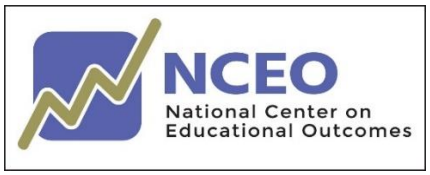
Feasibility of the Use of the ACT and SAT in Lieu of Florida Statewide Assessments

Volume 1: Final Report



Assessment Solutions Group

Ed Roeber
John Olson
Barry Topol

January 1, 2018

CITATION

This report was written by the Assessment Solutions Group with support from the Florida Department of Education

Citation: Roeber, E., Olson, J., & Topol, B. (2018). Feasibility of the Use of the ACT and SAT in Lieu of Florida Statewide Assessments: Volume 1: Final Report. Assessment Solutions Group.

The information in this report reflects the views of the authors. Publication of this document shall not be construed as endorsement of the views expressed in it by the Florida Department of Education

Table of Contents

Executive Summary

In the 2017 Florida legislative session, HB 7069 was passed and signed into law on June 15, 2017. The legislation states, in part:

> "The Commissioner of Education shall contract for an independent study to determine whether the SAT and ACT may be administered in lieu of the grade 10 statewide, standardized ELA assessment and the Algebra 1 end-of-course assessment for high school students consistent with federal requirements under 20 U.S.C. s. 6311(b)(2)(H). The commissioner shall submit a report containing the results of such review and any recommendations to the Governor, the President of the Senate, the Speaker of the House of Representatives, and the State Board of Education by January 1, 2018."

The Florida Department of Education (FDOE) issued RFP 2018-48 to solicit vendors to independently conduct studies, research, analyses, and Florida educator and expert meetings, and produce a final report to this effect. Assessment Solutions Group (ASG) and its team of subcontractors – Wisconsin Center for Education Products & Services (WCEPS), University of Minnesota's National Center on Educational Outcomes (NCEO), and University of Kansas's Center for Assessment and Accountability Research and Design (CAARD) – were selected to carry out the following studies:

1. Alignment – Evaluate the degree to which the ACT and the SAT align with Florida content standards and are, therefore, suitable for use in lieu of the grade 10 statewide, standardized English Language Arts (ELA) assessment and the Algebra 1 End-of-Course (EOC) assessment (Florida Standards Assessments, FSA).

2. Comparability – Conduct studies, research, and analyses required to determine the extent to which ACT and SAT test results provide comparable, valid, and reliable data on student achievement as compared to the Florida statewide assessments for all students and for each subgroup of students.

3. Accommodations – Determine whether ACT and SAT provide testing accommodations that permit students with disabilities and English learners the opportunity to participate in each assessment and receive comparable benefits.

4. Accountability – Conduct analyses to determine whether ACT and SAT provide unbiased, rational, and consistent differentiation among schools within the state's accountability system.

5. Peer Review – Conduct evaluations to determine whether the ACT and SAT meet the criteria for technical quality that all statewide assessments must meet for federal assessment peer review.

Results and Conclusions

1. Alignment (Criteria 1 and 2)
   a. Algebra 1 EOC – The analysis focused on the degree to which the assessments were aligned with the 45 Florida Algebra 1 content standards, which are a subset of the high school Mathematics Florida Standards (MAFS). Based on the results of the test forms analyzed, neither the SAT nor the ACT assessment is fully aligned to the Florida Algebra 1 standards. Both the ACT and SAT assessments would need to be augmented to assess the full breadth and depth of the Algebra 1 standards as called for by federal regulations. The analysis indicated that the ACT would need slight adjustment to attain the minimum level of full alignment: seven or eight items would need to be added to each ACT test form. The SAT test forms were found to have conditional alignment, depending on the test form: One test form was found to be acceptably aligned, needing four items added; and the other SAT test form was found to need slight adjustment, needing seven items added to the form to attain the minimum level of full alignment according to the criteria used in this study.

   b. Grade 10 English Language Arts – The results show that the ACT would need major adjustments—needing 10 or more items revised or replaced to be fully aligned with the Florida Grade 10 LAFS. The SAT test forms were found to have conditional alignment, depending on the test form: One SAT test form was found to need five items revised or replaced for full alignment while the other SAT test form was found to need slight adjustments—seven items revised or replaced—for full alignment.

      While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, augmentation adds cost and complexity to the administration of the tests, since items used to augment a test need to be developed annually and administered separately from the college entrance test. Without such augmentation, the ACT and possibly SAT tests might not meet the United States Department of Education (USED) peer review criteria for aligned tests, thus jeopardizing the federal approval of Florida's plan to offer choice of high school tests to its school districts.

2. Comparability (Criterion 3)

There were two significant concerns with the dataset available for analysis (students who took both the FSA and either the ACT or SAT) in the comparability studies.

a. First, only about half of Florida's tenth or eleventh grade students take either the ACT or the SAT before they graduate from high school, meaning that the matched samples of students who participated in the FSA assessments and took one or both of the college entrance tests seriously underrepresent the full sample of FSA test-takers (presumably the lower-scoring students that only took the FSAs).

b. Second, the data provided by the FDOE indicates that 83%of students took the Algebra 1 EOC in eighth or ninth grades, one to three years before they took the ACT or SAT (in spring of tenth grade, fall of eleventh grade, and/or spring of eleventh grade). Large

distances between time of testing in the two tests increases measurement error as learning likely occurred between those test administration times.

In short, the data provided for this study are neither representative of the full population of students nor were the tests taken close enough in time to assume little to no learning occurred. Several different statistical analyses were run to try to account for the data issues and determine if the tests are comparable. An important analysis performed was how often students would be placed at the same level of performance based on their scores on the three tests. The results of this classification consistency analysis indicate that many students would be placed at different performance levels on the three tests, some by as much as four out of the five performance levels. Thus, districts using the FSA option may have very different results than districts using either the ACT or SAT options. This casts serious doubt on the interchangeability of the three tests, and the soundness of making accountability decisions based on them. At this point, it appears the ACT and SAT do not produce results comparable to the FSA and should not be considered alternatives to them. This also indicates that the ACT and SAT will, likely, not meet USED peer review requirements.

3. Accommodations (Criterion 4)

This study concluded that in many ways, in terms of the provision of accommodations, the ACT and SAT could provide comparable benefit to the FSA for purposes of school accountability and graduation, although this was less evident on the SAT for English Learners (ELs). In general, both the ACT and the College Board indicated that they would provide greater numbers of the accommodations in the standard list of accommodations used in this study (previously developed by NCEO) than were provided for the FSAs. Whether these differences were appropriate for the Florida standards was not addressed in these studies.

Comparability in the process for accommodations requests was less clear and often relevant more to the use of the tests for college entrance; comparability to the FSA cannot be judged here because the FSA does not provide a score that can be used for college entrance. Still, if a district based a decision to use one of these tests in lieu of the Florida assessments on the possibility of having college entrance scores for all its students, this goal is unlikely to be realized for some students with disabilities and ELs.

The lack of transparency in the decision-making process about which specific accommodations would result in a college reportable score for which specific students is likely to result in non-comparability for some student groups compared to other student groups, which could be a concern when making the decision about whether to allow Florida districts to use either the ACT or the SAT in lieu of the Florida assessments.

4. Accountability (Criterion 5)

Using the sample of students with two years of data from the FSAs and an ACT or SAT score, simulated schools were created to examine the effects of calculating school-level indicators using the different tests.

Overall, differences are shown across all three indicators. The results show that the numbers going into the accountability determination would differ for many schools by the test selected.

Richer calculations can be done for ELA as there exists a state test for grades 8, 9, and 10. For mathematics, the time at which the Algebra 1 EOC test is taken varies by students and for many, there is no prior year score with which to base a growth calculation.

There are two important findings to consider from this accountability study: one data-based and one more theoretical. First, the differences shown for ELA vary by type of school. Larger schools with a greater number of lower performing students are advantaged by using the alternate tests (ACT/SAT). This finding has implications for policy, as districts could use these results to select a test, rather than making a more holistic determination about its students and what test best fits the population.

Second, there will often be very different students being compared in the growth models. For example, in mathematics, the learning gain using the FSA will be calculated based on grade 8 math and grade 9 Algebra 1. However, using the alternate test, a similar high school would be evaluated based on the learning gain between Algebra 1 EOC in grade 10 and the ACT or SAT in grade 11. Likewise, for the value added model, only schools using the FSAs will have a VAM score for ELA and only some of those for mathematics. With the elimination of the FSA Grade 10 ELA and Algebra 1 EOC tests, two years of prior data will not exist for students taking the ACT or SAT in grade 11.

Both of these findings indicate that the answer to the question on fairness is: "no – it is not fair to compare schools that use the state tests in their accountability system to those that use the alternate tests."

5. Peer Review (Criterion 6)

To test the acceptability of Florida's plan to offer its schools the option of using the ACT or SAT in lieu of the FSAs, ASG conducted a mock peer review, using evidence provided by ACT, the College Board, and the Florida Department of Education, as well as from ASG's studies of alignment, comparability, and accommodations. Experienced peer reviewers examined the evidence and prepared written notes similar to an actual peer review. A summary of the peer review results is shown in Table 1.

Table 1. Peer Review Critical Element Determinations by College Entrance Test

Peer Review Determinations	ACT	College Board
Met Mock Peer Review Requirements	23	20
May Not Meet Mock Peer Review Requirements	1	6
Did Not Meet Mock Peer Review Requirements	6	3
TOTAL	30	29*

*One Critical Element – related to online assessment – is not applicable to the current College Board SAT.

As can be seen, the ACT was judged to not meet 6 of 30 Critical Elements, and possibly not meet the peer review requirement for one additional element. The SAT was judged to not meet 3 of 29 peer review Critical Elements, and possibly not meet the peer review requirements for six additional elements.

Overall Conclusion

It is the opinion of ASG and its partners that due to the alignment, comparability, and accountability system issues associated with the ACT and SAT tests, allowing districts to pick which of the three tests to administer to its students is not appropriate and likely will not meet federal ESSA peer review requirements.

Detailed results from the five studies and a summary of the findings are provided in the following sections of the report and in the Summary and Conclusions chapter at the end. Details of the data and analyses used in the studies are provided in a separate document "Volume II – Appendices."

Background and Purpose of the Project

In the 2017 Florida legislative session, HB 7069 was passed and signed into law on June 15, 2017. The legislation states, in part:

> "The Commissioner of Education shall contract for an independent study to determine whether the SAT and ACT may be administered in lieu of the grade 10 statewide, standardized ELA assessment and the Algebra 1 EOC assessment for high school students consistent with federal requirements under 20 U.S.C. s. 6311(b)(2)(H). The Commissioner shall submit a report containing the results of such review and any recommendations to the Governor, the President of the Senate, the Speaker of the House of Representatives, and the State Board of Education by January 1, 2018."

The federal law referenced above is commonly known as the Every Student Succeeds Act (ESSA). The law provides flexibility for a state to approve a school district to administer, in lieu of the statewide high school assessment, a "locally selected," "nationally recognized" high school academic assessment that has been approved for use by the state, including submission for the U.S. Department of Education's (USED) assessment peer review process. At a minimum, ESSA requires that the state must determine that an assessment used for this purpose meets the following criteria:

1. Is aligned to and addresses the breadth and depth of the State's content standards;
2. Is equivalent in its content coverage, difficulty, and quality to the statewide assessments;
3. Provides comparable, valid, and reliable data on student achievement as compared to the Florida Standards Assessments (FSA), the statewide assessments used with all students and each subgroup of students. Final ESSA assessment regulations of December 8, 2016, clarify that comparability between a locally selected, nationally recognized high school academic assessment and the statewide assessment is expected at each academic achievement level;
4. Provides accommodations that permit students with disabilities and English learners the opportunity to participate in the assessment and receive comparable benefits;
5. Provides unbiased, rational, and consistent differentiation among schools within the state's accountability system; and
6. Meets the criteria for technical quality that all statewide assessments must meet (i.e., those specified by USED's assessment peer review).

The purpose of the project was to conduct a study to determine whether the state of Florida can allow districts to choose to offer its students the FSAs, ACT, and SAT and still have an assessment system that meets the above criteria.

