Presentation: Tips for Taking FSA ELA & Math Assessments (Grades 3-4, Paper-Based Tests)
134	Graduation Requirements for FL Statewide Assessments 2015
135	FSA ELA Writing Handscoring Specifications Spring & Fall 2015
136	FSA Help Desk Reports (03.16.2015 - 05.18.2015)
137	Monthly Emails from FLDOE to DAC
138	2015 Spring FSA Superintendent Certifications (30 school district records)
139	2015 Spring Irregularities (17 school district records)
140	AIR Report: Impact of Test Administration on FSA Test Scores
141	Utah ELA Informational Standards (Grades 3-12)
142	FSA ELA Text-Based Writing Rubrics - Grades 6-11 (Informative, Explanatory)
143	FSA ELA Text-Based Writing Rubrics - Grades 4-5 (Informative, Explanatory)
144	FSA Infrastructure Readiness Guide
145	Email Invitation to TAC Members to a WebEx Call for Reviewing Vertical Scaling Results
146	NCIEA Analysis of the Impact of Interruptions on the 2013 Admin. of the Indiana STEP-Plus Testing Program
147	FSA Item & Form Selection Process 2015 Operational Tests (Grades 3-10 ELA; Grades 3-8 Math, Algebra 1-2, & Geometry EOCs)
148	AIR Item Layouts & Answer Variations Guide
149	Utah State Office of Education SAGE Item Writing Process
150	FSA Meeting Agenda, AIR Offices, Washington DC
151	FSA Meeting Notes, AIR Offices, Washington DC
152	FLDOE Office of Assessment & K-12 Student Assessment Staff Lists
153	FSA Bias & Sensitivity Training and Activity Materials
154	Linking Reports (Algebra 1, ELA Grade 10)
155	Utah ELA Literary Standards (Grades 3-12)
156	Math Online Forms Review (Grades 5, 7)
157	Horizontal Linking Math (Algebra 1)
158	Math Summary Table (Item Descriptions & IDs, Standards, Keys & Rationale)
159	Vertical Linking - Math PPT (Grades 3-8)
160	Data: Math completion by session
161	Test Design Summary & Blueprint - Mathematics (Grades 3-8, Algebra 1 EOC, Algebra 2 EOC, Geometry EOC)
162	Math Content Rubric & Standards (Measurement and Data, 3-5)
163	FSA Statewide Assessments Production Specifications for Binder 2014-15
164	Data: Number of Students Who Took the Writing Assessment in the 2nd & 3rd Window
165	Math Content Rubric & Standards (Number System, 6-8)
166	Math Content Rubric & Standards (Numbers and Operations, Fractions, 3-5)

List of Evidence	
167	Math Content Rubric & Standards (Numbers and Operations in Base Ten, 3-5)
168	Operational Master Data Sheets
169	Math Content Rubric & Standards (Operations and Algebraic Thinking, 3-5)
170	ELA Text-based Writing Rubrics (Grades 4-5, Grades 6-11)
171	FSA Packaging & Distribution Specifications (ELA, Writing, Mathematics, Algebra 1, Algebra 2 and Geometry)
172	Letter to Pam Stewart, Commissioner of Education FLDOE from John Ruis, President FADSS (02102015)
173	Utah State Office of Education SAGE Parent Review Committee
174	Summary of Pasco Schools CBT Writing Test Issues
175	Statement from Pearson Regarding Service Interruptions & DDoS Attack on 21 April 2015
176	Math Content Rubric & Standards (Ratios and Proportional Relationships 6-7)
177	Spring 2015 Testing Issues due to Server Interruptions Documentation
178	Data: Reading completion by session
179	Utah Released Scoring Rubrics - Writing (Grades 3-11)
180	FSA Script for Administering the CBT Math, Grades 6-8, Sessions 2 & 3 - Spring 2015
181	FSA 2015 CBT Comment Form Reports (Reading, Math & EOC)
182	FSA 2015 PBT Comment Form Reports (Reading, Math & EOC)
183	Rule 6A-1.09422- Florida Comprehensive Assessment Test and End-of-Course Assessment Requirements
184	Rule 6A-1.094223 Comparative and Concordant Scores for the Statewide Assessment Program
185	Rule 6A-1.0943- Statewide Assessment for Students with Disabilities
186	Rule 6A-1.09432 Assessment for English Language Learners
187	SAGE Item Development Process
188	FLDOE Early Processing Sample Design
189	Secure Browser and TA Interface Demonstration Webinar PPT
190	Mathematics Standards Coverage (Grades 3 - 8, Algebra 1 and 2 EOC, Geometry EOC)
191	AIR Report: Students that completed testing "as expected"
192	02062015 Letter from Stewart to Gaetz
193	Smarter Balanced Assessment Consortium: 2014 Student Interaction Study - Design and Implementation Plan
194	Smarter Balanced Assessment Consortium: Cognitive Laboratories Technical Report
195	Spring 2015 Vertical Linking (Grades 3-8)
196	Spring 2015 ELA OP OO Rubric Items
197	2015 Test Administration Manual
198	Spring 2015 Math OP OO Rubric Items
199	Spring 2015 FSA Training Materials PPT

List of Evidence	
200	Spring 2015 Tips for Taking the CBT FSA ELA Reading Assessments (Grades 5-10) PPT
201	Spring 2015 Tips for Taking the CBT FSA Mathematics Assessments (Grades 5-8) PPT
202	Spring 2015 Directions for Completing ELA Reading Items (Grades 3-4)
203	Spring 2015 Directions for Completing Mathematics Items (Grades 3-4)
204	Spring 2015 Test Administrator Checklist for CBT (ELA Writing & Reading, Mathematics and EOCs)
205	Spring 2015 Braille Scripts (ELA Reading - Grades 3-10, Mathematics - Grades 3-8)
206	Spring 2015 Scripts and Instructions for Administering Accommodated CBT (ELA Reading - Grades 5-10, Mathematics - Grades 5-8, FSA EOC Assessments)
207	Spring 2015 Scripts and Instructions for Administering PBT (ELA Reading - Grades 5-10, Mathematics - Grades 5-8, FSA EOC Assessments)
208	FLDOE Staff Contact List
209	Data: Testing Completion Rates (State)
210	Math Statistical Summary (Grade 4, 6)
211	Math Content Rubric & Standards (Statistics and Probability, 6-8)
212	Data: Student Timeout Summary 031715
213	Data: How many students were active in both Reading sessions in one day?
214	Data: How many students were in a single session on multiple days (all exams)?
215	Data: How many students completed both math sessions (for those grade levels that had 2 sessions) or all 3 math sessions (for those grade levels that had 3 sessions) in one day?
216	Data: How many students completed both Reading sessions in one day?
217	Calibration - Summary of Daily Call Process
218	Florida Writing Supervisor and Scorer Numbers
219	Linking 2011 FCAT 2.0 Scores to the FCAT Vertical Scale: Legal and Policy Perspectives
220	SY2013-14 Utah Technical Report Volume 1 - Annual Technical Report
221	SY2013-14 Utah Technical Report Volume 3 - Test Administration
222	SY2013-14 Utah Technical Report Volume 4 - Reliability and Validity
223	SY2013-14 Utah Technical Report Volume 5 - Score Interpretation Guide
224	SY2013-14 Utah Technical Report Volume 6 - Standard Setting
225	SY2013-14 Utah Technical Report Volume 2 - Test Development
226	Alignment Study Recruitment List
227	Test Development Center Organization Chart
228	2015 HumRRO Quality Assurance Proposal
229	Test Administration Policy Email
230	AIR Test Development Staff Resumes
231	ELA Item Stats
232	Data: Tests that had a Segment Reopened

List of Evidence	
233	TIDE Online Training Module
234	TIDE User Guide 2014 – 2015
235	Utah Technical Report Volume IV Appendix A - Marginal Reliability Coefficients for Overall and by Subgroup
236	Utah Technical Report Volume IV Appendix B - SEM Curves by Subgroup
237	Utah Language Tasks Item Guidelines
238	Utah Listening Item Specifications
239	Utah SAGE Listening Guidelines - 2014
240	Utah SAGE Writing Task Guidelines
241	Utah Validity Summary
242	Utah Writing, Language Editing, and Listening Task Specifications
243	Math Item Card with Stats (Grades 3-6)
244	Verification of Computation of Raw Scores, Theta Scores, and Scale Scores (Math Grades 3-8, Algebra 1 and 2, Geometry)
245	Vertical Linking Design
246	Vertical Linking Master Data Files
247	Calibration Meetings - Weekly Action Log 03182015
248	FSA Range ALD Workshop Memo 07102015
249	Item Writer Training Materials
250	Writing Item Specifications
251	2015 Writing Operational - DAC, SAC, PBT TA, and CBT TA Comment Forms
252	2015 Writing Response Help Desk Cases

Appendix C: District Survey Results

As part of the evaluation of the FSA test administrations, the evaluation team sought input from district representatives about their experiences. To collect this information, the evaluation team created an online survey that included questions related to preparation prior to the administrations, support during the administrations, and the administrations for each of the three main FSA content areas: Writing, Reading, and Mathematics. Using a list of district assessment coordinators and contact information provided by FLDOE, the evaluation team distributed the survey via email on July 1, 2015 to representatives from all 76 Florida districts. The survey remained open through July 20, 2015, and two reminder emails were sent on July 8 and 13.

A total of 58 survey responses were received. Three responses were removed for incompleteness (no responses beyond survey question #5) leaving a total of 55 responses from the following 48 districts.

Baker	Highlands	Okeechobee
Bay	Hillsborough	Orange
Bradford	Holmes	Palm Beach
Broward	Jefferson	Pasco
Calhoun	Lafayette	Pinellas
Citrus	Lake	Polk
Collier	Lee	Putnam
Desoto	Leon	Santa Rosa
Dixie	Levy	Sarasota
Escambia	Liberty	Seminole
FL Virtual	Madison	St. Lucie
FSDB	Manatee	Sumter
Gadsden	Marion	Suwannee
Gilchrist	Martin	UF Lab School
Hamilton	Miami-Dade	Volusia
Hernando	Okaloosa	Washington

The following sections include each individual survey question along with the responses received. Where applicable, open-ended comments are also included.[2][3]

[2] Respondent comments were copied directly from the online survey results without correcting for errors in spelling or grammar.

[3] To protect confidentiality, names of individuals were removed.

Survey Instructions

On behalf of Alpine Testing Solutions and edCount, LLC, thank you for taking the time to complete this survey.

The purpose of this survey is to evaluate the test administration process for the Florida Standards Assessments (FSA) program, administration data, and administration successes and degree of interruptions across all test centers. This survey should take approximately 15-20 minutes to complete.

System Preparation (SP)

SP1

Prior to the Florida Standards Assessment (FSA) test administration, which of the following did the schools in your district engage in to prepare for the test administration? (Check all that apply)	
Test administration manuals were sent to all schools and school testing coordinators (individuals responsible for testing activities at each school) were required to review the user manuals.	96.36% (53)
All school testing coordinators were trained on the administration protocols with this individual responsible for training any other testing administrators at their school.	98.18% (54)
School testing coordinators conducted training with all individuals at the schools that were scheduled to serve as testing proctors.	98.18% (54)
The technology requirements for the FSA were reviewed at the school level to ensure that the school could support the test administration.	98.18% (54)
Prior to the administration, school testing coordinators engaged with the system and its functionality.	96.36% (53)
None of the above	0.00% (0)

SP2

Please review the following statements regarding your district's computer system preparation for the FSA administrations and indicate your level of agreement with each. Please indicate your level of agreement with the statement with 1 indicating strongly disagree and 5 is strongly agree.

