Independent Verification of the Psychometric Validity for the Florida Standards Assessment

Executive Summary

August 31, 2015

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

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

and

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Table of Contents

Acknowledgments

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

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

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

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

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

Ellen Forte
Elizabeth Towles
Lori Nebelsick-Gullett
edCount, LLC

Executive Summary

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

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

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

Summary of the Evaluation Work

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

The process of validation refers not to a test or scores but rather to the uses of test scores. By reviewing a collection of evidence gathered throughout the development and implementation of a testing program, an evaluation can provide an indication of the degree to which the available evidence supports each intended use of test scores. As such, the evaluation of the FSA program began with the identification of the uses and purposes of the tests. Per legislation and as outlined within FLDOE's *Assessment Investigation* (2015) document, FSA scores will contribute to decisions

made regarding students, teachers, schools, districts, and the state. These uses across multiple levels of aggregation incorporate FSA data taken from a single year as well as measures of student growth from multiple years of data.

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

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

Evaluation of Test Items

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

Findings

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

Commendations

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

Recommendations:

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

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

Recommendation 1.2 Conduct an independent alignment study

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

Recommendation 1.3 The FLDOE should conduct a series of cognitive labs

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

Evaluation of Field Testing

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

Findings

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

Commendations

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

Recommendations

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

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

Evaluation of Test Blueprints and Construction

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

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

Findings

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

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

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

Given the timing of this review, FLDOE and AIR have yet to develop interpretation guides for the score reports. These guides typically explicate a deeper understanding of score

interpretation such as what content is assessed, what the scores represent, score precision, and intended uses of the scores.

Commendations

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

Recommendations

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

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

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

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

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

Evaluation of Test Administration

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

Findings

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

Commendations

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

Recommendations

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

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

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

The problematic spring 2015 FSA administration has made many individuals involved with the administration of the FSA to be extremely skeptical of its value. Given this problem, the FLDOE and its partners should engage in an extensive communication and training program

throughout the entire academic year to inform its constituents of the changes that have been made to help ensure a less troublesome administration in 2016.

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

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

Evaluation of Scaling, Equating, and Scoring

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

Findings

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

Commendations

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

Recommendation

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

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

Specific Psychometric Validity Questions

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

Findings

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

Commendations

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

Recommendation

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

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

Conclusions

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

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

Study-Specific Conclusions

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

Conclusion #1 – Evaluation of Test Items

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

Conclusion #2 – Evaluation of Field Testing

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

Conclusion #3 – Evaluation of Test Blueprint and Construction

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

Conclusion #4 – Evaluation of Test Administration

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

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

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

Conclusion #6 – Evaluation of Specific Psychometric Validity Questions

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

Cross-Study Conclusions

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

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

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

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

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

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

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