ASG Partners and Plan of Study

In September 2017, Assessment Solutions Group (ASG) and its subcontractors were awarded the contract under Florida RFP 2018-48: *Feasibility of Use of ACT and SAT in Lieu of Statewide Assessments*. The ASG Team consisted of some of the most recognized names in their respective fields of statewide assessment and their expertise aligned perfectly with the six criteria that Florida outlined in the RFP. The ASG team and their assigned areas of study were as follows. The ASG individual assigned to coordinate the work is shown in parentheses for Criteria 1-5.

Criteria	Responsible Organization/(ASG Support Person)
1 and 2 – Alignment	WCEPS – Norman Webb/Sara Christopherson (Olson)
3 – Comparability	CAARD – Marianne Perie (Roeber, Olson)
4 – Accommodations	NCEO – Sheryl Lazarus/Martha Thurlow (Roeber)
5 – Accountability	CAARD – Marianne Perie (Olson)
6 – Peer Review	ASG – Ed Roeber/John Olson (Thurlow, Lazarus, Perie, Webb, and Christopherson)
ASG Team Members	John Olson Edward Roeber Barry Topol

The overall work by ASG and its partners was performed collaboratively with multiple groups working together to evaluate and respond to each of the criteria outlined in the RFP. There was a lead group and ASG support for each criterion, thereby allowing for multiple points of review and expertise to be brought to each area. The overall project direction and management was provided by ASG's Barry Topol, with technical support provided by ASG's John Olson and Ed Roeber.

The plan of study for each criterion is outlined below.

Criteria 1 and 2 – Alignment (Wisconsin Center for Education Products and Services)
The proposed methodology for the alignment analysis was based on processes developed and refined by Norman Webb over the past 20 years. These processes have been used to analyze curriculum standards and assessments in around 30 states to satisfy or to prepare to satisfy the Title I compliance as required by the United States Department of Education (USED). Many states that have met the USED requirements used this process to evaluate the alignment of their standards and assessments. The alignment analysis conducted in this study was designed to answer two key research questions:

1. How does content coverage in the ACT and SAT, for both mathematics and English language arts (ELA), compare with the content coverage of the current Florida grade 10 statewide standardized ELA assessment and the Algebra 1 EOC assessment for students?
2. What is the degree of alignment of the ACT and the SAT with the high school Language Arts Florida Standards (LAFS) and the mathematics Florida Standards (MAFS) with regards to the satisfying the federal requirements within the Every Student Succeeds Act (ESSA)?

The proposed alignment analyses involved two stages:

- **Stage I:** An analysis of assessment framework documents
- **Stage II:** An in-person content alignment institute

The final deliverables for the work outlined were two alignment reports, one for mathematics and one for ELA. The two reports provided findings from a reliable and replicable process using an item-level analysis. Results described the overlap in content targeted by Florida's FSAs and the ACT and SAT, as well as any content unique to each assessment, considering refined content topics and levels of content complexity. In addition, the results described the degree of alignment of the ACT and the SAT with the Florida standards for both language arts and mathematics.

The two final reports on alignment included a comparison of the blueprints and item specifications, or equivalent documents, for each of the ACT, SAT, and Florida assessments to provide more in-depth description of the content coverage by each assessment.

Criterion 3 – Comparability (Center for Assessment and Accountability Research and Design)
The main goal of the comparability analyses was to determine if the technical characteristics of the different measures under consideration (ACT and SAT; FSAs in Grade 10 ELA and Algebra 1 EOC) would provide comparable, valid, and reliable data on student achievement, for all students and each subgroup of students, to permit them to be used in lieu of the current FSAs in Florida's assessment system.

In carrying out this comparability work, the following research questions and associated analyses were addressed:

1. How similar are item types across the three tests? Item types were evaluated to determine whether the tests were comparable in terms of the types of items used.
2. How similar are the ranges of difficulty across the three tests at the item level? Average and range of item difficulties of the tests were examined to see if they were significantly different from one another.
3. How similar are the reliabilities across the three tests? Technical information for each assessment (e.g., standard errors of measurement and internal consistency) and the overall reliability of each measure were reviewed and compared.
4. When looking at matched samples, how similar are the distributions of performance across the tests? Comparisons were made of the distribution of test takers by county school system with comparable demographic information about student enrollments and, where necessary, sub-samples of students to best represent the entire state were selected. Once

matched/representative samples were obtained (adjusted if necessary for student sub-group underrepresentation), the performances of students on the different tests were examined.
5. What percentage of students can be expected to be categorized into different achievement levels with each test (and in which direction)? Comparisons were completed on whether the same percentages of students perform in different achievement levels on each test.
6. What is the probability that a student would be placed in the same Florida category of achievement when taking the FL test, the ACT, or the SAT? Analyses were completed to evaluate the comparability of whether every student would be placed in the same category of achievement on the ACT or the SAT as they are on the Florida tests.

To complete these studies, data were collected from recent ACT and College Board Technical Reports, test files, matched datafiles from Florida, other state data (those using ACT or SAT), and state reports of student performance and achievement level results. The intent was to build a body of evidence to be used to evaluate the degree of comparability of the tests.

Criterion 4 – Accommodations (National Center on Education Outcomes)
The National Center on Educational Outcomes (NCEO) organized and conducted in-person studies with Florida educators to evaluate the degree to which the ACT and SAT provide testing accommodations that permit students with disabilities and English learners (ELs) the opportunity to participate in each assessment and receive comparable benefits to participation in the FSA.

NCEO staff worked with the FDOE to identify 8 panelists who had familiarity with the Florida mathematics standards and the ways in which the FSA Algebra 1 EOC assessment is administered to students with disabilities and ELs. Similarly, NCEO also worked with the FDOE to identify 8 panelists who had familiarity with the Florida language arts standards and the ways in which the FSA grade 10 ELA assessment is administered to students with disabilities and ELs.

Each panel included a special educator (Exceptional Student Education); English language learner educator (English for Speakers of Other Languages); blind/low-vision educator; deaf/hard of hearing educator; and content educator. Each group of panelists went through a systematic study process during the in-person meetings and reviewed accommodations for the ACT, SAT, and FSAs with respect to:

1. The process and ease of signing up for accommodations
2. The availability of the accommodations themselves
3. The testing context for the accommodations provided to reach its conclusion about the suitability of the ACT or SAT to replace the current Florida assessments of grade 10 ELA and Algebra 1.

Criterion 5 – Accountability (Center for Assessment and Accountability Research and Design)
The accountability simulations called for in Criterion 5 were in many ways the "bottom line" of establishing the comparability of the college entrance tests with the FSAs in ELA and Algebra 1 EOC because of the consequences for schools and districts for the performance of their

secondary students. Florida includes ELA, math, science, and social studies scores in its school accountability system. In this study, the ACT or SAT scores substituted for the Grade 10 FSA ELA scores or the FSA Algebra 1 EOC scores. Other achievement scores and all other variables included in the Florida accountability system remained the same. In Florida's A-F accountability rating system, the ACT or SAT scores were included in the two indicators for high schools – Achievement and Learning Gains.

Using an adjusted matched sample, CAARD replicated the procedures used by Florida to hold schools accountable for the performance of their students in Florida's school accountability system. This included the substitution of the ACT- or the SAT-derived student scores for the FSA in ELA or the Algebra 1 EOC scores.

A series of simulations were conducted using adjusted matched samples of FSA and ACT or SAT student files. Results from the simulations were then carefully evaluated and the findings from using ACT and SAT were compared to those for FSA ELA and Algebra 1 EOC tests that are currently used in the accountability system reports for Florida schools. These simulations determined whether ACT and SAT provide unbiased, rational, and consistent differentiation among schools in Florida's accountability system.

Criterion 6 – Peer Review (Assessment Solutions Group)
The ASG team's approach under Criterion 6 included the following steps:

1. Summarizing current information and evidence from states that have used the ACT and the SAT as their high school assessments on how they have addressed the requirements in the USED peer review, in particular those Critical Elements related to alignment, test development, accommodations, technical quality, and validity. When this research had concluded, a determination was made that information for the College Board SAT was not yet available, while information for the ACT preceded ESSA and its peer review elements.
2. Creating a hybrid peer review template for use by the FDOE, ACT, and the College Board to submit their evidence of adequately addressing each peer review critical element.
3. Reviewing and commenting on each of the pieces of evidence submitted by ACT and the College Board in support of the use of these assessments in lieu of the FSAs.
4. Providing a professional judgment on the likelihood of ACT and/or SAT, when used as *an optional high school test* in place of the state's test, being approved by USED following peer review. This included providing comments on the strengths and weaknesses of the evidence that was provided and recommendations on the areas where improvements or additional evidence may be needed. Note: Not all evidence for peer review will be provided by ACT or the College Board. There are peer review critical elements for which evidence in support of the use of the ACT or the SAT will come from school districts after the initial administration of these assessments in Florida's districts. Collecting this "local-use" data will add to the complexity of FDOE's ultimate peer review submission to the USED.
5. Preparing the relevant parts of the actual peer review document for submission for the Department.

ASG and its partners used a "peer review-like" process to accomplish the purposes outlined above. The ASG team gathered evidence from the FDOE, ACT, and the College Board,

including the draft of the pertinent sections of the actual *State Assessment Peer Review Submission Cover Sheet and Index Template,* the document that each state uses to submit its evidence of the technical qualities of its proposed assessments. The peer review evidence compilation was split into two tracks: 1) the technical criteria for the Critical Elements responded to by ACT or the College Board, and 2) evidence related to supporting these Critical Elements from the work that ASG and its partners carried out in Criteria 1-5.

Sections outlining the analyses conducted and conclusions reached for each of the six criteria appear in the following sections of the report. Appendices providing detailed data and other information appear in a separate volume of the report.

Section 1
Alignment Studies (Criteria 1 and 2)

1A – Math Alignment Studies
Executive Summary

An alignment analysis was conducted as part of a comprehensive study to determine if Florida school districts might be able to use a college entrance test (the ACT or the SAT) in place of the Florida Standards Assessments (FSA): Florida's Grade 10 Statewide Standardized ELA Assessment and Algebra 1 end-of-course (EOC) exam. The larger study encompasses the alignment of the three tests with Florida's academic content standards, as well as an examination of the accommodations offered to students with disabilities and English learners, the statistical comparability of the measures, and potential impacts of using all three tests interchangeably on school accountability in Florida. Together this study is designed to reveal the degree to which the ACT or SAT could be used in lieu of the Florida Grade 10 Statewide Standardized ELA assessment and Algebra 1 EOC assessment in fulfilling requirements as stated in Federal statute. A separate report has been prepared to describe the alignment of the ELA assessments of each of the three tests.

A two-part alignment study was conducted as one of a concert of investigations to answer this question. The first stage of the alignment study compared the differences and similarities in the frameworks used to develop or interpret the findings from the three assessments. The framework analysis was conducted by a mathematics content expert, Professor Kristen Bieda, of Michigan State University. The second stage of the study was a two-day alignment institute, October 18-19, 2017, that was conducted in Orlando, Florida. Seven reviewers conducted the analysis, five of whom were from Florida, and invited to participate from a list provided by the Florida Department of Education, and two of whom were external reviewers from other states. All of these reviewers had backgrounds in teaching high school mathematics or serving as a mathematics coordinator. The project director and an additional reviewer, both with mathematics education backgrounds, coded some of the forms. This was done to have at least a total of five reviewers that coded each assessment form.

The analysis focused on the degree to which the assessments were aligned with the 45 Florida Algebra 1 standards, a subset of the high school Mathematics Florida Standards (MAFS). For use in the alignment institute, these standards were supplemented by additional ones, informed by the framework analysis, in order to be able to describe in more detail the content targeted by the ACT and SAT. The seven mathematics reviewers were trained in the alignment process at the institute. The reviewers entered their data into the Web Alignment Tool version 2 (WATv2).

The degree of alignment of a test form with the corresponding standards can be considered in terms of the degree to which specific alignment criteria are met as well as in terms of the total number of items, if any, that would need revision or replacement for full alignment. In terms of meeting the specific alignment criteria, both of the Florida test forms analyzed met all of the alignment criteria for all three reporting categories with one exception: both test forms only weakly met the criterion of Range of Knowledge (breadth) for one of the three reporting categories (RC3 Statistics & the Number System). The ACT test forms did not have items that

corresponded to a sufficient number of standards for any of the three of the reporting categories to be considered to have an acceptable breadth in coverage of the Algebra 1 standards. Breadth, as measured by the Range-of-Knowledge alignment criterion, was unmet for both ACT test forms for two reporting categories (RC1: Algebra & Modeling and RC3: Statistics and the Number System) and was only weakly met for the third reporting category (RC2: Functions and Modeling). The SAT test forms were found to not have items that corresponded to a sufficient number of standards to address the breadth of expectations within RC2 or RC3.