	Strongly disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly agree 5
My district was adequately prepared to administer the Florida Standards Assessments on computer.	0.00% (0)	10.91% (6)	18.18% (10)	30.91% (17)	40.00% (22)
My district was given sufficient information to prepare our systems for the computer-based test administrations.	1.82% (1)	21.82% (12)	20.00% (11)	40.00% (22)	16.36% (9)
My district was given sufficient time to prepare our systems for the computer-based test administrations.	3.64% (2)	14.55% (8)	21.82% (12)	40.00% (22)	20.00% (11)
My district had adequate resources to prepare for the computer-based test administrations.	3.64% (2)	23.64% (13)	14.55% (8)	36.36% (20)	21.82% (12)

SP3

If you rated any of the questions above with either a 1 or 2, please provide additional information about the challenges you encountered in the box below. If you answered 4 or 5 on any of the questions above, please provide additional information on any instrumental components of your district's preparation that you considered to be vital to your preparation.
While we received information regularly, many times we received information just before implementation. In addition the timelines for implementation with information were compacted because we were getting information.
We provided 8 mini trainings on how to administer and extra staff to make sure all school tech cons were ready.
I think that some of the "last minuteness" of information/system changes were really difficult
A major difficulty was the lack of timeliness for information regarding the administration. This includes late information on technology and administration. A primary issue is still the lack of an actual training site for test administrators that reflects the actual testing administration, with separate sessions. But, the main issue is still that we were essentially flying blind, not knowing what the screens would look like on testing day. With trainings not held until February, we were really in a bind to prepare our training materials and train our personnel in time for the tests that began in only a couple of weeks after that.   Our ITS team worked diligently to prepare as they received information, and I believe they received that information on time, but I am not sure.   Our district had adequate resources to prepare for the CBT administration as it was scheduled, but NOT resources to pull kids 4, 5, 6 times in for what was supposed to be only 1 session, due to FSA/AIR major malfunctions.
We have been preparing and refining our approach to CBT for several years and have a strong inter-District collaboration (5) We were not given accurate information on how peripherals would interact with our systems; we were given inaccurate information about saving routines and time-out routines; We were given inadequate/wrong information so that test to speech was not operable (2) We had sufficient time to prepare because we had significant preparation ahead of time; however, late notification on some of the above items hampered us slightly in implementation (4) We don't have enough computers to prevent testing from having a negative impact on instruction, but we are able to effectively schedule to meet all testing requirements (4)
Test administration changes were happening too close to the test administration window. Text to speech was suspended on 3/27 for the 4/13 assessment. Supplemental scripts were sent out late. There was not sufficient time to let school based test administrators practice in TIDE. TDS was taken down from 4/1 to 4/5 so teachers couldn't train in the system.   Better collaboration and communication between our IT staff and the vendor or state's IT staff would help identify issues when the system is down.
There were steps involved with the administration that were not in the manual (opening segments, etc.) or portal. In addition, we did not have grade level practice tests, only grade bands. The manual itself changed with more options leading up to the administration of the assessment (changes to scripts). The information on "readers" and what they could read created great confusion in the district when administering the CBT to a student with this accommodation even with the examples provided. There was not enough information released ahead of time regarding the availability of a writing passage booklets. We had prior experience with reading passage booklets, but didn't know about the option of the writing passage booklet until it was time to begin the administration (we were not part of the writing pilot).
Question 1 - Agree - My district went to great lengths to prepare testing profiles, set up hardware, test the infrastructure, and train school administrators, teachers, and students based on resources

If you rated any of the questions above with either a 1 or 2, please provide additional information about the challenges you encountered in the box below. If you answered 4 or 5 on any of the questions above, please provide additional information on any instrumental components of your district's preparation that you considered to be vital to your preparation.
provided by DOE and AIR. Question 2 - Strongly Disagree - The districts were not provide enough information to account for the intricacies of using the secure browser and java settings for Mac OS. Important power management and network settings were not communicated which resulted in frequent student and TA kick off. Question 3 - Disagree - Components of the cbt testing platform including Text to Speech were not communicated as to allow for sufficient prep and training. Question 4 - Disagree - While resources were plentiful, they were difficult to find on the portal and spread across more than 7 manuals. Aspects necessary to provide to schools so that successful administration could occur like scripts, etc were not readily available nor were trouble shooting documents to help for when things went wrong. As a large MAC district, we were forced to trouble shoot on our own with little to no AIR support and fix settings so that the platform was stable on MAC OS's.
The infrastructure trial was important and knowing the exact technical specifications that were needed for computer set up.
2 was marked because although the information was available it came out piece by piece and it made it difficult to keep up with the changes.
The Superintendent of Seminole County Public Schools had provided a written statement to Commissioner Stewart during the web certification process that provided a detailed account of concerns our district had prior to the first administration of FSA. Details included: * It is important to recognize that compared to Spring 2014, an additional 41,048 SCPS students are now required to take a CBT. This requires SCPS to schedule an additional 65,388 CBT test sessions. * There is a lack of time and computers for students to adequately practice using online tools.
The district assessment office was well trained and provided all necessary information to schools and departments. The challenge was the large number of assessments and the large testing windows.
Even though our technology teams followed the tech specs when preparing the computers for testing, we still ran into issues during the test administration regarding tech concerns (i.e., some students were kicked out of testing with an error message saying a program was running in the background, even though computers were correctly set up prior to testing following the directions from AIR).
Computer Specs were available early enough to adequately prepare for the administration. The CBT Certification process FLDOE has in place is an excellent tool to help ensure districts are prepared for CBT as each school must assess their readiness for each test administration.
TA training of the test delivery system was a vital concern. The schools participating in the field testing had a advantage from that experience the other schools did not have.
We had to replace ALL of the computers that had been used in our testing labs the past several years with ones that met the system requirements for the FSA. This was a sizable expense for our district, and we are concerned about the lack of recognition for additional funding needed to support the testing environment.
Our Information Technology Support division is in the same reporting structure as the district assessment office, which helps to ensure that we get the needed support.
I feel given the information we had at the time we were as prepared as we could have been. The amount of information we 'discovered' furing the assessment was disconcerting at best and in some cases hampered our administration. Even things as simple as nomenclature, we found out during the administration that the contractor was calling a session, a segment which caused much confusion out in the schools. the arrival of information was in some cases 'at the last minute' and really made it hard to ensure that all schools were up to speed with the latest information. In districts the size of

If you rated any of the questions above with either a 1 or 2, please provide additional information about the challenges you encountered in the box below. If you answered 4 or 5 on any of the questions above, please provide additional information on any instrumental components of your district's preparation that you considered to be vital to your preparation.
those in Florida, our only method of communication is email. You can send it out but you can't make them read it or know that they really understand it.
The information regarding administration was fine, the information regarding technical issues and potential "glitched" was not in place prior to testing. Once information was available it frequently changed throughout the course of the administration.
Orange County Public Schools made large improvements in student/computer ratios and increasing available bandwidth in the 12 months before the administration of the Florida Standards Assessments. Before these improvements, secondary schools in OCPS had student to computer ratios ranging from 1:1 to over 12:1. All middle schools were brought to no less than a 3:1 ratio and all high schools were brought to no less than a 4:1 ratio. Elementary schools were brought to around a 5:1 ratio depending on size. This provided flexibility for schools, ensured lower amounts of instructional disruption and continued to move the entire district toward digital curriculum goals. We also increased bandwidth by 33% to ensure consistent access and no interruption with other existing digital needs. The rapid transition between PARCC and the Florida Standards Assessments gave limited time for FLDOE and AIR to provide the resources that we needed on a schedule similar to prior years. With over 200 schools and sites, we need a reasonable amount of lead time with training materials and other related resources in order to train and prepare our systems. We felt that the FLDOE did well given their constraints, though we would not say that we received sufficient information or time to prepare systems and train.
Our district is quite small, and the assessment coordinator works closely with the Information Technology department on all computer-based assessments.
My district was given sufficient information to prepare our systems for the computer-based test administrations: Information was being sent out quickly via email but the FSA portal was never up to date. This caused confusion. My district had adequate resources to prepare for the computer-based test administrations: The directions for administering the paper-based test for students with accommodations were not provided in a timely manner. We were notified less than two weeks before testing that Text to Speech was not going to be available. My district had adequate resources to prepare for the computer-based test administrations. We did not have enough computers for testing nor did we have enough district technical report to manage ALL of the issues encountered. The helpdesk was not able to provide assistance in a timely manner and some issues were never solved.
Being in a poor rural county, we have great issues with bandwidth, connectivity and the ability to have enough computers to test students in a timely manner.
Resources are a loose term. School instruction was totally disrupted, especially at the high school level. Students at the high school level may take grade band exams, but they are scattered across multiple sections and not all students in those classes they are pulled from take the same exam due to their grade level. This is nothing new, the state is aware. Computer labs are disrupted of their normal educational instructional time in order to allow students access to the resources for online exams. At the end of the testing period, almost 6 weeks of instructional time in those labs are lost.
We participated in the Infrastructure Trial and utilized the practice tests with our students. This allowed us to ensure our computers would function properly and get the students and test administrators familiar with the new platform.
Our district technology team with hands-on with each of the schools to have technology support available.

If you rated any of the questions above with either a 1 or 2, please provide additional information about the challenges you encountered in the box below. If you answered 4 or 5 on any of the questions above, please provide additional information on any instrumental components of your district's preparation that you considered to be vital to your preparation.
Information seemed to change nearly daily during the school year...and even during testing. It was very difficult to follow information consistently and efficiently from one memo to the next with such constant change.
Computer-based testing is an unfunded mandate. Until we are 1:1, testing will be a burden instead of a typical part of a school year.
Resources came, but late. Often, we had already done something ourselves before a resource was available (ex. training). Also, some guidance to prepare us for any issues we might encounter before testing began would have been extremely useful. Calling FSA for help always resulted in extremely long wait times and I received questionable responses more than once. I would have to call DOE to verify before I acted, and usually found that what I was told by FSA help was incorrect. It should be noted that the DOE office was great to help and prompt to reply. My issues came with FSA help desk.
Some of the materials for administration came only several days before the assessment windows, We were told text-to-speech would be available and then were told last minute that it would no longer be available.

Overall FSA Test Administration (TA)

For the following questions, please consider your district's experience with the test system across all 2014-15 Florida Standards Assessments (FSA) administrations.

TA1

Across all tests and administrations, please estimate the degree to which the administration of tests to students was postponed or interrupted by technological challenges.	
No impact	0.00% (0)
Minor impact	29.63% (16)
Moderate impact	42.59% (23)
Major impact	27.78% (15)

TA2

Across all tests and administration dates, approximately what percentage of students in your district was impacted by technology issues related to the FSA?	
None, 0%	0.00% (0)
1-9%	18.52% (10)
10-19%	20.37% (11)
20-39%	16.67% (9)
40-59%	18.52% (10)
60-79%	14.81% (8)
80-100%	11.11% (6)

TA3

Based on your experience, do you feel that there were more technology issues during 2014-15 test administrations as compared to prior years?	
Yes	88.89% (48)
No	11.11% (6)

TA4

During the test administrations, did you reach out to the FSA Help Desk for any assistance?	
Yes	98.15% (53)
No	1.85% (1)

TA5

If yes, please rate the quality of the help desk experience.	
1 - Exceptionally poor customer service	33.33% (18)
2 - Poor customer service	40.74% (22)
3 - Neutral	22.22% (12)
4 - Good customer service	3.70% (2)
5 - Excellent customer service	0.00% (0)

TA6

Please provide further explanation for your rating of the FSA Help Desk and the assistance provided.
Could not answer the questions most of the time. Would call back the next day with no solution. They knew as much as we did, slowwwww responses.
We were given inaccurate guidance on one occasion. Also, there were delays in response and assistance. Could not reach a person on some occasions. Customer service was good at times and poor at times during heavy testing across the state.
At times customer service was easy to reach and very helpful and other times hard to reach and the solution took time.
Some of the representatives seemed unsure of how to rectify some of the issues that our district was having.
The people manning the helpdesk were more clueless than the ones above them. When we called, we were placed on hold for unacceptably long wait times. Remember, as they did not seem to realize, that we had students waiting for the resolution; Students sitting at computer screens, getting antsy. When someone finally answered, they had no idea what we were talking about. They had some manual in front of them that they were reading from, but had not actually been on the system. I had to provide my log in and password so that they could log in and see what we were talking about. The Manual would describe tabs to click - however, those tabs were not present in the actual system, or as one helpdesk person told me, "That functionality is not available yet". However, this person was instructing me to use the "functionality". Their solution was for us to just go pull the students out of class again, pull a teacher out of class again and see if the issue had been resolved. NO!! These kids and teachers and schools are not here to serve as FSA/AIR QA department. If the platform is not working, then suspend until it is. Do NOT tell us to go get these kids out one more time just to see if it's working.
Help Desk Agents did not have access to the system and could not see the problems being referenced; they could not make required adjustments (did not have the authority needed); they did not know how the system worked and we frequently had to teach them while on the phone with them; they told us it was our problem when they knew the issues were widespread and systematic; they frequently wanted to put the burden for correcting things back on the student testing.
The help desk was cordial but did not have the expertise needed to solve the issue. A student was being kicked out of the test several times and the help desk resolutions did not resolve the problem.
There was a call I made and the poor person assisting me did the best job they could, but it was obvious they were fumbling through instructing me. I had a bad feeling about the instruction I received so I called FDOE when I finished and sure enough, the steps provided were not correct.
Front line service agents were not knowledgeable of the test and often times (in the beginning) seemed like they were temps from a job service agency with no prior testing nor education background.
Wait times were ridiculous. Once you got someone on the phone, they provided absolutely no help and often required that you repeat the information already provided. They often did not call back or returned a call so far past the time of original notification that the student had gone home or the testing window had closed.
Long waits, too much information was collected just to ask a question. (name, district, etc.)
I felt that I knew more about solving the problem then Level 1 tech. Sometimes the call was just a waste of time. Level 2 tech. was more knowledgeable in recognizing the issue and solving it in a timely matter.

Please provide further explanation for your rating of the FSA Help Desk and the assistance provided.
There were some situations which they were not very helpful. Their instructions were a little confusing. But overall they were nice.
The help was decent once I received it. The wait time was a huge issue. Also, schools can't sit and wait. The expectation that schools can manage some of this is not realistic. Return calls was also another issue.
My experiences included: -- being hung up on -- being placed on holds of 20+ minutes whiles students waited at schools -- being asked to provide non-essential information while they completed their paperwork while students were waiting -- having to assist the help desk in fixing my own problems while they read the manual out load to me so I can explain to them what to do -- spending 5+ hours in one week providing the same information over and over to have student tests resumed and restored --being told multiple mornings that the testing computer issues were not from their end that it must be the district's issues
I had the same person at least 3 times and he was actually unhelpful. When I shared that the issues I was calling about were at the level 2 support; he refused to expedite my call to that level. He was rude and awful.
The help desk personnel were not familiar with the Florida assessments and not listening to callers. VERY poor response time - Help desk not answering the phone....Prompting people to call back later when you are in the middle of testing and have stressed students sitting in front of a computer waiting is every districts' worst nightmare. . Conflicting advice from help desk personnel about the steps to correct issues.
- Some incidents received case numbers, some did not; one incident may have multiple case numbers--- this was very confusing - Help desk attendees were unfamiliar with the platform (i.e., I was transferred multiple times to different tiers and/or peopl
I mainly worked with FDOE staff to help resolve my issues. I did report a few missing writing tests and called to follow up with the help desk. I did not receive any helpful information or follow up regarding the call. Since my interaction was limited, I cannot definitively describe the service. The schools who did make calls to the help desk reported extremely high wait times, so much so, some hung up or refused to call back when they had issues. In those cases, I reached out to FDOE staff to help resolve the issues.
Although I do not feel it was the fault of anyone at the FSA Help Desk, as they, too, were learning a new system and were faced with overcoming challenges beyond their control, they inability of customer service representatives to resolve problems was troubling. Additionally, the hold times were excessive. Districts don't have the time to hold the line for upwards of an hour while waiting on a representative.
Some representatives more helpful that others.
The Help Desk did not get back with schools/our district in a timely manner.
Customer service varied depending on who you spoke with - sometimes they could answer our questions, but other times it appeared that they had no idea what we were talking about!
In our district, we were given direct access to an individual in technical support to assist with an extremely heavy load of students whose tests needed to be reopened. This support was extremely helpful, and test administration could not have been carried out without it. Direct communication with the FSA help desk - by our office staff and schools - was extremely poor. Promises were made, but response time was extremely slow, and often resolution was not completed.
The Help Desk was unprepared for the job they were doing. In some cases I had to wait as the HD personnel were thumbing through the manuals, and in one case I had to direct the person to the

Please provide further explanation for your rating of the FSA Help Desk and the assistance provided.
correct document. They also were unprepared for the quantity of calls. Call backs were never made or if they were came days too late. Too much time was spent verifying information on the caller and not enough time on the problem.
They did not understand the importance of a call or provide any assistance. I did my own trouble shooting and if I couldn't figure out the problem I would email DOE and they would assist me.
The customer service agents were never able to give me a direct answer. I was often told that they would have to follow up with a project lead to get an answer, followed by we're still working on it when I called back for an update.
FDOE staff provided excellent support, as did some of the representatives from the FSA Help Desk. Otherwise, the support was poor at best. The primary issues were wait time and the efficacy of the support. Further, we were told by second tier tech support that they were not allowed to open tests. I am not sure how an issue that has risen to that level could be resolved without opening the test on their end.
The service of the FSA Help Desk was exceptionally poor at times, though it did improve later in the window. When there are problems with test administration (and some problems are to be expected), districts want clear information. The Help Desk would not provide clear information, would tell schools information that was different than what was told to district personnel and did not follow-up on issues in a timely fashion. During the administration of the computer-based writing assessment, this poor services caused many students to have their assessment administration extend multiple days with no understanding of what assessment information had already been collected. This happened on a scope that raises strong concerns about the validity of the assessment. We are particularly concerned that this happened with the Grade 10 ELA assessment that is a graduation requirement for students.
I rate the customer service as poor for the following reasons: Long wait times for telephone and email requests; Help desk personnel were often unable to answer questions and I had to call FLDOE for assistance; I was told several times that I must have done something incorrectly. More specifically, when students disappeared from the TIDE system, I was told that I must have deleted them, when I most definitely did not. When experiencing issues with the reporting system, a common help desk response was to wait and everything would update eventually. It rarely did, and I had to call again for assistance.
Wow.... There were people answering the phone but they were not knowledgeable about TIDE or TDS. Basically, our information was taken and then repeated back for clarification and "elevated to the next level." One major issue was the initial help ticket was assigned one number but then that number was changed without explanation or notification. The original ticket was not able to be tracked. This caused confusion when districts would send ticket numbers to schools so they could later track. Ordering paper based material was extremely frustrated. Different agents gave different deadlines for ordering and which included time zone discrepancies. This caused a delay in receiving materials which meant students with accommodations had to be rescheduled for testing. The customer service agents who were on the "front line" were mostly polite but unable to help! They did not have a basic understanding of FSA testing at all nor did they have any technical understanding. This was a waste of time.
We reported numerous issues while testing to the help desk. At times, we had representatives at the helpdesk who couldn't even begin to help us, who we couldn't understand due to dialect and had many tickets never responded to at all.

Please provide further explanation for your rating of the FSA Help Desk and the assistance provided.
FSAs help desk was very unresponsive and took, at times, two days for a response. Service was so bad, DOE finally had a separate phone line setup just for District Coordinators. This line was more user friendly, but should not have been needed.
They never know what to do; how to answer; how long it would take to fix. There were times when they blamed the problems on our systems and it was later proven not to be.
Many times callers were on hold for 45 mintues to an hour and when someone finally became available they didn't know how to assist the caller. I had an operator become frustrated and hang up on me and several of my testing coordinators reported the same thing to me. They gave misinformation to callers, which I reported to the state. One operator told my testing coordinator to suspend testing for the rest of the day, which they have no authority to do. I stopped asking my schools to call them because they were the "unhelpful" helpdesk and I didn't want to frustrate them further.
Most of the FSA Help Desk representatives were unable to solve the issue being reported. All issues had to be elevated to Level 2 status. The usual response was "we will have to get back to you." The amount of time required to provide answers to all the representatives' questions was extensive, especially given that no resolution was ever provided by the end of the phone call. Early in testing, there were long (>20 minute wait times) to reach a representative by phone. This was corrected by the end of the testing window. When emailing the FSA Help Desk, there were usually very long wait times for resolution steps; many of which did not work. One positive -- at the end of testing, the majority of operators (exception of 1) who were assisting with option #3 -- test resumes -- were efficient, friendly, and very helpful. They were usually able to get students testing again within 10 minutes.
Unreasonable wait time to contact with the help desk and answers were not readily available to the troubleshooting we were looking for.
It seemed to me that the individuals that I spoke with, didn't have the proper authority or understanding to expedite questions and issues. I will say that it seemed to get better the further into assessment that we got, but I don't think that it was adequate for the level of importance placed upon these exams.
I got redirected too many times and was not told accurate information. Ultimately, I had to involve DOE and work through the distributor.
The help desk personnel were cordial but lacked the basic knowledge and understanding of the platform. In addition, follow up to technical problems were inconsistent and not done in a timely manner.
They were not on the same page as the FLDOE. They would make statements which were inaccurate...or perhaps just made up to appease us when we called....(lost records)
Wait times were extremely long. Most representatives that I spoke with were hesitant and unsure with their responses. Some responses were questionable, so I had to call DOE to verify...sometimes to find the information I received from the Help Desk was incorrect.
We had both good and bad experiences calling in. At the beginning, it was worse. As testing progressed, it was better. DOE Assessment Office also helped facilitate the problems.
I only had to call a few times, but they could not answer my question. I would get a call from someone several days later saying they work working on it. By then, it was resolved or too late to matter.
They were very helpful and tried to resolve issues quickly.

Please provide further explanation for your rating of the FSA Help Desk and the assistance provided.
We called numerous times to the helpdesk and were sometimes on hold for up to 45 minutes. We were not able to get answers when we did finally get through to the helpdesk. We started just calling the state with issues and they were able to resolve them quickly and effectively.

Test Administration – ELA Writing (ELA-W)

Please answer the following questions regarding the **computer-based** test administration of the FSA ELA Writing within your district.

ELA-W1

Did schools in your district encounter any technology issues during the administration of these tests?	
Yes	94.34% (50)
No	5.66% (3)

ELA-W2

If yes, please answer the following: Approximately what percentage of students in your district was impacted by technology issues during the administration of these tests?	
None, 0%	5.66% (3)
1-9%	24.53% (13)
10-19%	22.64% (12)
20-39%	15.09% (8)
40-59%	7.55% (4)
60-79%	11.32% (6)
80-100%	13.21% (7)

ELA-W3

Please indicate the grade levels where difficulties with the FSA ELA Writing Test were encountered (check all that apply).	
Grade 5	24.53% (13)
Grade 6	28.30% (15)
Grade 7	28.30% (15)
Grade 8	90.57% (48)
Grade 9	81.13% (43)
Grade 10	73.58% (39)
None	7.55% (4)

ELA-W4

Please indicate the types of issues that were encountered (check all that apply).	
Challenges logging into tests	81.13% (43)
Test sessions were closed unexpectedly forcing students to log back in	86.79% (46)
The test session ran too slowly and interfered with student navigation across the test	39.62% (21)
Test system did not function as expected during testing	62.26% (33)
None of the above	5.66% (3)
Other, please describe (see next page for responses)	37.74% (20)

Other, please describe
Students were kicked out of testing during testing.
students losing content
Student work was continually lost. Many students's work was continually lost.
Use of writing tools and keyboard functions did not work as expected; student work did not save as we were told; text was often not recovered for students
White screens
Student responses not saved
Denial of Service Attacks, Loss of student work that was unrecoverable
Students being timed out with no indication because all they did was type and not interact in the system
System did not save student work. Overwritten ELA Writing tests, Performance task questions not able to take student responses
taken from email report from our middle school, We had several students who were kicked out of the test but were able to log back in. During the test, the font increased in size during the test. We had him log out and log back in, but the font was still very large. We then had to move him to another computer to log into the test. That corrected the problem. On Tuesday, March 3rd, some students had a difficult time getting into the test. When they clicked on the secure browser, the screen just went white. It was then frozen on the white screen or would time out and kick them out. On Tuesday, teachers had to wait up to 15 minutes to be able to log into the portal. This occurred while they were in the testing situation and reading the script, and they weren't able to get the Session ID during that time. It was very frustrating to the teachers and the students. During makeups on Wednesday, the FSA website was not available, and teachers and students could not log in. We were then told that the FSA was down for maintenance.
Students' essays would appear to be lost when logging back in requiring students to wait numerous days to continue so essays could be retrieved. Students were not sure if all data was retrieved.
Responses lost. No mechnism to recover lost responses. Responses not saved as often as was supposed to be. (Debrief provided explanation that students needed to have clicked on tools, not just typed, to have responses saved.)
Writing was lost in part or in total.
Lost student work
Students with tickets unable to log in due to "no test available."
Tests started "indenting" randomly within the student's writing
Student's lost work when test unexpectedly closed.
Students were not made aware of internet connectivity issues so they continued to type and then when the connection was restored everything that they typed while the connection was lost disappeared. If a student was kicked out of a test (for unknown reasons), when they logged back in some of their work was missing. It was too easy for students to mistakenly erase their work. Sometimes they highlighted a portion of what they wrote and perhaps hit a key and all of the text was lost. We were told that the system would save students work every 2 minutes, however later at the debrief we were told that it only saved if a student used certain functions like bold, italics, etc.and that it didn't have an auto save feature otherwise. We were also told that if students were actively working in a session it would not time out, however there were many sessions that closed "unexpectedly" and kicked the students out. It was later told to us that the timeout feature of 90 minutes was actually shorter and that may have caused some of these occurances.
Student's data was not able to be retrieved until later (or not at all in two cases). Therefore, some students saw the prompt, began their essays, and did not finish them until nearly 2 weeks later.