In terms of the number of items that would need revision or replacement for full alignment, both Florida test forms were found to be acceptably aligned—defined as needing 5 or fewer items revised or replaced. One Florida test form was found to need only one item revised or replaced and the other test form was found to need two items revised or replaced to meet the minimum cutoffs for full alignment. One SAT test form was also found to be acceptably aligned, needing four items added to meet the minimum cutoffs for full alignment. The second SAT test form was found to need slight adjustments—defined as needing six to 10 items revised or replaced to meet the minimum cutoffs for full alignment. That second SAT test form needed seven items revised or replaced to meet the minimum cutoffs for full alignment with the Florida Algebra 1 standards. Thus, alignment of the SAT was found to depend on the test form. The analysis indicated that about seven or eight items would need to be added to the ACT to meet the minimum cutoffs for full alignment according to the criteria used in this study.

About one-third of the ACT items and two-thirds of the SAT items corresponded to the 45 Florida Algebra 1 standards. The ACT had items that corresponded to a greater number of standards overall, including geometry and grades 4-8 standards. The SAT had items that corresponded to these topics as well, but in fewer numbers. The measures of agreement in assigning depth-of-knowledge levels to assessment items and items to curriculum standards were all in an acceptable range.

Whereas both Florida assessment forms were found to be acceptably aligned with the Algebra 1 standards, both ACT test forms were found to need some adjustments. One of the SAT test forms was found to be acceptably aligned while the other test form was found to need slight adjustments. Both the ACT and SAT would need to be augmented with additional items to meet the minimum cutoffs for full alignment with the Florida Algebra 1 standards. While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process and adds complexity to the administration of the tests, since items used to augment a test need to be administered separately from the college entrance test. Without such augmentation, however, these tests might not be viewed as meeting the United States Education Department (USED) criteria for aligned tests, thus jeopardizing the college entrance tests' approval in the federal standards and assessment peer review process.

Introduction and Methodology

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments

are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute solely of either of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). The corresponding methodology used to evaluate alignment has been refined and improved over the last 20 years, yielding a flexible, effective, and efficient analytical approach.

This is a report of a two-stage alignment analysis in the area of mathematics that was conducted during the month of October, 2017, to provide information that could be used to judge the degree that the ACT or SAT meet the Criteria 1 and 2 (related to alignment, from Florida RFP 2018-48) for their suitability to be administered in lieu of Florida's Algebra 1 end-of-course assessment, consistent with federal requirements under 20 U.S.C.s. 6311(b)(2)(H). More specifically, this study addressed the question of alignment between the ACT or SAT with the Mathematics Florida Standards (MAFS) used to develop the Algebra 1 EOC assessments administered in the spring of 2016 and 2017. As such, the study focused on the degree that the assessments, including the current Florida Algebra 1 EOC, addressed the full depth and breadth of the standards used to develop the Florida Algebra 1 EOC assessment. This alignment analysis is one of a concert of studies conducted in response to the Florida RFP 2018-48 requesting proposals by August 15, 2017. A parallel alignment study was done for the ELA assessments (described in a separate report).

The alignment analysis consisted of two stages:

- **Stage I:** An analysis of assessment framework documents; and
- **Stage II:** An in-person content alignment institute.

The Stage I framework analysis was done by mathematics education Professor Kristen Bieda, of Michigan State University. Dr. Bieda analyzed the specification of mathematics content in supporting documents for each of the three assessments including blueprints, item specifications, item type, calculator policy, and other relevant materials that were used in developing tests or interpreting scores. Her report is included as an attachment to this report (see **Appendix 1a.E**). Information from her report was used to increase the number of the MAFS included in Stage II. Although the charge for the alignment analysis was restricted to the Mathematics Florida Algebra 1 standards, these standards were supplemented with additional MAFS, including those in grades 4 through high school, in order to describe in more detail the content assessed by the ACT and the SAT.

The Florida Standards are a modified version of the Common Core State Standards (CCSS). The Common Core State Standards were developed in 2010 through the coordination of the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO). The standards were designed to provide a clear and

consistent framework to prepare pre-K through grade 12 students for college and the workforce. The standards were written to describe the knowledge and skills students should have within their K-12 education careers so that high school graduates will be able to succeed in entry-level, credit-bearing academic college courses and in workforce training programs. The CCSS have been widely used by over half of the states in the country to prepare students for college and careers. The MAFS are nearly identical to the CCSS for mathematics and can be considered as meeting the requirement of high quality standards related to college and career readiness.

This study included the 45 standards identified by Florida that defined the expectations for the Algebra 1 course. In addition, another 124 standards were added to the 45 Algebra 1 standards to be able to have standards that would correspond to items that may be on the ACT or SAT assessments. In particular, standards related to the topics of geometry, trigonometry, statistics, data, and proportions were included. The eight mathematical practices standards were not used by Florida for the Algebra 1 course and were not included in this study.

The 45 Algebra 1 standards were grouped under three reporting categories—Algebra and Modeling (N=17); Functions and Modeling (N=15); and Statistics and the Number System (N=13). Under the reporting categories, the standards were grouped by domain. For this analysis, 65 additional standards from the MAFS were added to the three Algebra 1 reporting categories along with standards grouped under two reporting categories—Geometry (N=43) and Grades 4-8 Mathematics Standards (N=16). These additional standards were included in the study to be able to better reflect the content included in the SAT and ACT assessments. The framework analysis provided information that suggested that content from pre-high school courses could appear on the assessments. These topics from grades 4-8 could be possible predictors for college and career performance.

Stage II of the study, an in-person content alignment institute for English Language Arts (ELA) and Algebra 1, was held over three days, October 18-20, in Orlando, Florida, at the Hyatt Place Orlando/Buena Vista. Both ELA and Mathematics assessments were reviewed at the institute. The content groups worked separately. The mathematics panel worked for two days, October 18 and 19. Seven reviewers served on the mathematics panel. The group leader, a retired mathematics curriculum coordinator from Pittsburg, Pennsylvania, had served as a leader and reviewer in numerous other alignment studies. A second external reviewer was a state mathematics assessment coordinator who had participated in one other alignment study. Five Florida Algebra 1 or mathematics coordinators participated as reviewers, invited from a list of highly qualified educators provided by the Florida Department of Education. In addition, Norman Webb (study director), whose background is in mathematics education and has participated in a multitude of alignment studies as far back as 1996, coded four of the six assessments. Webb also served as study director for this project. After the institute, a ninth reviewer, a retired district mathematics coordinator who has participated in numerous alignment studies, coded two assessments in order to have at least five reviewers analyze each of the six assessments. A total of five to eight reviewers coded each assessment.

Study director Norman Webb is the researcher who developed the alignment study procedures and criteria (through the National Institute for Science Education in 1997, funded by the

National Science Foundation, and in cooperation with the Council of Chief State School Officers) that influenced the specification of alignment criteria by the U.S. Department of Education. The Webb alignment process has been used to analyze curriculum standards and assessments in at least 30 states to satisfy or to prepare to satisfy the Title I compliance as required by the United States Department of Education (USED). Study Technical Director Sara Christopherson has participated in and led Webb alignment studies since 2005, for over 20 states as well as for other entities.

The Version 2 of the Web Alignment Tool (WATv2) was used to enter all of the content analysis codes during the institute. The WATv2 is a web-based tool connected to the server at the Wisconsin Center for Education Research (WCER) at the University of Wisconsin-Madison. It was designed to be used with the Webb process for analyzing the alignment between assessments and standards. Prior to the Institute, a group number was set up on the WATv2 for each of the two panels. Each panel was assigned one or more group identification numbers and the group leader was designated. Then the reporting categories and standards were entered into the WATv2 along with the information for each assessment, including the number of items, the weight (point value) given to each item, and additional comments such as the identification number for the item to help panelists find the correct item.

Training and Coding

In the morning of the first day of the alignment institute, reviewers in both the English Language Arts (ELA) group and the mathematics group received an overview of the purpose of their work, the coding process, and general training on the Depth-of-Knowledge (DOK) definitions used to describe content complexity. All reviewers had some understanding of the DOK levels prior to the institute. The general training at the alignment institute was crafted to contextualize the origins of DOK (to inform alignment studies of standards and assessments) and purpose (to differentiate between and among degrees of complexity), and to highlight common misinterpretations and misconceptions in order to help reviewers better understand and, therefore, consistently apply the depth of knowledge (DOK) language system. Panelists also practiced assigning DOK to sample assessment items that were selected to foster important discussions that promote improved conceptual understanding of DOK. Appropriate training of the panelists at the alignment institute is critical to the success of the project. A necessary outcome of training is for panelists to have a common, calibrated understanding of the DOK language system for describing categories of complexity.

Following the general training, the two groups went to separate rooms to receive more detailed training on the DOK levels for each content area. For mathematics, the group discussed the definitions for the four DOK levels for each content area. After the mathematics reviewers attained a common understanding of the DOK definitions, they reviewed the DOK levels assigned to the MAFS given to them. They were asked to identify any of the assigned DOK levels they thought did not accurately depict the appropriate level of content complexity. The group then discussed any standard identified by one or more of the reviewers. The group decided to change the DOK level of two standards (G-SRT 4.9 from DOK 3 to 2 and G-SRT 4.10 from DOK 2 to 3). The mathematics group then coded the first five assessment items from the Florida Algebra 1 EOC Spring 2016 assessment form. This was done to monitor that all the reviewers understood the process and to check on their coding of items to standards. Then the

reviewers coded the remainder of the items on the Florida Algebra 1 EOC Spring 2016 form independently.

In coding an assessment, reviewers were instructed to read the assessment item and to respond to the question. Then they were told to determine and enter the DOK level of the item into the WATv2 before deciding the matching standard. Next reviewers were to find the curriculum standard from the 169 standards they were given that best represented the content knowledge that was necessary for someone to know in order to answer the item correctly. If the reviewer felt that the knowledge required to answer an item correctly corresponded to two distinct standards, then they were to identify one or two additional standards. However, they were cautioned to use additional standards only when an item truly targeted multiple standards because doing so increased the weighting for that item.

Reviewers were instructed to consider the full statement of expectations to consider if an assessment item should be mapped to a standard. For a reviewer to code an item to a standard, all or nearly all, of the expected outcome as expressed in the standard had to be necessary for a student to perform to answer the item correctly. In some cases, reviewers could make reasonable arguments for a coding an item to different standards. If reviewers map an item to a variety of standards it may also indicate that the assessment task may be inferred to relate to more than one standard but that the item is not a close match.

Reviewers may have difficulty finding where an item best fits when an assessment is coded to a set of standards that were not used in developing the assessment. If an item did not closely fit any standard, then the reviewers were instructed to code the item to a standard where there was a partial fit or to a generic standard (domain or reporting category level). If the item did not match any of these, then the reviewer was instructed to indicate that the item was uncodeable. No items were considered uncodable on any of the test forms in this review.

If reviewers did not find a standard that explicitly matched an assessment item, they were instructed to code the item to a generic standard. A generic standard is the next level, either the domain or the reporting category. The supplementary standards to the Algebra 1 standards were added to reduce the number of items that would be assigned to a generic standard.

Reviewers were instructed to enter a note into the WATv2 for an assessment item to provide additional and helpful information about the item and the corresponding standards. For example, if the item only corresponded to a part of a standard and not the full standard, reviewers were requested to enter the letter indicating what part of a standard was targeted. Thus, the reviewers' notes reveal if assessment items only targeted a part of the individual standards (see **Appendix 1a.C**). Reviewers also could indicate whether there was a Source-of-Challenge issue with an item—i.e., a problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly. After finishing coding of all of the items on an assessment, reviewers were asked to respond to four debriefing questions. These questions sought additional information from the reviewers about their holistic view of the assessment, including qualitative feedback that was not captured in their standards codings, DOK codings, or earlier notes.

Reviewers' codings entered into the WATv2 were monitored by the project director as reviewers were entering the data. This was done to identify any potential problems in data entry. Once all the reviewers had completed entering data for an assessment—a DOK and standard for each assessment item and a response to debriefing questions—the director then identified what items should be adjudicated. The study director and group leader noted the assessment items that did not have a majority of reviewers in agreement on standard assignment or where the reviewers differed significantly on the DOK assigned to an item (e.g., three different DOK values were assigned). When these extreme disagreements occur, it suggests that reviewers are either interpreting the DOK definitions in very different ways or are interpreting the particular assessment item in very different ways. The WATv2 produces tables that show the standards assigned to an item by all of the reviewers along with a table of the DOK levels to help identify variation in coding among reviewers.

After discussing an item, the reviewers were given the option to make changes to their codings, but were not required to make any if they thought their coding was appropriate. If an item did not closely fit any standard, then the reviewers were instructed to code the item to a standard where there was a partial fit or to a generic standard (domain level or reporting category). For some items, reviewers could make reasonable arguments for coding an item to different standards. This was particularly the case when an assessment was coded to a set of standards that were not used to develop that assessment. In these situations, an item may measure a general part of more than one standard, but not the more specific details that distinguish the two standards. For example, two Florida standards both address quadratic equations: F.IF.2.4 expects students to interpret key features of graphs and tables of a quadratic equation such as the x-intercept while A.REI.2.4 expects students to solve a quadratic equation by a number of methods including factoring or identifying the x-intercepts. An item that requires students to identify the x solutions from a graph of a quadratic equation could be coded to either of these standards. It is likely such an item would not appear on an assessment that was explicitly written to target F.IF.2.4 and A.REI.2.4.

Reviewers completed the coding of one form of the Florida Algebra 1 assessment late in the afternoon on the first day. All of the reviewers then began coding the ACT Form 74H. The coding of this form and the adjudication process were completed by midmorning of the second day. At this time, reviewers were divided into smaller groups. This was done to allocate the coding of the remaining four assessments so that at least some reviewers will have coded each assessment. Three reviewers coded the second form of the Florida assessment (Spring 2017), two coded the first SAT test (April 2017), and two coded the second ACT form (74C). All but one reviewer then coded a fourth assessment. By the end of the two days allotted for coding, eight reviewers had coded the Florida Algebra 1 EOC (Spring 2016), three reviewers had coded the Florida Algebra 1 EOC (Spring 2017), seven reviewers had coded the ACT Form 74H, four reviewers had coded the ACT Form 74C, four reviewers had coded the SAT April 2017 form, and three had coded the SAT May 2017 form.

The two Florida Algebra 1 EOC assessment forms were viewed via a secure online browser on a separate computer than the one that reviewers used to enter data into the WATv2. The online interface required reviewers to move sequentially through the items and did not allow

reviewers to jump back and forth to check or compare items and codings. Many reviewers found they needed to record their codings on a piece of paper, and then transfer these codings into the WATv2. Consequently, the mathematics assessment review process took nearly double the usual time to analyze an assessment of around 60 items (six hours rather than the planned three hours). Reviewer coding speed varied, with the group leader coding all six forms, six of the reviewers coding four of the forms, one reviewer coding three forms, and the extra reviewer coding two forms. The additional reviewer was engaged after the institute in order to have at least five reviewers for each of the six assessment forms. Eight reviewers coded at least one form of each of the three assessments. From previous experience, reasonably high agreement statistics are attained with five reviewers.

Data Analysis

To derive the results from the analysis, the reviewers' responses were averaged. First, the value for each of the four alignment criteria is computed for each individual reviewer. Then the final reported value for each criterion is found by averaging the values across all reviewers. Any variance among reviewers was considered legitimate; for example, the reported DOK level for an item could fall somewhere between the two or more assigned values. Such variation could signify a lack of clarity in how the standards were written, the robustness of an item that could legitimately correspond to more than one standard and/or a DOK that falls in between two of the four defined levels. After the adjudication, reviewers were not required to change their results based on the discussion. Any large variations among reviewers in the final results represented true differences in opinion among the reviewers and were not because of coding error. These differences could be due to different standards targeting the same content knowledge or may be because an item did not explicitly correspond to any standard, but could be inferred to relate to more than one standard. Reviewers were allowed to identify one assessment item as corresponding up to three content expectations—one primary match (the expectation was for a single content match) and up to two secondary matches.

The results produced from the institute pertain only to the issue of alignment between the Mathematics Florida Standards and the six assessments that were analyzed. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met.

Alignment Criteria Used for This Analysis

This report describes the results of an alignment study of six assessments with the MAFS for Algebra 1 EOC supplemented by additional standards. Results are reported for the alignment of each assessment with the MAFS for Algebra 1 EOC as well as for the alignment of each assessment with the MAFS for Algebra 1 EOC supplemented by additional standards. Two forms of each of the three assessments were analyzed. The study addressed specific criteria related to the content agreement between the standards and assessments. Four criteria received major attention:

- Categorical Concurrence,
- Depth-of-Knowledge Consistency,
- Range-of-Knowledge Correspondence, and
- Balance of Representation.

Details on the criteria and indices used for determining the degree of alignment between standards and assessments are provided below. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had reasonably met the expectations within each reporting category. In the mathematics study, the Algebra 1 standards has three reporting categories: Algebra and Modeling (RC1); Functions and Modeling (RC2); and Statistics and the Number System (RC3). The analyses included considering the degree of alignment of each assessment form with the 45 Algebra 1 standards under these three reporting categories. In addition, this report describes the content coverage including standards other than the Algebra 1 standards. In the descriptions below, the term "standards" may be used as an umbrella term to refer to expectations in general. In addition to judging alignment between reporting categories and assessments on the basis of the four key alignment criteria, information is also reported on the quality of items by identifying items with Source-of-Challenge and other issues.

Categorical Concurrence

An important aspect of alignment between standards and assessments is whether both address the same content categories. The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. *The criterion of categorical concurrence between standards and assessments is met if the same or consistent categories of content appear in both documents.* This criterion was judged by determining whether the assessment included items measuring content from each conceptual category. The analysis assumed that the assessment had to have at least six items for measuring content from a conceptual category for an acceptable level of categorical concurrence to exist between the conceptual category and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable scale for estimating students' mastery of content for a conceptual category. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the scale, the mean score, and cutoff score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of a single item is 0.1, it was estimated that six items would produce an agreement coefficient of at least 0.63. This indicates that about 63% of the group would be consistently classified as masters or non-masters if two equivalent test administrations were employed. The agreement coefficient would increase to 0.77 if the cutoff score were increased to one standard deviation from the mean and, with a cutoff score of 1.5 standard deviations from the mean, to 0.88.

Usually states do not report student results by standards or require students to achieve a specified cutoff score on expectations related to a conceptual category. If a state did do this, then the state would seek a higher agreement coefficient than 0.63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a conceptual category, and as a basis for making some decisions about students' knowledge of that standard. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would

produce an agreement coefficient of 0.77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-knowledge consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the standards, as judged in this analysis, at least 50% of the items corresponding to a conceptual category had to be at or above the depth-of-knowledge level of the corresponding standard; 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one conceptual category of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding standards. For example, assume an assessment included six items related to one conceptual category and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three (50%) of the six items were at or above the depth-of-knowledge level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one expectation. Some leeway was used in this analysis on this criterion. If a conceptual category had between 40% and 50% of items at or above the depth-of-knowledge levels of the expectations, then it was reported that the criterion was "weakly" met.

DOK Levels for Mathematics

Interpreting and assigning depth-of-knowledge levels to both standards and assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in mathematics.

Level 1 (Recall)

DOK 1 is defined by the rote recall of information or performance of a simple, routine procedure. For example, repeating a memorized fact, definition, or term; performing a simple algorithm, rounding a number, or applying a formula are DOK 1 performances.
Performing a one-step computation or operation, executing a well-defined multi-step procedure or a direct computational algorithm are also included in this category. Examples of well-defined multi-step procedures include finding the mean or median or performing long division.
Reading information directly from a graph, entering data into an electronic device to derive an answer, or simple paraphrasing are all tasks that are considered a level of complexity comparable to recall. A student answering a DOK 1 item either knows the answer or does not: that is, the item does not need to be "figured out" or "solved."

At a DOK 1, problems in context are straightforward and the solution path is obvious. For example, the problem may contain a keyword that indicates the operation needed. Other DOK 1 examples include plotting points on a coordinate system, using coordinates with the distance formula, or drawing lines of symmetry of geometric figures.

At more advanced levels of mathematics, symbol manipulation and solving a quadratic equation or a system of two linear equations with two unknowns are considered comparable to recall, assuming students are expected or likely to use well-known procedures (e.g. factoring, completing the square, substitution, or elimination) to derive a solution. Operating on polynomials or radicals, using the laws of exponents, or simplifying rational expressions are considered rote procedures.

Verbs should not be classified as any level without considering what the verb is acting upon or the verb's direct object. "*Identify* attributes of a polygon" is recall, but "*identify* the rate of change for an exponential function" requires a more complex analysis. To *describe* by listing the steps used to solve a problem is recall (i.e., *Show your work*) whereas to *describe* by providing a mathematical argument or rationale for a solution is more complex.

Level 2 (Skills and Concepts)
DOK 2 involves engaging in some mental processing beyond a habitual response as well as decision-making about how to approach the problem or activity. This category can require conceptual understanding and/or demonstrating conceptual knowledge by explaining thinking in terms of concepts.

DOK 2 tasks includes distinguishing among mathematical ideas, processing information about the underlying structure, drawing relationships among ideas, deciding among and performing appropriate skills, applying properties or conventions within a relevant and necessary context, transforming among different representations, and interpreting and solving problems and/or graphs. When given a problem statement, formulating an equation or inequality, deriving a solution, and reporting the solution in the context of the problem fit within DOK 2. Processes such as classifying, organizing, and estimating that involve attending to multiple attributes, features, or properties also fall into this level.

Verifying that the number of objects in one set is larger or fewer than the number of objects in a second set by matching pairs or forming equivalent groups is a DOK 2 activity for a kindergartener. A first grader modeling a joining or separating situation pictorially or physically also is at this level.

Skills and concepts include constructing a graph and interpreting the meaning of critical features of a function, beyond just identifying or finding such features as well as describing the effects of parameter changes. Note, however, that using a well-defined procedure to find features of a standard function, such as the slope of a linear function with one variable or a quadratic, is a DOK 1. Graphing higher order or irregular functions is a DOK 2. Basic computation, as well as converting between different units of measurement, are generally a DOK 1, but illustrating a computation by different representations (e.g., equations and a base-ten model) to explain the results is a DOK 2. Computing measures of central tendency (applying set procedures) is a DOK 1, but interpreting such measures for a data set within its context or using measures to compare multiple data sets is a DOK 2. Performing original formal proofs is beyond DOK 2, but explaining in one's own words the reasons for an action or application of a property is comparable to a DOK 2.

Activities at a DOK 2 are not limited only to number skills, but may involve visualization skills (e.g., mentally rotating a 3D figure or transforming a figure) and probability skills requiring more than simple counting (e.g., determining a sample space or probability of a compound event). Other activities at this category include detecting or describing non-trivial patterns, explaining the purpose and use of experimental procedures, and carrying out experimental procedures.

Level 3 (Strategic Thinking)
DOK 3 requires reasoning and analyzing using mathematical principles, ideas, structure, and practices. DOK 3 includes solving involved problems; conjecturing; creating novel solutions and forms of representation; devising original proofs, mathematical arguments, and critiques of arguments; constructing mathematical models; and forming robust inferences and predictions. Although DOK 2 also involves some problem solving, DOK 3 includes situations that are non-routine, more demanding, more abstract, and more complex than DOK 2. Such activities are characterized by producing sound and valid mathematical arguments when solving problems, verifying answers, developing a proof, or drawing inferences. Note that the sophistication of a mathematical argument that would be considered DOK 3 depends on the prior knowledge and experiences of the person. For example, primary school student arguments for number problems can be a DOK 3 activity (e.g., counting number of combinations, finding shortest route from home to school, computing with large numbers) as can abstract reasoning in developing a logical argument by students in higher grades.