Student force out and answers not retained when logging back in.
tests were lost
A student's writing test closed out and could not be recovered. After phone calls to FSA Help Desk, and DOE, the students test was recovered, and the student was able to continue the test, but this took days to accomplish.
Students tests were cleared out and had to wait to either write their assessment again or get their assessment recovered.

ELA-W5

How would you classify the impact of these issues?	
No impact	5.66% (3)
Minor impact	20.75% (11)
Moderate impact	35.85% (19)
Major impact	37.74% (20)

ELA-W6

How did your district respond to test administration issues that were encountered? (check all that apply)	
Waited for issue to be resolved and then continued testing as scheduled	94.34% (50)
Postponed testing to a later date	79.25% (42)
Other, please describe	11.32% (6)
We kept following instructions and pulled our students out of class over and over and over and over again. We were told that their missing work had been recovered, which many times, it had not. So, again, the kids were sent back to class and then brought back to test again. Ironically, the security on this test is supposed to ensure that the students do not see the prompts and then provide their answers on a separate day, yet that is what happened to at least 50% of our students.	
Many students had to sit for the test multiple times	
Students had to re-enter responses	
Due to logistics, sharing of cafeteria, displaced classes for CBT tests to be administered , proctors and test administrators needed. Postponing the testing caused a major disruption. Also, some of our students had access to the writing prompt but could not complete their test for several days. A few of our students did complete their test and it wasn't saved, so they had to take the writing test again.	
Called FSA Help Desk and/or FDOE representatives	
A few instances where students were testing and having technology issues, we continued testing as students had already begun the test.	
Initially tried to wait for issues to be resolved, then technical difficulties, followed by denial of service in our district caused all testing to be delayed.	
Because the information about the problems we were facing came in so slow to the districts we mostly had schools waiting for the problem to resolve. But in at least some of these cases we should have postponed because the problem was not going to be resolved.	
We advised schools to do their best to encourage students during the testing issues. However, most students remained frustrated.	
We did postpone on one day...after waiting all morning for the issue to be resolved.	

ELA-W7

Did your district encounter any challenges related to student's work being lost or not saved during the writing test?	
Yes	77.36% (41)
No	22.64% (12)

ELA-W8

If yes, please answer the following: How would you classify the impact of this issue?	
No impact	10.87% (5)
Minor impact	28.26% (13)
Moderate impact	21.74% (10)
Major impact	39.13% (18)

ELA-W9

Approximately what percentage of students in your district was impacted by this issue?	
None, 0%	15.09% (8)
1-9%	52.83% (28)
10-19%	13.21% (7)
20-39%	5.66% (3)
40-59%	11.32% (6)
60-79%	0.00% (0)
80-100%	1.89% (1)

ELA-W10

Please describe any other test administration issues related to the FSA ELA Writing test here.
Screen would black out had to constantly reopen the test.
There are multiple concerns regarding this issue. For a test that supposedly relies on students answering the question in the moment, and not having time to go home and construct an answer, this test failed miserably for probably at least 50% of our students. For the ones whose work wasn't lost, there was anxiety regarding whether their work was actually saved, what exactly did AIR receive. In addition, the continual interruptions of people running in and out of testing labs had an incredibly negative impact on all the students, not just the ones testing. Then, there are the students, who just shut down after the 3rd or 4th time and said, "I'm not writing this again". They would type in a few words and submit and say they were good to go. I cannot imagine that any of these tests can be considered reliable.
Practice tests were not available in sufficient time, so students had limited opportunity to work with the tools. There was a mismatch between the Practice test and real test (spell check available) The students work was to save every 2 minutes, and it did not. It saved whenever a student used a test tool. Because the test tools were not seeming to work properly, they were avoided by many students - afraid to use them for fear of losing their work. Teachers got bumped out while students were active, interfering with their ability to monitor the testing
Our assessment calendar started the writing administration with elementary grades that were paper-based. We were supposed to begin CBT of Writing on Tuesday. But when we heard about all of the problems, we postponed and reworked our administration schedule to begin on Thursday and pushed back our calendar for writing from there. We had substitutes lined up, etc.

Please describe any other test administration issues related to the FSA ELA Writing test here.
Students were kicked off and when they went back in they would find missing paragraphs or all there work was gone.
Lost writing results - but were retrieved.
Students being timed out without any indication that it had occurred.
No matter if you lose one or one million responses, that is a major impact... the trust in the system to capture responses is gone.
As previously mentioned, postponing the test caused a major disruption for our middle school. Also, students who had the opportunity to begin the writing test but not complete it had a few more days to consider to the material and the writing prompt. The few students who had a complete test lost were very upset. The school had a concern if they made the same level of attempt the second time.
Students kicked out of test and not able to log back in to complete their essays for numerous days (Congrats box would appear instead).   -Sound for "text-to-speech" did not work for students who needed the prompt read to them (accommodations).   -Stu
It's hard to estimate the amount of tests that were lost or did not save as we don't have the results for tests to date. We were provided with the list of students who took Writing and it appears a large number of students did not take the test. I cannot confirm that the numbers yet as school personnel are not able to verify the numbers until they return from summer break. The numbers are high which is a cause for concern and we plan to follow up to determine how many, if any, were not reported because the test was lost.
While the estimated percent impacted seems quite low, hundreds of students in our district experienced difficulty in completing the writing assessment. We have not way to effectively quantify the full number of students who may have been impacted.   Students' responses that were lost were rarely recovered.   Some students retyped their responses, rather than going through the help desk request process, which may have impacted the validity of the responses.   Also, some students who initially signed in were unable to complete the response, and signed in days later (in isolated cases, during the additional makeup sessions), which may have impacted the validity of the responses.
Bottom line is we really don't know what percentage of writing was lost. It was discovered so late that we at the district feel that many students writing was lost but they never knew it because they wrote and then submitted and logged out, never knowing that because of the way the system was saving (differnt than what we were told) was not what we thought.
The fact that students could continue to type while disconnected from the FSA system (local network or Internet connection) I believe lead to the students lost work. The notification that the computer is disconnected is not noticeable and allowing students to continue typing when the response is not being saved is an issue.
The test administration issues were adequately addressed in the response choices. We experienced these problems for all schools who attempted to administer in the first week.   Please note that the percentages in the responses above assume as the denominator the total number of students in Grades 8 through 10. Because of the student computer ratios that we improved before the administration, we had a fairly large percentage of students impacted by these concerns. We also did our best to continue with administration throughout the first week.
While the number of students impacted was not large, one student is one too many! We had students who worked for an hour only to learn none of their work had been saved, Some of these students had to return to attempt testing MULTIPLE times only to learn that very little, if any, of their work had been recovered.   We were told by the helpdesk that Tier II technicians were unable to view actual responses; rather, they uploaded the file that looked to be the right size of file....
Test data was lost during the test administration. Several days later the data was recovered.

Please describe any other test administration issues related to the FSA ELA Writing test here.
We had two students whose work was never recovered by AIR and we were told to have them take the test again in the next makeup window. I have to say that I am really shocked at the amount of problems students encountered during the writing test, especially since we conducted a field test prior to the actual test. I don't remember student work being lost during the field test like we had in the actual test. I have to say that the lost work caused much frustration and angst to the students. I can't imagine writing an answer only to have it disappear either in whole or part and then be asked to rewrite it all over again. Also, we have never allowed students to come back to a test on a different day to finish and that had to be done a countless number of times due to the problems encountered. This definitely gives an unfair advantage to students because they have more time to think about how to answer than their peers who took the test in one sitting. This administration was riddled with challenges and inconsistencies.

Test Administration – ELA Reading (ELA-R)

Please answer the following questions regarding the **computer-based** test administration of the FSA ELA Reading within your district.

ELA-R1

Did schools in your district encounter any technology issues during the administration of these tests?	
Yes	90.57% (48)
No	9.43% (5)

ELA-R2

If yes, please answer the following: Approximately what percentage of students in your district was impacted by technology issues during the administration of these tests?	
None, 0%	7.55% (4)
1-9%	24.53% (13)
10-19%	16.98% (9)
20-39%	30.19% (16)
40-59%	13.21% (7)
60-79%	1.89% (1)
80-100%	5.66% (3)

ELA-R3

Please indicate the grade levels where difficulties with the FSA ELA Reading Test were encountered (check all that apply).	
Grade 5	69.81% (37)
Grade 6	83.02% (44)
Grade 7	83.02% (44)
Grade 8	81.13% (43)
Grade 9	81.13% (43)
Grade 10	81.13% (43)
None	9.43% (5)

ELA-R4

Please indicate the types of issues that were encountered (check all that apply).	
Challenges logging into tests	76.92% (40)
Test sessions were closed unexpectedly forcing students to log back in	82.69% (43)
The test session ran too slowly and interfered with student navigation across the test	42.31% (22)
Test system did not function as expected during testing	67.31% (35)
None of the above	9.62% (5)
Other, please describe (see next page for responses)	23.08% (12)

Other, please describe
Major issue was students going into Session 2 accidentally due to the flaws in the system.
Navigation through the test was fraught with problems, both during the test (jumping to a particular item) and at the end when students tried to go back a check their work
students getting into session 2 prematurely
Sound problems
TA's unexpectadly forced out of test, Denial of Service Attacks
Some students were unable to review written responses in the reading text box. We also had the reverse with other students that were not able to see the letter response but were able to see the text response when they reviewed there reading test.
Audio did not work
Students having to take both sessions in one day due to accidentally accessing and answering items in the next session.
Students were able to access day two of the test on day one. This meant numerous students completed the entire test in one day. Some started on the day two test, stopped and it was not discovered until the next day. This meant numerous students recieved extra time in session two.
Problems moving from session to session and problems with audio.
Access to Segment 2 was not working as planned
Listening passages did not work even when the initial sound check did work. Students would begin testing able to hear but the sound would diminish or have static as the students progressed.
tools did notr work correctly
if students chose white text on black background, which is a choice for them before entering the test, some students were not able to see the passages. They had to log out and back in and choose a different combination to be able to see the writing in the passages, which interrupted the flow. Students reported some question types didn't work as they should, for instance drop and drag didn't work properly, some questions wouldn't allow the student to move the item to the appropriate place.
Students were able to get into Session 2 on Day 1.
System would show questions unanswered but navigation back to unanswered questions were confusing and frustrating to students.
Students being locked out of session 2 even though they never accessed the session.
When tests had 2 sessions, it was confusing to pause session one for the students and TAs.

ELA-R5

How would you classify the impact of these issues?	
No impact	7.84% (4)
Minor impact	19.61% (10)
Moderate impact	47.06% (24)
Major impact	25.49% (13)

ELA-R6

How did your district respond to test administration issues that were encountered? (check all that apply)	
Waited for issue to be resolved and then continued testing as scheduled	97.96% (48)
Postponed testing to a later date	65.31% (32)
Other, please describe (see next page for responses)	8.16% (4)

Other, please describe.
We had students log back in if they were kicked off. We also encountered that students would highlight items, but highlight would disappear when they would revisit the item.
Contacted FSA Help Desk and/or FDOE contacts
Many schools continued to test or try and login to test without distict approval as they felt strongly that they would run out of time or further impact their instructional time and schedules negatively if they did not complete testing. There was a high level of frustration among students and staff as some students were held over an hour trying to login to test.
Fixed issue ourselves if we could
contacted FDOE directly to re-open sessions, etc.
We had some where they logged into the wrong sessions.

ELA-R7

Did your district encounter any challenges related to the use of headphones during the listening items on the FSA ELA Reading test?	
Yes	65.38% (34)
No	34.62% (18)

ELA-R8

If yes, please answer the following: **How would you classify the impact of these issues?**	
No impact	23.91% (11)
Minor impact	50.00% (23)
Moderate impact	17.39% (8)
Major impact	8.70% (4)

ELA-R9

Approximately what percentage of students in your district was impacted by this issue?	
None, 0%	30.77% (16)
1-9%	36.54% (19)
10-19%	13.46% (7)
20-39%	11.54% (6)
40-59%	1.92% (1)
60-79%	1.92% (1)
80-100%	3.85% (2)

ELA-R10

Please describe any other test administration issues related to the FSA ELA Reading test here.
Reading passage booklets, items looked like a passage but was not and confused students felt they were not provided correct booklet.   No speech to text as told weeks before testing, was planning to use before but was pulled the program a week before testing.   Going back to review student had a hard time finding the question that was unanswered due to computer program.
Vary rare and sporadic issues.
Again, as with the Writing (perhaps not as much as with the Writing), students accessing portions of the test prematurely, being stopped and then being allowed to go back in the next day, seems to negate the reliability that this test is measuring what it should. How do you tell one student who did not have overnight to think about the questions that his/her test is equal to the other students who had time to think about it? This is especially true for the Grade 10 ELA. How can a student be told he/she did not meet his/her graduation requirement, while another student who was provided multiple times over multiple days to answer did meet his/her? In addition, with the Writing debacle being considered as part of this ELA score, this report as to pass/not pass seems even more suspect. I imagine there will be lawsuits by parents if this stands.
Losing access to test to speech at the last minute was a major issue for Students with Disabilities.
We did get some calls about not having a listening item on the test. It seems that not every session had a listening item which confused students.
Described previously.
Volume settings where difficult to manage for mac devices
Sound would not function correctly. Would have to work with it to fix it.
We had to change settings to the devices.
Students having to take both sessions in one day due to accidentally accessing and answering items in the next session. -Students who needed to re-access a session to continue working needed to wait for sessions to be reopened (this took from 5 minutes
One school had their headphones on mute. Once that was determined the issue was resolved.
The reason that all our answers are none or 0% is that we did not administer the computer-based Reading test to any of our sensory-impaired students due to accessibility concerns. All of our Deaf/Hard-of-Hearing and Blind/Low Vision students took paper-based Reading tests.
Problem for Reading and Math encountered during transition between sessions. Delay in reopening sessions and students being re-entered into the session started the prior day could impact the validity of the results.   Again, while the percentage of students impacted seems small, this amounts to hundreds, possibly thousands of students in our district. We have no way to effectively quantify the full number of students impacted.   Should specify in manual that headsets must be plugged in and volume need to be set prior to login.   Also, was confusing that some sessions/students had audio components and others did not - it should be specified which session have audio, and which do not.   Hard of hearing/deaf students should have form that allows them to participate without audio.   Text to speech was cancelled at the last moment before administration.   Issue of what could and couldn't be read for students with that accommodation was confusing.   alidity of the results.
Schools would check all settings to make sure headphones were working AND would do the sound check and headphones would work but then something would switch and they wouldn't be working!
The fact that headphones had to be in place prior to the secure browser being started was not communicated well up front. Second, some test segment had no audio. This was confusing to students and took a lot of setup time for the schools where it was unnecessary.
These issues were a problem the most during the week of April 20 when updates were made to the administration platform.

Please describe any other test administration issues related to the FSA ELA Reading test here.
We contacted helpdesk first and then advised to make sure headphones were properly installed, asked to change computers, or adjust volume as guided by FLDOE. A helpdesk agent kept advising us that the Text to Speech was not available even though we confirmed that we were aware. He then asked us why it was important that students hear during testing!!
Computers would need to be repeatedly restarted to achieve audio connectivity. This was an FSA issue and was validated by testing audio capabilities with other audio functions on the computer.
While this administration didn't lose student work like the writing administration did, it had its own problems. The test was poorly designed and it was too easy for students to get into the next session. Even though it required test administrator approval, many test administrators were confused because of unfamiliar terminology--session vs. segment. Many students were allowed into the second half of the test accidentally which caused one of two problems, either they had to log out, if it was caught right away and continue the next day, which again is an exposure issue, or two, they were allowed to finish the second half on the same day as the first half and the test was not designed to be administered that way. We have been told that psychometrically this is wrong and the test should be administered over two days.
At least twice, there were server errors on the part of AIR that resulted in our entire district (along with other districts in the state) being unable to test for significant periods of time. We also had ongoing issues with TDS and ORS communicating efficiently and accurately, making it even more difficult to determine when technological errors had occurred. The steps required to get students back into sessions was tedious. Only the DAC could make these requests -- I am 1 person for 160 schools. Most issues involved making multiple requests. I spent all of the testing window trying to get the students back into the correct sessions.
Sometimes it was hard to ear the listening items on the test even with volume at high level.
There were problems with volume. the instructions provided were inadequate, but as testing were on, we discovered ways to overcome problems. It would have been great if these solutions were available in the directions. I feel that the directions were often vague or inadequate.
There were issues with students who couldn't hear the audio questions even after the sound check was verified at the beginning of the test. Students logged out and logged back in to retry the audio questions.
When headphones were plugged in or unplugged during the administration it caused issues with testing. We were not told ahead of time that this would cause an issue for students.

Test Administration – Mathematics and End of Course (EOCs) Exams (M-EOC)

Please answer the following questions regarding the **computer-based** test administration of the FSA Mathematics and EOC (Algebra 1, Algebra 2, Geometry) within your district.

M-EOC1

Did schools in your district encounter any technology issues during the administration of these tests?	
Yes	90.38% (47)
No	9.62% (5)

M-EOC2

If yes, please answer the following: Approximately what percentage of students in your district was impacted by technology issues during the administration of these tests?	
None, 0%	9.62% (5)
1-9%	32.69% (17)
10-19%	13.46% (7)
20-39%	30.77% (16)
40-59%	7.69% (4)
60-79%	1.92% (1)
80-100%	3.85% (2)

M-EOC3

Please indicate the grade levels where difficulties with the FSA Mathematics/EOCs Test were encountered (check all that apply).	
Grade 5	50.00% (26)
Grade 6	59.62% (31)
Grade 7	67.31% (35)
Grade 8	80.77% (42)
Grade 9	78.85% (41)
Grade 10	82.69% (43)
None	9.62% (5)

M-EOC4

Please indicate the types of issues that were encountered (check all that apply).	
Challenges logging into tests	65.38% (34)
Test sessions were closed unexpectedly forcing students to log back in	75.00% (39)
The test session ran too slowly and interfered with student navigation across the test	38.46% (20)
Test system did not function as expected during testing	67.31% (35)
None of the above	7.69% (4)
Other, please describe (see next page for responses)	21.15% (11)

Other, please describe
Again, the same issue as with Reading - students being moved accidentally into subsequent sessions.
Test administrators were unable to effectively monitor student progress through the test
students using handheld calculators for non-calculator sessions
I believe this was the test when AIR did an update and we could not access the interface the morning of testing.
Students were able to choose the incorrect test. Pre-ID was not extremely helpful.
Denial of Service Attacks
Problem stated previously with moving from sesison to session, especially for 6 - 8 graders, who had three sessions.   EOCs couldn't specify which subject they were taking, and some students logged into wrong test.
problems accessing the calculator,
Transitioning between segments did not work as expected.
Venn diagrams did not display correctly, student worked "stacked" on top of previous work, students were unable to select answers or save answers, they had to go back in and retry time and time again,
Students indicated that when trying to choose an answer choice, nothing would indicate that an answer was chosen. They clicked on all choices and none of them filled in black.
Students were able to get into sessions they were not supposed to be in, which resulted in a multitude of tedious steps to reopen tests, reopen test segments, etc.
Students given message that they completed test when in fact they never even answered a single item in session 2.
Confusing to pause session 1.
Some logged into the wrong sessions

M-EOC5

How would you classify the impact of these issues?	
No impact	9.62% (5)
Minor impact	32.69% (17)
Moderate impact	48.08% (25)
Major impact	9.62% (5)

M-EOC6

How did your district respond to test administration issues that were encountered? (check all that apply)	
Waited for issue to be resolved and then continued testing as scheduled	95.92% (47)
Postponed testing to a later date	57.14% (28)
Other, please describe	8.16% (4)
Had the students log out and then log back in.	
Called FSA Help Desk and/or FDOE representatives	
Domino effect from earlier delays caused scheduling issues for schools with difficulty completing testing within the window, no matter how long the window was.	
Fixed most issues ourselves when we could	
Contacted FDOE to re-open test session	

M-EOC7

Did your district encounter any challenges related to calculator use on the FSA Mathematics assessments?	
Yes	59.62% (31)
No	40.38% (21)

M-EOC8

If yes, please answer the following: Please indicate the types of 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)

M-EOC9

How would you classify the impact of this issue?	
No impact	34.69% (17)
Minor impact	32.65% (16)
Moderate impact	14.29% (7)
Major impact	18.37% (9)

M-EOC10

Approximately what percentage of students in your district were impacted by calculator-related issues?	
None, 0%	38.46% (20)
1-9%	32.69% (17)
10-19%	9.62% (5)
20-39%	7.69% (4)
40-59%	1.92% (1)
60-79%	5.77% (3)
80-100%	3.85% (2)

M-EOC11

Please describe any other test administration issues related to the FSA Mathematics/EOCs test here.
Was very unfair to invalidate students that used calculators on day one when no directions said, NO CALCULATORS can be used in Session 1. All student in our district used hand held calculators, we told our test coordinators. IT was not a problem in our district but I heard many others that did and thought it was very unfair to punish student for teachers mistakes.
The TEI that malfunctioned caused major issues with schools and students becoming frustrated and upset. Being told that the "items are functioning as they should, but students must not be reading the directions" seems to indicate that the question was a trick question insofar as technology is concerned, but not based on Math standards. Finally, those questions that students across the State

Please describe any other test administration issues related to the FSA Mathematics/EOCs test here.
had issues with were dropped, but how many kids became frustrated and confused, which would affect the rest of their performance.
Major difficulty with teachers allowing calculators for non-calculator sessions.   Major difficulty getting the state to approve a model calculator that we selected.   Major difficulty getting enough calculators into the schools hands.
Selection was difficult. Parameters were clear but knowing which to choose was time consuming. I would prefer that we were told one approved calculator. We made sure that all students were given a hand held as well as informed about the calculator on the computer.
invalidated a few classes because of calculator issues
Students received an error box stating that an answer was not submitted when trying to navigate to the next item, even though an answer was typed in the box.   - Students kicked off during sessions for multiple reasons and multiple times due to various
It was a struggle to provide calculators as the policy changed midyear to allowed them, by that time, funding was an issue. Some schools were able to purchase them for students and some were not.
Invalidations were relatively minor due to calculator issues, but the full impact is hard to determine. Students prefer using a hand held calculator, so it is difficult to say what the impact would have been if handhelds had not been allowed.   There must be conformity in the calculators used AND reference sheets used across tests for the same subject area. It makes no sense that some facts and/or functions are considered critical for different tests covering the same subject.   In addition, allowing the use of calculators for just one session of the test invites human error.   Three sessions for middle school students causes fatigue for the student, and a scheduling nightmare for schools. Particularly for schools with large populations of ELL and ESOL students, who must proved extended time.
We had to invalidate an entire class of math scores due to them using the calculator fro session 1 -
Since the onscreen calculator was not a replication of an existing calculator it was nearly impossible to find handheld equivalents. This resulted in students having limited practice with the calculator they would see on the test.   Also, for the non-calculator segments the instructions were very specific to test administrators, but there was nothing in the script read to students saying explicitly no calculators. We had a few students invalidated beacuse they took out their own calculator unknowingly.
This is an area where having more time to train and repeat new information to schools is critical. Providing this information to thousands of teachers and hundreds of schools takes time, particularly when the process is different from the last few years of administration.
We needed a list of approved calculators not approved features.
Many issues with the online calculator. There seemed to be some confusion between the prior approved calculators and the current handheld calculators.   There is currently a request specific to our district pending a response from DOE in regards to exact calculators for use during the exam.
Because the pre id file for EOCs didn't identify which test a student should take it made all 3 available when they logged in. This caused some students to take the wrong test. These tests had to be invalidated and then the student had to take the correct test. This wasted the student's time.
We had a couple of students whose tests were invalidated, because they accessed their own personal calculators during testing (not approved calculator and/or during Session 1).
Confusion on calculator use for certain sessions of the test and what was considered approved functions on the scientific calculators. This problem lead to numerous invalidations.