DOK 3 problems are those for which it is not evident from the first reading what is needed to derive a solution and so require demanding reasoning to work through. Such problems usually can be solved in different ways and may even have more than one correct solution based on different stated assumptions. Paraphrasing in one's own words or reproducing a proof that was previously demonstrated is a DOK 2. Applying properties and producing arguments in proving a theorem or identity not previously seen is a DOK 3. Also in the DOK 3 category is making sense of the mathematics in a situation, creating a mathematical model of a situation considering contextual constraints, deriving a new formula, designing and conducting an experiment, and interpreting findings.

Level 4 (Extended Thinking)
DOK 4 demands are at least as complex as those of DOK 3, but a main factor that distinguishes the two categories is the need to perform activities over days and weeks (DOK 4) rather than in one sitting (DOK 3). The extended time that accompanies this type of activity allows for creation of original work and requires metacognitive awareness that typically increases the complexity of a DOK 4 task overall, in comparison with DOK 3 activities. Category 4 activities require complex reasoning, planning, research, and verification of work. Conducting a research project, performance activity, an experiment, and a design project as well as creating a new theorem and proof fit under Category 4.

The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, collecting water temperature from a river each day for a month and then reporting the findings by constructing a graph is a DOK 2 activity. Developing a mathematical model of

the flow of water in a river for all four seasons using a number of variables would be a DOK 4 activity. It is likely that a DOK 4 activity will require making connections among a number of ideas or variables within the area of mathematics or among a number of content areas. Category 4 activities require selecting an appropriate approach among many alternatives to produce a product, conclusion, or finding, such as critiquing a body of work, synthesizing ideas in a new way, or creating an original model.

Range-of-Knowledge Correspondence

For reporting categories and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The Range-of-Knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a reporting category and an assessment considers the number of standards within the reporting category with one related assessment item/activity. Fifty percent of the standards for a reporting category must have at least one related assessment item for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a reporting category. This assumes that each expectation for a reporting category should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for a reporting category increases the likelihood that students will have to demonstrate knowledge on more than one expectation per reporting category to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectations that can be assessed. Range-of-Knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of reporting categories and a large number of expectations. If 50% or more of the objectives for a reporting category had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If between 40% and 50% of the objectives for a reporting category had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned standards and assessments require that knowledge be distributed equally or proportionally in both. The range-of-knowledge criterion only considers the number of expectations within a conceptual category that have a match (a standard with a corresponding item); it does not take into consideration how the assessment items/activities are distributed among these expectations. *The balance-of-representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the expectations for a conceptual category that have at least one related assessment item per expectation.

The index is computed by considering the difference in the proportion of expectations and the proportion of items assigned to the expectations. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a conceptual category are equally distributed among the expectations for the given conceptual category. Index values that approach 0 signify that a large proportion of the items are on only one or two of all of the expectations. Depending on the number of expectations and the number of items, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining expectations) has an index value of less than 0.5. A bimodal distribution has an index value of around 0.55 or 0.6. Index values of 0.7 or higher indicate that items/activities are distributed among all of the expectations at least to some degree (e.g., nearly every expectation has at least two items) and is used as the acceptable level on this criterion. Index values between 0.6 and 0.7 indicate the balance-of-representation criterion has only been "weakly" met.

Source-of-Challenge Criterion
The source-of-challenge criterion is only used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted mathematics standard or expectation. Bias and sensitivity issues as well as technical issues and error could all be reasons for an item to have a source-of-challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed. It was not anticipated that reviewers would find any source of challenges in this study.

Cutoffs for Alignment Criteria
For overall alignment, an assessment form is reported as *fully aligned* if no items need replacement to meet the conditions for all of the criteria described above. A test form is considered *acceptably aligned* if it needs between one and five items replaced or revised to meet the conditions for all alignment criteria. A test form is reported to *need slight adjustments* if six to ten items need to be replaced or revised to meet the criteria and is reported to *need major adjustments* if more than ten items need to be replaced or revised. These categories represent typically used cutoff levels.

Findings

Framework Analysis for Mathematics
Prior to conducting the Alignment Institute, October 18 and 19, Professor Kristen Bieda, a mathematics educator at Michigan State University, conducted a review of the design documents and other explanatory materials found for each of the three assessments. This report is included as **Appendix 1a.E**. Information from this report was used to identify additional MAFS that should be included in the analysis to better reflect content that the ACT and SAT assessments may address that were not included in the Algebra 1 standards. The design documents included test blueprints, test specifications, and curriculum standards as were available.

About 20 of the 45 Algebra 1 standards (44 percent) did not have comparable standards in any of the documents found for the ACT or SAT assessments. For example, the Algebra 1 standards include standards related to students understanding exponents, radicals, and rational and

irrational numbers (RC3: N-RN1.1-1.2). These content topics were not found in the SAT materials. These topics were found in ACT College and Career Readiness Standards, but not among the benchmarks to be assessed at the level of college and career readiness.

Another difference among the frameworks was in the area of statistics and probability. The Florida Algebra 1 standards included standards RC3: S-ID.3.8 and 9, computation and interpretation of correlation coefficients for a linear line of best fit. Neither of the SAT or ACT documents reviewed in the framework analysis considered this topic as essential understanding for college and career readiness. Also, some differences were found in the description of items. For example, the Florida Algebra 1 assessment specifications explicitly noted that items written for certain standards should be embedded in a problem context. No such explicit statements were found for the ACT or SAT. Thus, the framework analysis did reveal some design differences and variation in the content intended to be assessed.

Assessments

The mathematics assessments differed in their structure and the type(s) of items. The Florida Algebra 1 EOC test was administered over two sessions, one session per day, for a maximum time of 180 minutes. Scientific calculators are provided during the second session of the assessment. Nearly all of the items were assigned one point. One item on the Spring 2017 form was given a point value of two (Table 1a-1). Less than 50 percent of the items on the Florida assessments were multiple choice items (Table 1a-2). The majority of the items were technology-enhanced including those where the students select letters, numbers, or symbols to generate an answer (e.g., an equation); enter a replacement word or phrase; complete a graph using point, line, or arrow button; and other formats. A few of the items, 2 to 5 percent, were multi-select items where students were expected to select all of the appropriate responses from a list. Ten field test items were included on each of the Florida Algebra 1 EOC assessments. These were excluded from the analyses.

The ACT mathematics assessment consisted of 60 items completed in 60 minutes. All 60 items were multiple choice with four choices. Calculators were permitted for use when taking the ACT mathematics test but not required. Students could use most calculators, including four-function, scientific, or graphing calculators except for those explicitly prohibited such as those with built-in or downloaded algebra computer system functionality.

The SAT mathematics assessment had 58 items administered in two parts, including 20 items where calculators were not permitted and 38 items where students were permitted to use a calculator. Students were allotted 80 minutes to complete the mathematics proportion of the assessment. The SAT assessments had two types of items, multiple choice (78 percent) and grid-ins (22 percent), in which students fill in a grid to enter a positive whole number, decimal, or fraction (Table 1a-2).

Table 1a-1. Number of Items, Point Value, and Average Time per Item per Assessment for the Florida Algebra 1 Analysis

Test	Number of Items	Number of Extra Point Items	Total Point Value	Assessment Time	Average Time per Item
Florida Spring 2016	68	0	58	180 min	2.6 min
Florida Spring 2017	68	1	59	180 min	2.6 min
ACT Form 74H	60	0	60	60 min	1 min
ACT Form 74C	60	0	60	60 min	1 min
SAT Apr 2017	58	0	58	80 min	1.4 min
SAT May 2017	58	0	58	80 min	1.4 min

Table 1a-2. Number and Percent of Items by Type for Each Assessment for the Florida Algebra 1 Analysis

	Item Type								
	Multiple-choice		Multiple-select		Technology-enhanced		Fill-in-the-grid		Total Number
Test	N	%	N	%	N	%	N	%	
Florida Spring 2016	21	36	1	2	36	62			58
Florida Spring 2017	24	41	3	5	31	54			58
ACT Form 74H	60	100							60
ACT Form 74C	60	100							60
SAT Apr 2017	45	78					13	22	58
SAT May 2017	45	78					13	22	58

Standards

For all but two standards, DOK levels for the MAFS assigned by the state were used as the DOK levels in this study (http://www.fldoe.org/core/fileparse.php/12087/urlt/G9-12_Mathematics_Florida_ Standards.pdf). As noted before, the group decided to change the DOK level of two standards (G-SRT 4.9 from DOK 3 to 2 and G-SRT 4.10 from DOK 2 to 3). A summary of the levels of complexity are given in Tables 1a-3 and 1a-4. Of all the 169 standards included in the study, the majority of them (67 percent) were considered a DOK level 2, skills and concepts. About 20 percent of the standards were judged to have a DOK level 1, recall, and 14 percent to have a DOK level 3, strategic thinking. The distribution by content complexity of the 45 Algebra 1 standards was nearly the same with a slightly higher percentage of standards at a DOK level 2 (Table 1a-4). Thus, most of the standards in the analysis expected students to apply skills and to have a conceptual understanding of the mathematics.

Table 1a-3. Percent of Expectations by Depth-of-Knowledge (DOK) Levels for the Mathematics Florida Standards for Algebra 1 Supplemented with Additional Standards

Standard	Total Number of Standards	DOK Level	Number of Standards by Level	Percent within Conceptual Category by Level
RC1 Algebra & Modeling	24	1 2 3	7 12 5	29.17 50.00 20.83
RC2 Functions & Modeling	28	1 2 3	3 22 3	10.71 78.57 10.71
RC3 Statistics & the Number System	58	1 2 3	16 41 1	27.59 70.69 1.72
RC4 Geometry	43	1 2 3	5 25 13	11.63 58.14 30.23
RC5 Grades 4-8 Mathematics Standards	16	1 2 3	2 13 1	12.5 81.25 6.25
Total	169	1 2 3	33 113 23	20 67 14

Table 1a-4. Percent of Expectations by Depth-of-Knowledge (DOK) Levels for the Mathematics Florida Standards for Algebra 1

Standard	Total Number of Standards	DOK Level	Number of Standards by Level	Percent within Conceptual Category by Level
RC1 Algebra & Modeling	17	1 2 3	5 9 3	29.41 52.94 17.65
RC2 Functions & Modeling	15	1 2 3	1 12 2	6.67 80.00 13.33
RC3 Statistics & the Number System	13	1 2	1 12	7.69 92.31
Total	45	1 2 3	7 33 5	15.56 73.33 11.11

Mapping of Items by Standards

If no particular grade-level standard is targeted by a given assessment item, reviewers were instructed to code the item at the cluster, domain, or reporting category. This coding to a generic standard generally indicated that the assessment item did not target one of the standards included in the study. However, if the item is grade-appropriate, then this situation may instead indicate that there is a part of the content not expressly or precisely described in the standards, or that there is a part of the content within the standards that is being interpreted differently by different parties. Items coded to generic standards may highlight areas in the standards with missing content or where the statement of the standard is not as precise as it should be as well as a mismatch with an assessment.

Table 1a-5. Items Assigned to Generic Content Expectations by Assessment and Number of Reviewers for the Mathematics Florida Standards Alignment Analysis

Test	Generic Content Expectation	Item Number (N Reviewers)	Comments
FL Spr. 2016	RC3: S-ID	49(7)	[Information subject to nondisclosure agreements has been omitted for public release.]
FL Spr. 2017	None	--	--
ACT 74H	RC2: F-LE	12(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	RC4: G-GMD	36(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	RC5: N	50(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
ACT 74C	RC3: N-Q	3(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	RC5: N	4(2), 16(3), 57(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
SAT Apr 2017	None	--	--
SAT May 2017	RC1: A-REI	13(3), 38(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	RC4: G-GMD	26(3)	[Information subject to nondisclosure agreements has been omitted for public release.]

Very few items for any of the six assessment forms were coded to generic standards by two or more reviewers (Table 1a-5). Seven of eight reviewers indicated that Item 49 on the Florida Spring 2016 assessment did not precisely match any of the standards. [Information subject to nondisclosure agreements has been omitted for public release.] Two or three of the reviewers found three items on the ACT Form 74H and four items on the ACT Form 74C that did not precisely match any of the standards included in the study. Most of these items corresponded to middle school standards. Three reviewers found three items on the SAT May 2017 form that required students use equations in a different way than specified in the standards or mathematics identified by standards in a lower grade. Overall, nearly all of the items on the six assessment forms matched in some way the supplemented Florida Algebra 1 standards included in the study.