Other (O)

O1

Please feel free to provide any additional information about the impact of technology on the 2014-15 Florida Standards Assessments administrations.
Students had to wait 20-30 minutes for computer let them in to test. A big problem at the beginning but did get better. Screens blacking out and had to reboot many times. Student running over a session and had to get approval to reopen test and many times it did not work so student are sitting over 20 minutes to get correct session to open.
Frustrating dealing with re-opening of test sessions. Wait time for assistance.
The majority of technical issues encountered in Lee County occurred during the first week of FSA Writing testing. After that, we experienced sporadic but continuing issues.
When administering more than one session in Pearson, students are required to enter a "Seal Code" in order to proceed to the next session. AIR may want to consider the same type of procedure.
Our schools need accurate screen shots, or a realistic training site that will prepare them for the actual testing day(s). School Coordinators and District Coordinators need to have access to the testing sessions without having to actually go into the testing room. This is extremely critical. There should be seal codes, or some type of barrier, to prevent students from moving to the next session. School Coordinators should be able to assign tests to the Test Administrators, rather than have all the tests listed on a drop down. This will eliminate TAs picking the wrong test. Accurate information needs to be shared. Being told that students' writing work was saved every 2 minutes turned out to be false. After the testing window was closed, information came out that ONLY if students clicked on features such as Italics, Bold, Highlight, etc. would their work be saved. This is even more ridiculous as during the test, so many issues happened when students did use those features, so students were instructed to try to NOT use those features.
The technology (software) should support the student working - rather than requiring the student to also master the technology and the content
The most disruptive part of the administration was students getting into sessions too early. This was a long process to correct and occupied all the district assessment staff's time, while students sat in front of a computer with nothing to do.
Below are some important dates that reflect the level of issues we experienced. In completing this form, I realize that the problems never stopped for the entire window, we just got better at resolving them sometimes even at the school level. Thu 3/5 - Writing Network Outage 7:43 Malicious Attack State indicates AIR server outage in the AM WRHS suspends some am testing. Finds discrepancies with TIDE and student information used of accommodations and ticket generation as well as reversions back to original data after changes to tested grade level have been made GHS tests am and pm sessions, issues with TIDE reverting student demographic info back to original version after changes have been made to tested grade JIEC missing students entered into TIDE that have already tested (10 students) Fri 3/6 7:28 Network Issues. Schools report sporadic students even in same lab cannot access Secure Browser and some computers cannot access TA site. LOLHS, AHS, PHS, JWMHS cannot access secure browser or TA site. GMS, PMS, RBSMS can access both sites. After 3+ hours, DSBPC IT staff determines issue with login4.cloud1.tds.airast.org dns root server has corrupted data. Any computer pointing to that DNS as a primary with no secondary DNS will not be able to access AIR testing resources. Testing resumes by 10:45 at schools. The ability for a student to enter the second session of the test with such ease created havoc for scheduling and more students than I have ever seen (over 3 years of testing) took both sessions of a 2 day 2 session test in just one day.

Please feel free to provide any additional information about the impact of technology on the 2014-15 Florida Standards Assessments administrations.
The ability to start tests and log on was the biggest disruption to testing. With the disruption to the school day testing offers, to postpone a test or start a test later was difficult. The test directions and time it took to start a test was ann issue.
In summary, SCPS experienced systematic failures of the testing platforms, flaws in test design/construction, delays in vendor responses, concerns over the assessments reliability and validity, and the extreme loss of instructional time (ranging from 10-15 instructional periods.) Specific details include: The last minute cancellation of the Text-to-Speech functionality caused schools to reschedule students with oral presentation accommodation, and provide a test-reader for each session. Several reports from schools indicating that younger students and ESE students had challenges answering technology enhanced questions. Students stated that there were issues with drag and drop items and having to log in and out of the testing system in order to have items work properly. Approximately 100+ students were forced to complete two sessions in a single day due to the flaw in design which allowed students access to both segments without teacher approval. There was no visual cue to stop students at the end of a session and many students continued on to the next session. By the time a testing administrator notice the issue, it was too late to stop and students had to complete both sessions in one day. This situation happened on all FSA tests and EOCs. Students were not able to be tracked on how much time was utilized in the testing session, since there was no time stamp indicating when the student had begun 2nd session. Loss of instructional time - Due to the testing schedule our campus experienced disrupted classes for at least 4 weeks. We used a varying schedule of block classes and teachers may have only seen partial classes twice in a week. Testing pulled some students out on morning tests, and other students out in the afternoon. If a teacher had mixed grades (6-8), then it was possible he/she didn't have a full class for 4 weeks. ESE testing occurred daily (except on the outage day 4/20) and these students missed the most time in classes. The computer classes (Video/Web Design) lost at least 3 weeks of being in their lab (classroom), and minimal access to computers was available to those classes held during the morning for testing. Teachers were pulled to test when they had a few students in their class, sending students to be covered in other classroom with other teachers. All in all, testing displaced students and teachers so that our daily routine was not possible for 4 weeks. Rock Lake Middle School
Please note that many more Students were impacted by these issues than what was reported. Our data in this survey convey actual incidents that were reported and Students who were directly affected. However, when each incident occurred to specific student
Although I recognize that technology issues are the primary focus of this survey and your study, I hope you will also include concerns re: the paper-based FSA administration this year, especially the Braille tests.
Online Reporting System was not updated daily as intended. Because EOC eligible students were not assigned to a specific test via the pre-id upload, it was not possible to determine the percent of students who had tested, or were pending. This was also the case for 7th and 8th grade students who were not going to take the grade-level math test. Some font/background colors did not allow highlighting to show. Need a stop sign at the end of a session, and eliminate the "next" button on the review screen that takes a student to the next session. Use terminology consistently ... session vs. segment. Online teachers saw "segment", where everywhere else it is "session. Some students remained as "paused" and never moved to completed status until the end of the window, when it was supposed to be completed programatically. Question grouping caused students to return the the first question in a group, rather than the one they wanted to review. Until and unless all districts and schools have one-to-one ratios of computers, scheduling students into labs without strict schedules for each content area exposes test content. The extended window cuts into instructional

Please feel free to provide any additional information about the impact of technology on the 2014-15 Florida Standards Assessments administrations.
time and displaces students from instructional labs. Writing and ELA administration during separate windows is problematic for students missing one component will not get a score - especially for Grade 10 graduation tests. While the later makeup sessions are needed for this reason, it was very difficult to coordinate these, simultaneously with all the other components. Would also like the opportunity to register concern with the ordering and shipping of special paper materials at some point.
One of the biggest problems was the technology did not prevent students from going forward into sessions they were not supposed to access until the next day. We at the district spend hours and days reopening students into previous sessions, giving students extra time and generally approving everything because there was no time to look into the problems students were experiencing. We also had no way to check what the issue was even if there was time to do so.
As a district, it appears that we did better than most BUT that is because WE figured out how to fix our issues because of having a top notch team. WE also had a major internet issue where a cable was cut and was a NIGHTMARE for me to get thousands of students reset in the TIDE system. TIDE is not user friendly - I figured out its quirks but it is NOT designed for the way districts need to use it.
The FSA system does not seem robust enough to handle large scale assessment.
The administration of the writing portion of the ELA assessment in particular was poor. Many students were kicked out of the assessment multiple times or lost large portions of their responses. Some students had to be assessed over multiple days due to problems with the administration platform. This raises large concerns with the security and validity of the assessment. Students were very frustrated with the process, and this also likely impacted the administration and scores. This is particularly critical for Grade 10 ELA students who were taking this assessment as a graduation requirement.
How many testing issues are too many when considering students are negatively affected by this testing fiasco? One. One issue is too many because of the enormous consequences on students, teachers and schools. FLDOE continued to quote numbers of how many students had completed computer-based testing. How every day got better and the state was on track to complete the FSAs. What FLDOE failed to address were the number of students who had to make repeated attempts to participate in tests. Students were sent to the computer lab, sat in front of a computer with a ticket in hand, only to get an error message, blank screens, or slow to load tests. Approximately $200 million was spent by FLDOE on an assessment program that was ineffective. One consistent worry has always been secure and proper testing situations. Test administrators have to ensure that test items are kept secure; the testing environment is conducive for a relaxed location and free of distractions. This was IMPOSSIBLE with the FSA. With SO many error messages and issues, frustration and stress levels were through the roof!
A very trying year of exams. Student instruction is totally disrupted.
The very high stress level actually began with the students. They came prepared to test, became frustrated, fatigued and discouraged. The discussion they overheard from school staff, and the rumors of scores not counting impacted their effort on postponed attempts to test. In other words, they stopped taking it seriously. The security of the content had to be compromised because there were students taking the same test over a number of days. Several got to attempt the same questions multiple times. Some had to re-write the same essay multiple times as well.
After looking back at my notes from the administration I forgot to mention in the reading portion that the line reader feature wasn't working for some students and there was also a passage in the listening portion that wouldn't play. We were also told that there would be text-to-speech provided in the platform and then right before the reading/math administration we were told that it wasn't working and would not be availble to students, which forced schools to provide adult readers, which was a

Please feel free to provide any additional information about the impact of technology on the 2014-15 Florida Standards Assessments administrations.
problem for some schools because they didn't have enough personnel. The listening portion caused issues for our deaf students because interpretors had no way to prepare beforehand and then it was reported that some of the content had to be interpreted by spelling the words out because there was no signage for them, which was labor intensive for the interpretors. In addition, the students who wear hearing aids were not able to wear the headphones because of feedback caused by them in the hearing aids so those students had to be put in a location by themselves with a test administrator so they could take the test without the headphones. This again, caused manpower issues for schools. It doesn't seem that a lot of things were well thought out before the administration. It seems like there should have been a better field test done much earlier so that these problems could have been discovered and corrected before students were put in a high stakes testing situation. The technology issues encountered during live testing caused much distress and led to a bad testing environment for students.
There were a lot of glitches in the system: 1.) State-wide server issues 2.) Poor resolution process for reopening test segments 3.) ORS inaccuracies 4.) Lag time between the various AIR systems (TIDE, ORS, TDS) 5.) Students were being pulled from our TIDE to other districts' (scores reported to those districts). 6.) The testing platform told numerous students they had not answered a question, when it was answered.
Due to technology staff members needing to work on troubleshooting in the testing rooms, although it does not have a great impact, it was distracting to students to have any type of additional movement in rooms.
Some of the technology enhanced questions were not straightforward....a technology "savvy" individual could have trouble executing some of the technology enhanced math questions. These tests should not measure how a student can execute a question using a mouse and keyboard. The test should measure what the student knows about the content of the course/subject.
Our district was very fortunate with any major issues with any of the new FSA assessments.
There was a issue that became a major issue in our district during testing. Students were able to move ahead in the same session code and were not always caught until it was already approved. The next day the students session or test had to be reopened to be able to continue testing. This caused a lot of testing delays. We feel that each session should require a new session code.

Appendix D: District Focus Group Meeting Results

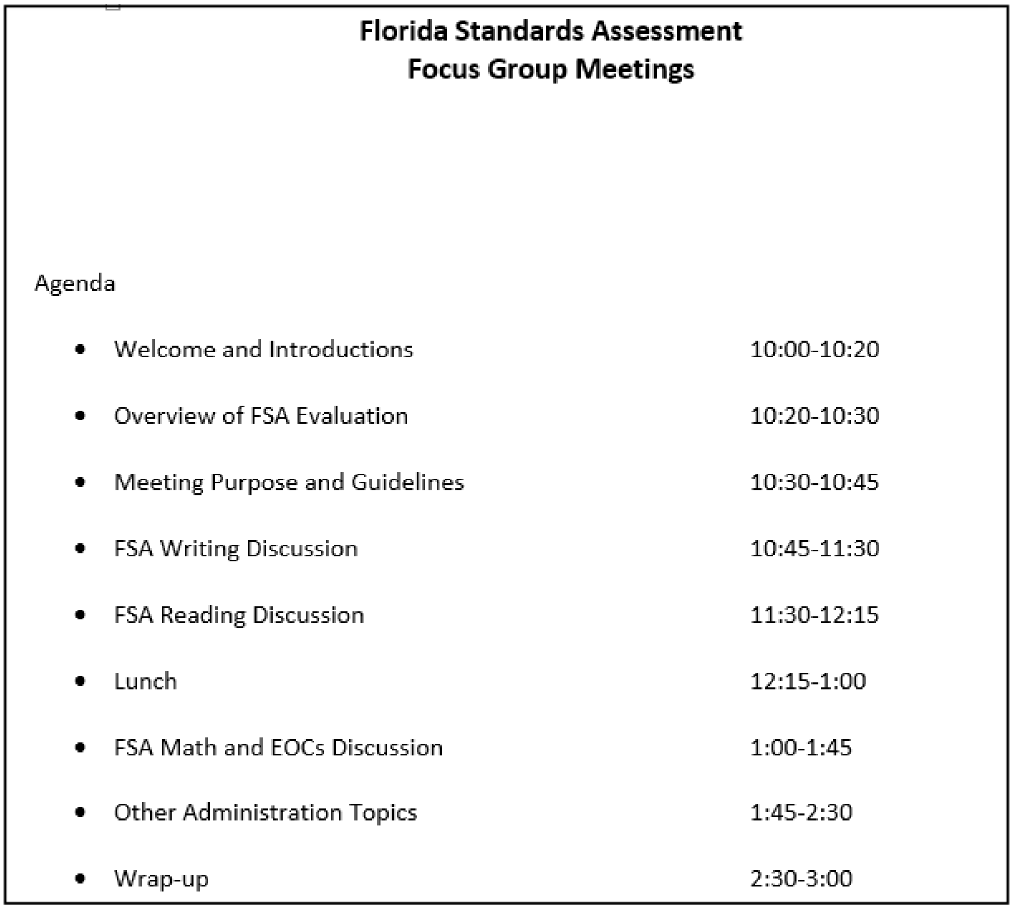
To gain additional insight into the spring 2015 Florida Standards Assessments (FSA) administrations, the evaluation team conducted three focus group meetings with district assessment coordinators and other district representatives. The team worked with members of the Florida Association of Test Administrators (FATA) to coordinate the meetings. Miami, Orlando, and Tallahassee were selected as the meeting locations to make attendance feasible for all districts. Using a list of district assessment coordinators and contact information provided by FLDOE, invitations were emailed on July 6, 2015, to all 76 Florida districts. Up to two representatives per district were invited to attend. A reminder email was sent on July 8, 2015. No compensation was offered for attendance, and participation was voluntary for districts and their staff.