Table 1a-6. Number and Percent of Mathematics Florida Standards with at least One Item Found by a Majority of Reviewers as Corresponding to Algebra 1 Standards and Supplemented Standards

Test	Number of Items	Number of Algebra 1 Standards	Number of Supplement Standards	Total Standards with at least One Item
Florida Spr. 2016	58	32 (71%)	0 (0%)	32 (19%)
Florida Spr. 2017	58	32 (71%)	1 (1%)	33 (20%)
ACT Form 74H	60	15 (33%)	28 (22%)	43 (35%)
ACT Form 74C	60	13 (29%)	32 (26%)	45 (36%)
SAT Apr 2017	58	20 (44%)	13 (10%)	33 (20%)
SAT May 2017	58	18 (40%)	13 (10%)	31 (18%)

Table 1a-7. Number and Percent of Mathematics Items for Six Assessments Judged by Majority of Reviewers as Corresponding to Algebra 1 Standards and to the Supplemented Standards

Test	Total Items	Algebra 1 Standards		Supplement Standards	
		Number	Percent	Number	Percent
Florida Spring 2016	58	58	100	0	0
Florida Spring 2017	58	57	98	1	2
ACT Form 74H	60*	21	35	39	65
ACT Form 74C	60	19	32	41	68
SAT Apr 2017	58	35	60	23	40
SAT May 2017	58	40	69	18	31

* Item 29 was assigned by four reviewers to each group of standards.

All of the assessment forms had very nearly the same number of items, 58 or 60 (Table 1a-6**)**. However, the assessments varied by the number of Algebra 1 standards with corresponding items and the number of items that targeted Algebra 1 standards. The two Florida assessment forms each had items that corresponded to 32 of the 45 Algebra 1 standards (71%) (Table 1a-6). The two ACT forms had items that corresponded to 13 or 15 of the Algebra 1 standards (about 30%) and the two SAT forms had items that corresponded to 18 or 20 of the Algebra 1 standards

(about 42%). Some items on all six forms corresponded to the same standards. Nearly all items on the two Florida assessment forms mapped to the Algebra 1 standards. The majority of reviewers found that Item 28 on the Florida Spring 2017 form mapped to a standard that was not included as an Algebra 1 standard. Of the 60 items on the two ACT forms, from 32 to 35 percent of the items mapped to the Algebra 1 standards (Table 1a-7). Of the 58 items on the two SAT forms, from 60 to 69 percent of the items mapped to the Algebra 1 standards.

Comparison of Overall DOK Distribution

A comparison of the overall DOK distribution for each assessment, averaged across the two test forms, is shown in Table 1a-8. The average DOK level among the three assessments were very similar. All three assessments had a majority of items with a DOK 2, skills and concepts, about 70 percent. Another quarter of the items on all three assessments were rated as a DOK 1. One form of the Florida Algebra 1 EOC assessment had one item judged to have a DOK 3, one form of the ACT had two items as a DOK 3, and each of the SAT forms had one item as a DOK 3.

Table 1a-8. DOK Distribution, averaged across two test forms for Florida Algebra 1 EOC, ACT, and SAT

Test	DOK 1	DOK 2	DOK 3
FL Algebra 1 EOC	26%	73%	1%
ACT	26%	72%	2%
SAT	27%	71%	2%

Alignment of Mathematics Assessments with the Mathematics Florida Algebra 1 Standards

The results of the analysis for each of the four alignment criteria are summarized in Tables 1a-9.1 to 1a-9.6. More detailed data on each of the criteria are given in **Appendix 1a.B** in the first three tables for each assessment. The reviewers' notes and debriefing comments (**Appendices 1a.C** and **1a.D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In Tables 1a-9.1 to 1a-9.6, "YES," indicates that an acceptable level was attained between the assessment and the MAFS mathematics standards on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error in the system. "NO" indicates that the criterion was not met by a noticeable margin—10% over an acceptable level for Depth-of-Knowledge Consistency, 10% over an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation.

Florida Algebra 1 End-of-Course Assessment Alignment Study Results

Results of the alignment analysis for the two Florida Algebra 1 EOC assessment forms with the 45 targeted Mathematics Florida Standards indicate the assessment forms and the standards were acceptably aligned (defined as needing 5 or fewer items revised or replaced for full alignment). The alignment results for both forms were the same (Tables 1a-9.1 and 1a-9.2). The content coverage by both forms in depth and breadth was essentially the same. For each of the three reporting categories, each assessment had at least 10 corresponding items. This was a sufficient number of items to have an acceptable level on the Categorical Concurrence criterion. This indicates that the assessment had an adequate number of items for each of the three reporting conceptual categories, six or more, to make a reasonably reliable judgment about a

student's proficiency on each conceptual category. The distribution of items among the three reporting categories for the Spring 2017 form was identical to what was expected from the framework analysis—41 percent for RC1, 40 percent for RC2, and 19 percent for RC3. However, the earlier form, Spring 2016, over emphasized RC1 (Algebra and Modeling) by about three items with 48 percent of the items corresponding to that reporting category.

Table 1a-9.1. Summary of Acceptable Levels on Alignment Criteria for the Florida Algebra 1 Spring 2016 assessment and the Algebra 1 Standards *(N=58 Items)*

Florida Algebra 1 Spring 2016	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (27.50)	YES (76%)	YES (79%)	YES (0.74)
RC2 Functions & Modeling	YES (19.63)	YES (78%)	YES (78%)	YES (0.76)
RC3 Statistics & the Number System	YES (9.88)	YES (95%)	WEAK (40%)	YES (0.73)

Table 1a-9.2. Summary of Acceptable Levels on Alignment Criteria for the Florida Algebra 1 Spring 2017 assessment and the Algebra 1 Standards *(N=58 Items)*

Florida Algebra 1 Spring 2017	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (24.6)	YES (73%)	YES (74%)	YES (0.87)
RC2 Functions & Modeling	YES (24.0)	YES (74%)	YES (81%)	YES (0.85)
RC3 Statistics & the Number System	YES (10.2)	YES (88%)	WEAK (48%)	YES (0.84)

The content complexity of the items was very comparable to the content complexity of the corresponding standards. Sixty-four percent or more of the items had a DOK level the same as the DOK of the corresponding standards (see **Appendix 1a.B** and tables for Florida

assessments). Over three-quarters of the items had a DOK that was the same or higher than the corresponding standard. This is well over the minimum requirement to have an acceptable level of the Depth-of-Knowledge Consistency criteria of 50 percent. Both Florida assessments varied some from the expected DOK levels expressed in the framework analysis. From the framework analysis, the proposed distribution of items by DOK was DOK 1 (10-20 percent), DOK 2 (60-80 percent), and DOK 3 (10-20 percent). The results from the alignment analysis indicate that the Spring 2016 form had 28 percent at DOK 1 and 72 percent at DOK 2. Spring 2017 had 24 percent at DOK 1, 74 percent at DOK 2, and 2 percent at DOK 3. Both forms matched the framework for the proportion of DOK 2 items, but included a higher percentage than proposed of DOK 1 items and a lower percentage of DOK 3 items. Only one item on either of the forms was judged by the majority of reviewers to have DOK 3 – Item 8 on the Spring 2016 form.

Range-of-Knowledge Correspondence criterion was the only alignment criteria with a small issue for the Florida tests, where the criterion for alignment was weak. Both of the Florida Algebra 1 assessment forms had items that corresponded to less than 50% of the 13 standards in RC3 (Statistics and Number System). These issues in range could be resolved by replacing two items on the Spring 2016 form and one item on the Spring 2017 form that corresponded to a standard not currently assessed. For the other two reporting categories, both assessments targeted a very high percentage of the standards, over 70 percent. Both of the Florida Algebra 1 assessment forms had acceptable balance. On both assessments, items were distributed fairly evenly among the standards. Any one standard generally had one or two items. One standard on the Spring 2016 form did have five corresponding items, emphasizing this standard a little more than the others.

Reviewers' notes in **Appendix 1a.C** identified a few items that only targeted a part of the corresponding standard. [Information subject to nondisclosure agreements has been omitted for public release.]

Overall, the two Florida Algebra 1 assessment forms and the 45 Mathematics Florida Standards designated for the course were found to be acceptably aligned.

ACT Alignment Study Results

Only about one-third of the items on the two ACT forms were found to correspond to the 45 Florida Algebra 1 standards (Table 1a-7). Another 11 or 12 items (20 percent) corresponded to the standards related to the three reporting categories (algebra, functions, statistics and number), but were not among those designed for the Algebra 1 course (See **Appendix 1a.B** for the ACT forms). Another 22 percent of the ACT items corresponded to geometry and 25 percent corresponded to standards from lower grades (e.g., proportions, computation, word problems, etc.). Considering only the third of the ACT items that mapped to one of the 45 Florida Algebra 1 standards, the alignment between the assessment and standards needed slight adjustment.

Table 1a-9.3. Summary of Acceptable Levels on Alignment Criteria for the ACT Form 74H assessment and the Florida Algebra 1 Standards *(N=60 Items)*

ACT Form 74H	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (8.0)	YES (77%)	NO (39%)	YES (0.86)
RC2 Functions & Modeling	YES (8.57)	YES (68%)	WEAK (46%)	YES (0.88)
RC3 Statistics & the Number System	NO (4.57)	YES (87%)	NO (24%)	YES (0.79)

Table 1a-9.4. Summary of Acceptable Levels on Alignment Criteria for the ACT Form 74C assessment and the Florida Algebra 1 Standards *(N=60 Items)*

ACT Form 74C	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (8.6)	YES (83%)	NO (36%)	YES (0.87)
RC2 Functions & Modeling	NO (4.8)	WEAK (46%)	WEAK (40%)	YES (0.85)
RC3 Statistics & the Number System	NO (5.8)	YES (87%)	NO (28%)	YES (0.84)

As shown in Tables 1a-9.3 and 1a-9.4, the Categorical Concurrence criteria was acceptably met by the ACT Form 74H for two of the three reporting categories, but for only one of the three reporting categories by the ACT Form 74C. Three of the six criteria (across the two tests) did not show adequate alignment. For the reporting categories with at least six or more items, the assessment had about eight items on the average. For the other reporting categories, the average number of items was about five items.

The one-third of the items that corresponded to the Florida Algebra 1 standards generally had an appropriate DOK level compared to the corresponding standards. Only Form 74C for RC2 had a weakness in the DOK Consistency. For this reporting category, 53% of the corresponding items had a DOK that was lower than the DOK of the matching standard. For the other

reporting categories from 68 percent or higher of the items had a DOK that was the same or higher than the DOK of the corresponding standard (see **Appendix 1a.B**).

Considering all 60 items on the ACT, the majority of the items on both forms were judged to have a DOK level 2, 77% on ACT 74H and 68% on ACT 74C (see **Appendix 1a.B**). Only two items on Form 74C were judged to have a DOK 3 by the majority of reviewers – Items 41 and 54. The other items were judged to have a DOK 1, 23% on ACT 74H and 28% on ACT 74C. The distribution of the items by DOK levels varies some from the intended distribution of DOK levels for the ACT forms in the framework analysis (12-15 percent DOK 1, 53-70 percent DOK 2, and 26-34 percent DOK 3). This suggests that the DOK definitions used by ACT were different from the original Webb definitions, which have been updated and were used in this study.

The main alignment issue of the two ACT forms with the Florida Algebra 1 standards is with Range-of-Knowledge Correspondence or the breadth of coverage, which were found to not be aligned or weakly aligned per the criteria. With only 21 and 19 items on the two ACT forms that were found to correspond to the Algebra 1 standards, it is difficult to reach the acceptable level of 50% of the 45 Florida Algebra 1 standards with at least one item. Considering only nearly one-third of the items that mapped to the Florida Algebra 1 standards, 36-39 percent of the standards under RC1 had at least one item, 40-46 percent of the standards under RC2 had at least one item, and 24 to 28 percent of the standards under RC3 had at least one item. Each of these ranges is lower than the 50% of the standards cutoff used in this study to indicate an acceptable coverage of the standards. Even when the Algebra 1 standards in RC1-3 are supplemented by the additional standards, the proportion of standards hit were between 16 to 34 percent of the standards.

On both of the ACT forms, only one or two items were mapped to any one standard. Thus, no standard was overemphasized compared to any other standard. As such, the Balance of Representation was acceptably met by both ACT forms.

Overall, the alignment between the two ACT forms and the Florida Algebra 1 standards would need slight adjustment to be considered aligned according to the criteria used in this study. Full alignment could be attained by having eight more items with the 21 items on Form 74H and seven more items with the 19 items on Form 74C that currently map to the 45 Florida Algebra 1 standards. These additional items are needed to increase the number of items to six for reporting categories RC3 (Form 74H) and RC2 and RC3 (Form 74C) and to increase the number of standards with at least one item for all three reporting categories for each form. Thus, about 12-13 percent of the 60 items on a form of the ACT would need to be replaced for alignment with the Florida Algebra 1 standards.

Two reporting categories were added to the analysis, RC4 Geometry and RC5 Grades 4-8 Mathematics Standards. The two ACT forms had items that were mapped to 21 to 24 percent of standards under the geometry reporting category (RC4) and from 26 to 28 percent of standards under the Grades 4-8 standards (RC5) (see **Appendix 1a.B**). Nearly half of the items on the ACT forms mapped to geometry or were below high school level. In their debriefing comments, a few reviewers noted having difficulty finding a standard that matched some items, in part, because they targeted content addressed in lower grades.

SAT Alignment Study Results

About two-thirds of the items on the SAT April and May 2017 forms corresponded to one of the 45 Algebra 1 Florida Standards. Another eight (April 2017 form) and three (May 2017 form) items targeted standards under the three reporting categories for the Algebra 1 course, but not one of the standards designated as related to the Algebra 1 course. The other one-third of the items included about six items that targeted RC4 (geometry) and about ten items that targeted RC5 (Grades 4-8 standards). Different from the other two assessments analyzed, the two SAT forms differed some in the allocation of items among the reporting categories and the level of content complexity. If only the 37 and 41 items that targeted one of the 45 Algebra 1 standards are considered, then the alignment for one form (April 2017) was acceptable, but for the other form (May 2017) alignment would require slight adjustment.

The Categorical Concurrence criterion was acceptable for all three reporting categories for both SAT forms. The number of items mapping to each reporting category ranged from six items (RC3 April 2017) to 24 items (RC1 May 2017). The two forms differed some in the number of items found to match standards under RC1, 19 items on the April 2017 form and 24 items on the May 2017 form. But still the number of items on either form was sufficient to make a reliable estimate of a student's proficiency on the reporting category. The largest number of items corresponded to RC1 (Algebra and Modeling), then RC2 (Functions and Modeling), and then RC3 (Statistics and the Number System).

Table 1a-9.5. Summary of Acceptable Levels on Alignment Criteria for the SAT April 2017 assessment and the Florida Algebra 1 Standards *(N=58 Items)*

SAT April 2017	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (19.2)	YES (68%)	YES (57%)	YES (0.81)
RC2 Functions & Modeling	YES (11.6)	YES (88%)	WEAK (48%)	YES (0.81)
RC3 Statistics & the Number System	YES (6.2)	YES (94%)	NO (31%)	YES (0.82)

Table 1a-9.6. Summary of Acceptable Levels on Alignment Criteria for the SAT May 2017 assessment and the Florida Algebra 1 Standards *(N=58 Items)*

SAT May 2017	***Alignment Criteria***			
Reporting Categories	***Categorical Concurrence (Avg. # items)***	***Depth-of-Knowledge Consistency (Percent at or above)***	***Range-of-Knowledge Correspondence (Percent of standards assessed)***	***Balance of Representation (Index 0-1)***
RC1 Algebra & Modeling	YES (24.2)	YES (63%)	YES (60%)	YES (0.75)
RC2 Functions & Modeling	YES (9.2)	YES (69%)	NO (36%)	YES (0.71)
RC3 Statistics & the Number System	YES (7.8)	YES (95%)	NO (26%)	YES (0.75)

The 37 and 41 items on each form and the Algebra 1 standards had an acceptable DOK consistency. Over 50 percent of the items on each form and for each of the three reporting categories had a DOK level that was at least as high as the DOK for the corresponding standard. The majority of items on each of the SAT forms had a DOK level 2, skill and concepts, 77 percent on the April 2017 form and 65 percent on the May 2017 form. Reviewers were in agreement that each of the forms had one item with a DOK level 3 that required strategic thinking, Item 49 on the April form and Item 54 on the May form.

Range-of-Knowledge Correspondence was the only alignment issue for the SAT, with the criteria being either weakly met or not met. The April 2017 form had an acceptable level of over 50 percent of the standards under a reporting category with at least one corresponding item for two of the three reporting categories, RC1 and RC2. This form only had items that corresponded to four of the 13 standards under RC3, about 30 percent. The SAT May 2017 form had items that corresponded to 11 of the 17 standards under RC1 and was judged to have an acceptable range. The May 2017 form for the other two reporting categories did not have as much breadth. The form had items that corresponded to five of 15 standards under RC2 (36 percent) and three of 13 standards under RC3 (26 percent), both with items corresponding to fewer than half of the standards within the reporting categories.

The Balance of Representation Index only takes into consideration those standards with at least one assessed item. Most of these standards on either SAT forms had one or two corresponding items. As a consequence, the Balance of Representation was acceptable for each form. However, the May 2017 form did have seven of the standards with three or four corresponding items. This lowered the Balance Index for this form, but not below the 0.70 level that is used as an acceptable criterion. [Information subject to nondisclosure agreements has been omitted for public release.]

Overall, the SAT April 2017 and the Algebra 1 standards were considered acceptably aligned. Four additional items would need to be added to the 37 items that corresponded to the Algebra

1 standards to adjust the assessment to have full alignment. One of the items would need to target an additional standard under RC1 and three items would need to target three additional standards under RC3. The SAT May 2017 and the Algebra 1 standards would need more items, or what would be considered slight adjustment, to reach full alignment. At least seven items, three for RC2 and four for RC3, would need to be added to the 41 items that now target Algebra 1 standards to have full alignment. These items would need to target additional standards to improve on the range of content assessed.

Reliability among Reviewers

The overall intraclass correlation among the mathematics reviewers' assignment of DOK levels to items was reasonably high for five to eight reviewers for all six analyses (Table 1a-10). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers. The intraclass correlations for assigning DOK levels to items for all six analyses were 0.83 or higher. A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other four to seven reviewers. With eight reviewers a total of 28 comparisons were computed for each item. A pairwise reporting category agreement of 0.90 is the desired level. For three of the six analyses, the pairwise reporting category agreement met the desired level by one form for each of the three assessments. For the other three assessments, the reporting category was reasonably high for the Florida Spring 2016 (0.88) and the SAT April 2017 forms (0.84), but was low for the ACT Form 74H (0.72). The desired pairwise standard agreement of at least 0.50 was met for five of the six assessment forms. Only the ACT Form 74H was lower than this value (0.46).

Table 1a-10. Intraclass and Pairwise Comparisons for the Alignment Analysis of the Mathematics Florida Standards for Algebra 1 with Supplement Standards and Six Assessments

Grade	Intraclass Correlation	Pairwise: Comparison	Pairwise: Reporting Categories	Pairwise: Standard
Florida Spring 2016	0.92	0.72	0.88	0.67
Florida Spring 2017	0.95	0.83	0.92	0.80
ACT Form 74H	0.83	0.63	0.72	0.46
ACT Form 74C	0.90	0.80	0.90	0.69
SAT Apr 2017	0.84	0.72	0.84	0.52
SAT May 2017	0.84	0.71	0.90	0.74

Reviewers did engage in an adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics were computed after adjudication. If the intraclass correlation and pairwise agreements are low after adjudication, then this could be an indication of a misfit between the standards and the assessment items. Reviewers will vary in their codings the more they have difficulty in finding a precise match between an assessment item and the standards.

Summary of Comparisons of the Three Assessments

Both the ACT and SAT assessment forms covered content addressed by the Florida Algebra 1 End-of-Course assessment along with additional content. The similarities and differences in the content coverage by each of the assessments are summarized in Table 1a-11. The domain of standards and the three reporting categories targeted by the Florida assessment are highlighted in the left column under Domain. From Table 1a-6, it is apparent that the Florida assessment covers in some way 71 percent of the 45 Algebra 1 standards. By reporting category, the 58 items on a Florida assessment were distributed about 43 percent measuring content under RC1 (Algebra and Modeling), 36 percent measuring content under RC2 (Functions and Modeling), and 20 percent measuring content under RC3 (Statistics and the Number System). Overall, with 71 percent of the Algebra 1 standards with at least one item, the items on the assessment had sufficient depth and breadth to be aligned with the Algebra 1 standards.

The ACT and the SAT assessment forms each included items that targeted similar content as assessed by the Florida Algebra 1 EOC assessment, but not to the same degree. About two-thirds of the items on the SAT targeted 40 to 44 percent of the Algebra 1 standards (Table 1a-6). About one-third of the items on the ACT targeted 29 to 33 percent of the Algebra 1 standards (Table 1a-6). The other items on the SAT and ACT assessments corresponded to Florida standards other than the Algebra 1 standards.

When the three assessments are analyzed at a more general level of content—domain and reporting category—then some differences are apparent (Table 1a-11). The SAT and the Florida assessments had a comparable proportion of items for reporting categories RC1 (Algebra and Modeling) and RC3 (Statistics and the Number System). The ACT and the Florida assessments had a comparable proportion of items for RC3. For RC1 (Algebra and Modeling), the Florida and SAT assessments had items that targeted all four of the underlying domains. The ACT assessments had items that targeted three of the four underlying domains, but not A-SSE. For RC2 (Functions and Modeling), at least one form for all three assessments had an item that corresponded to a standard under each Algebra 1 domain. As such, both the ACT and the SAT assessments addressed similar range in content, but just with fewer items, about half the number of items as the Florida assessment. The ACT assessments had three or four items that related to trigonometric functions that were not addressed by the Florida Algebra 1 standards. Only one item on one form of the SAT was found to correspond to a standard under the trigonometric functions domain.

Table 1a-11. Number of Items by Domain and Reporting Categories for Each of the Six Assessment Forms *(* indicates domains included in the Florida Algebra 1 Standards)*

Domain	FL2016	FL2017	ACT 74H	ACT 74C	SATApr17	SATMay17
	Number of Items a Majority of Reviewers Coded to a Standard Under the Domain					
RC1						
A-SSE*	4	4			1	4
A-APR*	2	3		2	3	
A-CED*	9	6	4	2	8	10
A-REI*	11	11	3	5	9	12
Total RC1	26(45%)	24(41%)	7(12%)	9(15%)	21(36%)	26(45%)
RC2						
F-IF*	13	13	4	3	6	8
F-BF*	3	3	1	3	3	
F-LE*	4	6	1		1	1
F-TF			3	2	1	
Total RC2	20(34%)	22(38%)	9(15%)	8(13%)	11(19%)	9(16%)
RC3						
N-RN*	5	3	2	3	2	3
N-Q*				2		
N-CN			2	1		
N-VM			1	1		
S-ID*	6	9	2	1	5	4
S-IC					1	2
S-CP			4	3	1	
Total RC3	11(19%)	12(21%)	11(18%)	11(18%)	9(16%)	9(16%)
RC4						
G-CO			5	6	2	
G-SRT			1	2	1	2
G-C			1	1		1
G-GPE			3	4	2	2
G-GMD			1	1		1
G-MG			1	1		
Total RC4	0	0	12(20%)	15(25%)	5(9%)	6(10%)
RC5						
N			6	6	7	5
EE			1	1		1
SP			3	5	2	1
G			5	3		1
Total RC5	0	0	15(25%)	15(25%)	9(16%)	8(14%)
Overall Total	57	58	54	58	55	58
Items without a majority std.	Item 65		Items 13, 29, 32, 46, 49, 50	Items 4, 40	Items 38, 41, 44	Item 27

For RC3, the SAT had a similar coverage of content to the Florida assessments on two of the domains. The ACT also targeted these two domains, but also had items that corresponded to three domains not targeted by the other two assessments (N-Q, including Algebra 1 standards, N-CN, and N-VM). Overall, both the ACT and SAT targeted to some degree similar domains under the three Algebra reporting categories but with a fewer number of items.

Both the ACT and the SAT assessments included items that measured content not expected by the Algebra 1 standards. The ACT had about 20-25 percent of its items that targeted geometry and 25 percent of its items that targeted content expected to be learned by students in grades 4-8 (proportions, computations with fractions, solving word problems with whole numbers, etc.). The SAT assessment also had items that targeted similar content, but in a lower proportion, 10 percent related to geometry and 15 percent related to grades 4-8 content.