Across the three focus group meetings, a total of 56 participants from 33 districts attended as shown in Table 27.

Table 27 District Focus Group Participation

Location	Date, Time	# of Participants	# of Districts
Miami	July 15, 10am-3pm	9	4
Orlando	July 16, 10am-3pm	30	21
Tallahassee	July 16, 10am-3pm	17	8
Total		56	33

Each meeting was facilitated by two Alpine staff members. Drs. Tracey Hembry and Andrew Wiley facilitated in Miami and Orlando. Drs. Chad Buckendahl and Brett Foley facilitated in Tallahassee. The agenda that was shared with participants and used to guide conversations is shown in Figure 13. For each agenda topic, the facilitators reviewed the preliminary survey responses (i.e., those responses received on or before July 13), asked follow-up questions related to these responses, and asked participants to comment as to whether the survey information accurately represented their experiences. Participants were also asked to share information that was not included within the survey (e.g., other administration issues experienced). At the conclusion of each focus group meeting, the facilitators reviewed key themes and common feedback with the group to confirm accuracy and understanding.



Florida Standards Assessment
Focus Group Meetings

Agenda

• Welcome and Introductions	10:00-10:20
• Overview of FSA Evaluation	10:20-10:30
• Meeting Purpose and Guidelines	10:30-10:45
• FSA Writing Discussion	10:45-11:30
• FSA Reading Discussion	11:30-12:15
• Lunch	12:15-1:00
• FSA Math and EOCs Discussion	1:00-1:45
• Other Administration Topics	1:45-2:30
• Wrap-up	2:30-3:00

Figure 13. Focus Group Meeting Agenda.

The following sections, organized by agenda topic, list feedback and experiences shared by the districts at the focus group meetings. Unless otherwise noted, the comments were heard at each of the three meetings.

Florida Standards Assessment Writing

- Many of the system-related issues occurred early in the writing window and impacted the schedules of these test administrations.
- Districts reported that many students lost work during the administration. Some of these cases of lost work could have been related to the inactivity timer issue that AIR experienced with its system. In other cases, districts reported that this could not have been the cause. Many districts reported that students lost work after attempting to use one of the system tools (i.e., the highlighter or line reader).
- For the students who lost work, the resolutions were not consistent. AIR was able to recover work for some students. In some cases, the recovered work was only a small portion of the student’s response or the recovered response was gibberish (i.e., a mixture of random symbols and letters). The time it took to recover lost work also varied greatly and, in some cases, took weeks.

- While districts are aware of many students for whom work was lost, the districts felt that likely more students experienced the same issue and did not report it.
- Various testing delays, both system related and those related to individual student issues, led to increased students' exposure and knowledge of the prompt as compared to prior years.
- Some students had difficulties with the "Submit" button at the end of the test. Students reported that the "Submit" button did not appear or that they were not able to select the option. Instead, students had to close out their session and log back into it. When this was done, some students lost work.
- Some districts were confused by the time limit for writing. Originally, the time limit was set at 90 minutes. FLDOE subsequently allowed an additional 30 minutes if needed.
- Grades 6 and 7 paper-based administrations were delayed because of the challenges experienced with the computer-based administrations (Tallahassee only).
- Significant administration challenges were not encountered with the elementary writing paper-based administrations.

Florida Standards Assessments Reading

- Challenges were encountered with the listening items. Some districts that tested early in the window noticed that issues could be avoided by plugging in the headphones prior to launching the secure browser. This information was circulated using the FATA listserv.
  - Districts felt that some students may have skipped the listening items or guessed the answers rather than reporting any issue encountered with the headphones.
  - Some test administrators learned that students had issues with headphones only after the students had completed the test.
  - Challenges with these items were more manageable and less widespread than other challenges encountered during the administrations.
- Significant administration challenges were not encountered with grades 3 and 4 Reading paper-based administrations.

Florida Standards Assessments Mathematics

- There was confusion related to the calculator policy. The initial policy did not permit handheld calculator use. Only the onscreen calculator within the testing system was permitted. FLDOE then permitted calculator use but released a list of calculator functions that were not allowed on handheld calculators. Districts reported that there was confusion around identifying acceptable calculators and that districts had limited time to select and purchase these calculators prior to the test administration window given the timing of the changes. Differing calculator policies between the FSA and the FCAT 2.0 caused additional confusion.
  - Districts reported that they could not identify a calculator for the large print accommodation that fit within the FLDOE requirements (Orlando only).

- During test administrations, there was confusion regarding the sessions for which calculators were permitted. Therefore, some students used calculators on sessions for which calculators were not permitted. Per test administration guidelines, these sessions were then invalidated.
    - Most districts reported entire classrooms of scores being invalidated for this reason.
    - In a few cases, schools invalidated scores for an entire grade level because of unpermitted calculator use.
- Significant administration challenges were not encountered with grades 3 and 4 math paper-based administrations.

Cross-Subject Issues

Movement Across Sessions

- Based on district assessment coordinator feedback, the most challenging issue encountered during spring 2015 related to students moving across test sessions. For those tests with multiple sessions (Reading and Math), districts reported that students were able to move into a later test session earlier than scheduled. Districts mentioned several ways in which they experienced this occurring:
    - Students unknowingly requested permission to move into the next session. The test administrator unknowingly approved the request.
    - Students unknowingly requested permission to move into the next session and the test administrator rejected or ignored the request.

  This movement across test sessions caused another challenge in that students were commonly locked out of the initial or subsequent sessions. This prevented the student from completing the test session during the originally scheduled testing time because both AIR and FLDOE had to be involved in reopening these sessions.

  This movement across sessions had an additional complexity of calculator use in math. Calculators were permitted on some sessions but not others in math; inadvertent movement across sessions meant that students either had a calculator when one was not permitted or that the student did not have the calculator when one was permitted.

  This movement across test sessions was also challenging to manage for students who had extended test time as an accommodation.

- Districts estimated that 10-20% of students experienced an interruption during testing. Beyond those students directly impacted, students who sat next to, near, or in the same room as a student who experienced an interruption also could have been impacted while the test administrators attempted to resolve the issue.
- In previous years, the testing system was set up to save student work on the local machine. However, the AIR system would not save if connectivity was lost. Districts felt

the issue of lost work could have been prevented if the system were set up to save locally, as was the practice in other systems (Orlando only).

- Students were kicked out of testing sessions for unknown reasons. Sometimes the students were able to resume testing; in other cases, AIR and/or FLDOE actions were needed to permit testing to continue.
  - When some students logged back into the test, they received a message that read "Congratulations on completing your test" or something similar.
- As the test administrations continued throughout the spring, districts reported that student motivation and patience continued to decrease given the challenges that were encountered.
- Student motivation may have also been impacted for the end-of-course (EOC) assessments related to changes in the policy to use test scores as part of course grades. A meeting was held with FLDOE and assessment coordinators on May 4. During this meeting, district assessment coordinators reported that they were made aware that the requirement to use EOC test scores within course grades would likely be eliminated. While the formal announcement of this change was not released until May 18, districts reported that the change was known and may have impacted student motivation during the May EOC test administrations.
- Districts reported that some of the computer-based testing tools (e.g., color contrast, the line reader, and the highlighter) did not function as expected.
- The pop-up warning related to loss of connectivity was small and easy to miss.
- Districts reported that they identified a small number of cases where students in another district were logged in and testing as a student in their district.
- Because testing was commonly rescheduled and delayed, students lost more instructional time than anticipated.
- Districts reported that the FSA spring test administration was the worst they could remember.

Read Aloud Accommodation

- Not long before the test administration window opened for Reading and Math, FLDOE announced that the text-to-speech tool would not be available. Instead, a read aloud administration would be used as a testing accommodation. Districts reported that the timing of this change left them with little time to prepare and train test administrators. For some districts, this time was further reduced by their spring break, which occurred between the FLDOE announcement and the test administration window.
- A script was not provided for the read aloud accommodation. Instead, FLDOE shared a list of what could and could not be read during the administration. Districts did not find this information to be clear, especially because the rules differed from previous years.

Specific Item Issues

- In the items where students were asked to “check all that apply”, if a student selected only one option and two options were correct, the system would not let the student continue to the next item. This cued the student to select another option (Tallahassee only).
- Some of the math technology-enhanced items did not function as expected. For example, some students experienced difficulties with the items that required interaction with and graphing on a coordinated plane.
- Drag and drop items had issues with the zoom functionality (Tallahassee only).

Help Desk

- The Help Desk was not helpful.
- Districts experienced long wait times. During this time, students were commonly sitting at computers waiting to test.
- Help Desk staff were not knowledgeable of the testing systems or the FSA program.
- Some districts reported that the support staff at the Help Desk did not have login information to access the FSA testing system. The district assessment coordinators provided the staff at the Help Desk with their login information so that the staff could see the system and the encountered issue.
- Some districts received instructions from the Help Desk that directly contradicted test administration policies.
- Districts stopped calling the Help Desk and instead, either called FLDOE or relied on the help of peers (through the Florida Association of Test Administrators).
- Help Desk tickets do not represent all of the issues experienced.

Administration Support and Communication

- Communication related to system-wide issues was inadequate and not timely. This made it challenging for districts to determine the appropriate action.
- Alpine explained the inactivity timer system issue that was related to students’ loss of work, as AIR had explained it. Many districts reported that they had been made aware that such an issue occurred during the administration.
- Many districts created their own troubleshooting guides to support staff during the administration.
- Several terms caused confusion, including “Pause” within the system as well as “Test Session” and “Test Segment.”
- Districts found it hard to navigate the various documents and email communications related to the test administrations.
- The online system for test administrators did not provide real-time monitoring of testing. This made it challenging for school and district assessment coordinators to

monitor testing programs and issues. School administrators had to enter the testing room in order to troubleshoot and resolve issues. This led to disruptions for all students in the testing room, not just the student encountering the issue. This was not an issue in years past.

- For students who transferred, it was difficult to determine if the student had already tested.
- Test administrators were told that their session would time out after 90 minutes of inactivity and that student activity within test sessions would prevent this time out. Test administrators experienced timeout issues, when students were actively testing, after 20 minutes. This timeout closed the test session for all students whose sessions were being proctored by the impacted test administrator (Orlando and Tallahassee).
- As the test administration issues mounted, the districts reported that FLDOE instructed them to wave typical test administration policies in order to complete testing. For example, students were permitted to complete testing over multiple days although this had not been permitted in the past.
- As districts shared their experiences, districts realized that the instructions they received to manage the test administration issues were not consistent.
- One district shared recent communication from FLDOE where they learned that there was a 2- day time period for which any test invalidations submitted within the system were not recorded. Instead, these students' scores were scored and reported normally. Only a few other districts were aware of this issue. This issue was discovered after scores were reported for these students (Orlando only).
- One district reported that they felt AIR did not provide adequate support or directions related to testing with Macs (Orlando only).
- The system error codes did not align to the issues encountered. For example, several districts experienced what they referred to as the "iPad error"; the error message reported that an extra program was running when this was not the case (Orlando only).

Training and Preparation

- Districts felt that they had been prepared for the administration if it had gone relatively smoothly. They were not prepared to handle the variety of issues that occurred.
- Districts mentioned that all districts were required to complete the preparation/readiness certification. Districts did not have the option of saying that they were not prepared for the test administration. Instead, many Superintendents expressed their concerns through separate letters to the state.
- AIR was supposed to provide a demonstration of the testing system during a kickoff meeting in late August, but it did not work. This was a missed opportunity for districts to provide input into any potential issues with the system (Miami only).
- Districts expressed concern about several administration-related issues at the kickoff meeting in August. These concerns included the level of monitoring and control that was

lacking in the AIR system for school administrators as well as the lack of a “seal code” on the test sessions to prevent movement across sessions. They felt that no resolutions or changes were offered for their concerns and then their concerns amounted to issues during the test administration (Miami and Orlando only).

- The training test was not an authentic representation of the actual test.
- The training test did not include multiple sessions, so the issue related to students inadvertently moving across sessions could not be anticipated.
- The training test was unavailable in the week prior to testing when some schools were planning to use it.
- Districts felt they did what they could to train test administrators, but the timing of resources and changes from FLDOE made training difficult.

Appendix E: Concurrent Users per day of the FSA CBT Test Administration Window

FSA Writing Component Concurrent Users Daily Comparison

Time	Mon 3/2 Users	Tues 3/3 Users	Wed 3/4 Users	Thurs 3/5 Users	Fri 3/6 Users
8:00:00 AM	8,956	12,697	13,873	9,068	10,866
8:30:00 AM	22,738	22,091	27,219	30,525	24,333
9:00:00 AM	29,779	30,854	31,059	43,344	29,763
9:30:00 AM	31,382	31,971	32,499	48,704	31,583
10:00:00 AM	28,704	37,063	33,242	50,617	29,798
10:30:00 AM	25,249	38,593	30,172	48,759	24,615
11:00:00 AM	22,143	33,395	25,519	40,438	18,523
11:30:00 AM	19,543	27,372	19,239	32,555	13,173
12:00:00 PM	16,496	19,249	12,857	22,807	10,273
12:30:00 PM	13,359	14,827	10,131	16,537	8,794
1:00:00 PM	10,509	12,879	10,021	14,252	8,089
1:30:00 PM	9,121	11,084	9,113	12,961	6,879
2:00:00 PM	6,599	8,295	7,773	9,877	4,159
2:30:00 PM	4,430	6,842	6,044	7,488	2,806
3:00:00 PM	2,350	4,235	3,064	4,400	1,193
3:30:00 PM	1,067	2,303	1,615	2,553	640
4:00:00 PM	386	1,021	618	813	176
Max Concurrent	**31,832**	**38,930**	**33,389**	**52,453**	**31,923**

FSA Writing Component Concurrent Users Daily Comparison

	Mon 3/9	Tues 3/10	Wed 3/11	Thurs 3/12	Fri 3/13
Time	Users	Users	Users	Users	Users
8:00:00 AM	6,912	10,107	5,138	2,284	713
8:30:00 AM	20,748	25,630	15,332	7,584	2,058
9:00:00 AM	26,727	33,513	21,046	10,736	3,406
9:30:00 AM	29,572	37,438	22,312	11,166	3,300
10:00:00 AM	30,459	42,935	21,082	9,146	2,596
10:30:00 AM	25,823	38,600	16,797	6,193	1,929
11:00:00 AM	20,693	33,559	12,217	3,991	1,280
11:30:00 AM	15,128	25,664	8,574	2,859	916
12:00:00 PM	10,956	15,913	5,442	2,406	612
12:30:00 PM	8,803	10,946	4,940	2,509	502
1:00:00 PM	8,188	9,208	4,140	2,338	376
1:30:00 PM	6,790	8,401	3,701	1,917	295
2:00:00 PM	5,080	5,265	2,316	1,022	172
2:30:00 PM	3,246	3,315	1,339	534	108
3:00:00 PM	1,813	1,754	694	256	48
3:30:00 PM	1,241	906	284	159	40
4:00:00 PM	210	221	112	71	17
Max Concurrent	**30,499**	**43,297**	**22,592**	**11,432**	**3,469**

FSA ELA and Mathematics, FSA EOCs Concurrent Users Daily Comparison					
	Mon 4/13	Tues 4/14	Wed 4/15	Thur 4/16	Fri 4/17
Time	Users (Grades 3-10 R, 3-8 M)	Users (Grades 3-10 R, 3-8 M)	Users (Grades 3-10 R, 3-8 M)	Users (Grades 3-10 R, 3-8 M)	Users (Grades 3-10 R, 3-8 M)
8:00:00 AM	14,194	30,525	20,884	29,648	17,365
8:30:00 AM	41,397	60,886	50,368	60,217	35,688
9:00:00 AM	68,165	89,534	80,499	85,394	46,200
9:30:00 AM	86,775	111,676	102,050	107,822	61,611
10:00:00 AM	103,160	137,133	130,094	140,424	79,363
10:30:00 AM	105,403	129,343	125,590	133,233	77,781
11:00:00 AM	82,205	96,853	93,974	96,297	55,268
11:30:00 AM	58,085	62,222	58,862	56,808	31,857
12:00:00 PM	41,708	45,201	40,772	38,571	20,864
12:30:00 PM	38,923	40,731	38,943	38,666	21,542
1:00:00 PM	36,717	42,580	39,333	39,960	21,338
1:30:00 PM	34,281	40,728	39,512	39,470	21,008
2:00:00 PM	28,716	35,176	31,706	34,562	20,249
2:30:00 PM	22,119	29,268	23,307	26,650	16,307
3:00:00 PM	11,762	17,234	12,570	15,096	10,889
3:30:00 PM	4,697	7,042	4,932	6,727	4,032
4:00:00 PM	779	1,380	991	1,440	720
Max Concurrent	**108,392**	**140,092**	**134,086**	**144,716**	**82,140**

FSA ELA and Mathematics, FSA EOCs Concurrent Users Daily Comparison					
	Mon 4/20	Tues 4/21	Wed 4/22	Thur 4/23	Fri 4/24
Time	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)
8:00:00 AM	457	21,284	25,123	17,572	16,577
8:30:00 AM	253	58,087	57,345	44,796	37,006
9:00:00 AM	94	95,916	93,205	70,557	56,943
9:30:00 AM	116	131,894	131,636	102,839	81,497
10:00:00 AM	839	165,880	161,985	131,530	110,232
10:30:00 AM	8,741	158,543	145,842	121,221	98,789
11:00:00 AM	20,454	113,522	94,224	76,456	61,008
11:30:00 AM	30,536	70,413	56,057	48,422	41,282
12:00:00 PM	30,957	44,501	37,071	33,794	29,795
12:30:00 PM	28,824	38,393	32,920	29,824	24,220
1:00:00 PM	26,866	39,683	31,639	28,336	22,456
1:30:00 PM	28,072	40,359	31,591	29,435	22,654
2:00:00 PM	26,722	36,130	27,150	26,651	20,561
2:30:00 PM	21,349	28,933	21,881	20,637	16,283
3:00:00 PM	10,199	15,148	11,508	11,950	9,034
3:30:00 PM	3,270	6,670	4,457	5,033	3,673
4:00:00 PM	623	1,454	1,002	982	676
Max Concurrent	**31,901**	**170,132**	**161,985**	**134,710**	**111,426**

FSA ELA and Mathematics, Writing Makeup Concurrent Users Daily Comparison					
	Mon 4/27	Tues 4/28	Wed 4/29	Thur 4/30	Fri 5/1
Time	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)	Users (Grades 3-10 R, 3-8 M; EOC)
8:00:00 AM	23,093	32,837	22,022	28,277	18,995
8:30:00 AM	49,010	62,393	48,969	51,323	34,969
9:00:00 AM	66,227	83,158	69,965	66,100	45,023
9:30:00 AM	86,306	108,752	87,524	82,832	51,141
10:00:00 AM	110,448	141,244	110,581	109,008	66,567
10:30:00 AM	104,489	126,632	98,516	96,957	61,852
11:00:00 AM	64,741	81,061	59,601	62,650	42,377
11:30:00 AM	39,870	56,585	38,431	45,539	28,968
12:00:00 PM	31,977	45,979	29,347	34,663	21,892
12:30:00 PM	30.553	39,309	25,301	27,525	18,683
1:00:00 PM	26,523	35,126	23,154	24,954	16,301
1:30:00 PM	25,876	32,564	22,934	24,912	15,164
2:00:00 PM	21,404	28,035	20,027	24,403	15,163
2:30:00 PM	17,079	23,046	16,091	19,637	12,502
3:00:00 PM	9,413	14,683	10,124	12,433	7,918
3:30:00 PM	3,622	6,465	4,191	4,908	3,216
4:00:00 PM	583	1,600	933	1,052	707
Max Concurrent	**111,600**	**143,299**	**112,745**	**110,754**	**68,146**

FSA ELA and Mathematics, Writing Makeup Concurrent Users Daily Comparison					
	Mon 5/4	Tues 5/5	Wed 5/6	Thur 5/7	Fri 5/8
Time	Users (Grades 3-10 R, 3-8 M; 8-10 W; EOC)	Users (Grades 3-10 R, 3-8 M; 8-10 W; EOC)	Users (Grades 3-10 R, 3-8 M; 8-10 W; EOC)	Users (Grades 3-10 R, 3-8 M; 8-10 W; EOC)	Users (Grades 3-10 R, 3-8 M; 8-10 W; EOC)
8:00:00 AM	20,608	28,156	20,740	20,075	13,363
8:30:00 AM	42,133	52,958	42,667	37,657	22,041
9:00:00 AM	50,590	62,231	50,717	42,138	25,016
9:30:00 AM	53,633	63,004	50,695	42,511	23,984
10:00:00 AM	67,696	74,359	55,919	44,144	24,464
10:30:00 AM	63,488	67,943	48,804	38,848	21,301
11:00:00 AM	43,472	47,736	33,439	27,953	16,352
11:30:00 AM	29,290	33,448	23,711	19,112	12,109
12:00:00 PM	22,404	29,120	20,000	15,823	9,363
12:30:00 PM	18,659	25,466	17,511	13,846	8,510
1:00:00 PM	16,735	21,690	14,898	11,754	7,075
1:30:00 PM	16,030	18,557	13,337	10,033	5,539
2:00:00 PM	13,025	15,291	9,502	8,086	3,919
2:30:00 PM	9,309	11,283	7,385	6,181	2,898
3:00:00 PM	4,829	7,137	5,002	3,388	1,735
3:30:00 PM	2,386	3,586	2,312	1,476	868
4:00:00 PM	553	955	473	572	267
Max Concurrent	**69,665**	**75,023**	**56,244**	**44,518**	**25,328**

FSA ELA and Mathematics, Writing Makeup Concurrent Users Daily Comparison

	Mon 5/11	Tues 5/12	Wed 5/13	Thur 5/14	Fri 5/15
Time	**Users (Grades 3-10 R, 3-8 M; EOC)**	**Users (Grades 3-10 R, 3-8 M; EOC)**	**Users Algebra 1, Geometry, Algebra 2**	**Users Algebra 1, Geometry, Algebra 2**	**Users Algebra 1, Geometry, Algebra 2**
8:00:00 AM	13,007	17,886	8,072	6,489	3,218
8:30:00 AM	27,288	33,838	17,793	13,750	5,232
9:00:00 AM	32,235	40,294	22,190	16,112	5,909
9:30:00 AM	33,388	44,495	24,793	16,852	5,397
10:00:00 AM	39,121	55,900	30,596	18,396	5,072
10:30:00 AM	37,302	52,086	28,455	16,072	4,131
11:00:00 AM	30,808	42,791	22,531	12,667	3,047
11:30:00 AM	19,652	28,285	14,796	8,189	2,348
12:00:00 PM	14,534	19,771	9,946	6,083	1,902
12:30:00 PM	11,858	15,517	7,277	4,973	1,934
1:00:00 PM	8,847	12,260	5,884	3,768	1,648
1:30:00 PM	7,310	10,171	5,110	2,840	1,418
2:00:00 PM	4,874	6,993	3,215	1,965	925
2:30:00 PM	3,322	4,747	2,252	1,159	590
3:00:00 PM	1,243	1,988	1,092	531	226
3:30:00 PM	477	934	556	307	106
4:00:00 PM	120	269	190	121	39
Max Concurrent	**39,691**	**17,886**	**30,678**	**18,406**	**5,974**