Items from all three types of assessments were comparable in the level of content complexity (Table 1a-8). Most of the items on all three assessments expected students to apply skills and conceptual understanding (DOK 2). At most only one or two items on any of the assessment forms were judged to require students to do significant reasoning (DOK 3).

The 45 Mathematics Florida Algebra 1 Standards and the two forms of the Florida assessments (2016 and 2017) were found to be acceptably aligned on all four of the major alignment criteria. For full alignment, the Florida test forms would need only one or two items revised or replaced. For both of the ACT and the SAT to reach full alignment with the 45 Mathematics Florida Algebra 1 standards, a greater number of revised or replaced items would be required to supplement each. The SAT had 37 items that mapped to Algebra 1 standards. These items would need to be supplemented with an additional four to seven items per form to attain full alignment. The additional items would need to target Algebra 1 standards not assessed under RC2 and RC3. The ACT had about 20 items that mapped to Algebra 1 standards. These items would need to be supplemented with seven or eight items per form to attain full alignment. These additional items would be needed to increase the number of items targeting standards under RC2 and RC3 and to increase the number of standards targeted under all three reporting categories to provide better breadth. While augmenting the ACT or SAT to attain full alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process. It also adds to the complexity of the assessment administration process since augmented items must be administered separately from the ACT or the SAT.

The three assessments also varied by the item types included on each form. The Florida Algebra 1 assessments used multiple-choice items for fewer than 50 percent of its items. The SAT used 78% multiple-choice items and the ACT had only multiple-choice items. The majority of items on the Florida assessment were technology-enhanced items that required students to produce an answer, usually by dragging and dropping the appropriate symbol or character. The SAT used grid-in items in addition to multiple-choice items. These items required students to darken the appropriate digit(s) among those listed. Varying the item type did not influence the content complexity among the assessments. The distribution of items by DOK was essentially the same across all three assessments. However, it is important to note that varying from a multiple-choice format did require additional time allocated for the testing. The Florida assessments

were allocated 180 minutes in two sessions, the SAT is administered in 80 minutes, and the ACT is administered in 60 minutes.

Conclusion

The main question for the alignment analysis was to what degree the ACT or SAT can be used in lieu of the Florida Algebra 1 assessment designed to assess student proficiency on 45 Mathematics Florida Algebra 1 standards to meet federal requirements. The two Florida test forms were found to be acceptably aligned with the Algebra 1 standards, needing only one or two items revised or replaced for full alignment.

Averaging across the two test forms of each, neither the ACT nor SAT were found to be acceptably aligned without need of some adjustments. Neither the ACT nor the SAT had items that corresponded to a sufficient number of standards within the three reporting categories to be considered to have an acceptable breadth in coverage of the Algebra 1 standards. The analysis indicated that both ACT test forms needed slight adjustment; about seven or eight items would need to be added to each ACT test form to meet the minimum cutoffs for full alignment. One SAT test form was found to be acceptably aligned, requiring addition of four items to meet minimal full alignment, while the other test form was found to need slight adjustment, requiring addition of seven items to attain the minimal full alignment according to the criteria used in this study. Thus, the SAT test and the Florida Algebra 1 standards were found to be conditionally aligned, depending on the test form considered. While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, it should be noted that augmentation adds costs and complexity to the assessment administration process.

Even though all three assessments had a similar number of items, 58 or 60, the assessments varied in the allocation of those items among topics. All of the items on the Florida assessments corresponded to the Algebra 1 standards. Only one-third of the items on the ACT, 19-21, and two-thirds of the items on the SAT, 37, corresponded to the Algebra 1 standards. The SAT assessment was more comparable to the Florida assessments in coverage of the Algebra and Modeling reporting category (RC1). Both the ACT and SAT had about the same number of items that targeted Functions and Modeling reporting category (RC2), but with fewer items than did the Florida assessment. For the Statistics and Number reporting category (RC3), the ACT had the same number of items as did the Florida assessment, but some of those items targeted standards not included in the Algebra 1 standards. The SAT had slightly fewer items corresponding to RC3, but these items were more similarly allocated by standards as were the items on for the Florida Algebra 1 EOC.

Overall, the ACT had items that corresponded to a greater number of standards than either of the other two assessments. However, nearly a quarter of these items targeted geometry and another quarter targeted topics corresponding to standards in grades 4-8. The SAT also targeted geometry and grades 4-8 standards but to a lesser degree. All assessments were comparable on the content complexity of the items. A large percentage of items on all assessments had a DOK 2 and were judged to require students to apply mathematical skills and conceptual understanding.

Based on the results of the test forms analyzed, neither the SAT nor the ACT assessment is fully aligned to the Florida Algebra 1 standards. The current Florida Algebra 1 EOC assessment is nearly fully aligned, requiring revision or replacement of only one or two items to be fully aligned. Both the ACT and SAT assessments would need to be augmented to have the breadth and depth with the Algebra 1 standards called for by federal regulations. More items on the SAT corresponded to the Algebra 1 standards than did the items on the ACT, about two-thirds compared to one-third.

References

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and Mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

1B – English Language Arts Alignment Studies
Executive Summary

This is a report of a two-stage content analysis in the area of English Language Arts (ELA) that was conducted during the month of October, 2017, to provide information that could be used to judge the degree to which the ACT and/or SAT meet Criteria 1 and 2 (related to alignment, from Florida RFP 2018-48) for their suitability to be administered in lieu of Florida's Grade 10 Statewide Standardized ELA Assessment for high school students, consistent with federal requirements under 20 U.S.C.s. 6311(b)(2)(H). More specifically, this content analysis addressed the question of alignment between the ACT or SAT with the Language Arts Florida Standards (LAFS) used to develop the Florida Grade 10 ELA assessment administered in the spring of 2016 and 2017. As such, the study focused on the degree to which the assessments, including the current Florida Grade 10 Statewide Standardized ELA Assessment, addressed the full depth and breadth of the LAFS used to develop the Florida Grade 10 ELA assessment. This alignment analysis is one of a concert of studies conducted in response to Florida RFP 2018-48.

The alignment analysis consisted of two stages:

Stage I: An analysis of ELA assessment framework documents; and
Stage II: An in-person content alignment institute.

The first stage of the analysis provided information about the ELA assessment structure and design similarities and differences. This analysis was conducted by literacy expert Dr. Erin Quast of Illinois State University. The report from the framework analysis can be found in **Appendix 1b.E** of this document. The second stage of the analysis was a three-day in-person alignment institute that was held from October 18-20, in Orlando, Florida, to analyze the agreement between the LAFS and two forms of each of three assessments: Florida's Grade 10 Statewide Standardized ELA Assessment for high school students, the ACT, and the SAT. A group of seven Florida educators and three external reviewers participated in the analysis of the ELA assessments. All panelists were selected because of their notable K-12 education experience and content expertise.

The degree of alignment of a test form with the corresponding standards can be considered in terms of the degree to which specific alignment criteria are met as well as in terms of the total number of items, if any, that would need revision or replacement for full alignment. In terms of meeting the specific alignment criteria, both of the Florida test forms analyzed met all of the alignment criteria for all reporting categories with one exception: neither test form was found to meet the Depth of Knowledge (DOK) expected by the standards within Reporting Category 4 (RC4: Language and Editing). Both test forms for the ACT and SAT also failed to meet this criterion for RC 4. Both of the ACT test forms and one of the SAT test forms also only weakly met or did not meet the Depth of Knowledge expected by the standards within Reporting Category 2 (RC2: Craft and Structure). Thus, none of the test forms could be considered to address the full depth of the LAFS, but the Florida test forms addressed the full depth of the standards to a greater degree than either the ACT or the SAT test forms did. In addition, neither the ACT nor the SAT had items that corresponded to a sufficient number of standards for one of

the reporting categories (RC3: Integration of Knowledge and Ideas) to be considered to have an acceptable breadth in coverage of the Language Arts Florida Standards.

In terms of the number of items that would need revision or replacement for full alignment, both Florida test forms were found to be acceptably aligned—defined as needing 5 or fewer items revised or replaced to meet the minimum cutoffs for full alignment. One SAT test form was also found to be acceptably aligned. Both of the Florida test forms and the one SAT test form just barely met the cutoff for "acceptable" alignment; all would need five items revised or replaced to meet the minimum cutoffs for full alignment with the Florida Grade 10 LAFS. The second SAT test form was found to need slight adjustments—defined as needing six to 10 items revised or replaced to meet the minimum cutoffs for full alignment. That second SAT test form needed seven items revised or replaced to meet the minimum cutoffs for full alignment with the Florida Grade 10 LAFS. Thus, alignment of the SAT was found to depend on the test form. Study results show that the ACT would need major adjustments—defined as needing 10 or more items revised or replaced—to meet the minimum cutoffs for full alignment with the Florida Grade 10 LAFS.

In addition to computer-scored items, each assessment included a single weighted writing prompt that was evaluated according to a three-part or four-part rubric (scoring key). The writing prompts for all test forms were considered to target appropriate corresponding writing standards at an appropriate level of complexity. Reviewers commented on the time difference for the essay component of the Florida assessment (120 minutes) compared with the ACT (35 minutes) and SAT (50 minutes) essays and noted that limited time affords less of an opportunity to meet the full depth of some of the expectations within the Text-Based Writing reporting category.

While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, it should be noted that augmentation tends to be a rather expensive process and adds complexity to the administration of the tests, since items used to augment a test need to be administered separately from the college entrance test. Without such augmentation, however, and in particular for the ACT, these tests might not be viewed as meeting the United States Education Department (USED) criteria for aligned tests, thus jeopardizing the approval of the use of the college admissions tests in the federal standards and assessment peer review process.

Introduction and Methodology

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. As such, alignment is a quality of the relationship between expectations and assessments and not an attribute solely of either of these two system components. Alignment describes the match between expectations and an assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the multiple criteria described in detail in a National Institute for Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997). The corresponding methodology used to evaluate alignment has been refined and improved over the last 20 years, yielding a flexible, effective, and efficient analytical approach.

This is a report of a two-stage alignment analysis in the area of English Language Arts (ELA) that was conducted during the month of October, 2017, to provide information that could be used to judge the degree that the ACT or SAT meet Criteria 1 and 2 (related to alignment, from Florida RFP 2018-48) for their suitability to be administered in lieu of Florida's Grade 10 Statewide Standardized ELA Assessment for high school students, consistent with federal requirements under 20 U.S.C.s. 6311(b)(2)(H). More specifically, this study addressed the question of alignment between the ACT or SAT with the Language Arts Florida Standards (LAFS) used to develop the Florida Grade 10 ELA assessment administered in the spring of 2017. As such, the study focused on the degree to which the assessments, including the current Florida Grade 10 Statewide Standardized ELA Assessment, addressed the full depth and breadth of the LAFS used to develop the Florida Grade 10 ELA assessment. This alignment analysis is one of a concert of studies conducted in response to the Florida RFP 2018-48. A parallel alignment study was done for the mathematics assessments (described in a separate report).

The alignment analysis consisted of two stages:

- **Stage I:** An analysis of ELA assessment framework documents; and
- **Stage II:** An in-person content alignment institute.

The Stage I framework analysis was done by teacher education (literacy focus) Professor Erin Quast, of Illinois State University. Dr. Quast analyzed the specification of ELA content in supporting documents for each of the three assessments including blueprints, item specifications, item type, and other relevant materials that were used in developing tests or interpreting scores. The framework analysis yielded a comparison of overall test claims and assessment targets, descriptions of how specific terms and concepts were used in each of the frameworks, and identification of any relevant structural variation among the three frameworks for each content area including any differences in item types, emphasis in content topics, type of reading passages used, sizes of numbers used, and other factors. Contextual factors such as the

allotted time for essay writing were also considered. Dr. Quast's Stage I report is included in **Appendix 1b.E** of this report. This information, along with information about passages, item types, and other details, was used to prepare for the in-person alignment study. Findings from the framework analysis are also summarized in the Findings section of this report.

The Stage II in-person content alignment institute was held over three days, October 18-20, in Orlando, Florida, at the Hyatt Place Orlando/Buena Vista. Both ELA and math assessments were reviewed at the institute. Ten reviewers served on the ELA panel. The ELA panel leader, a retired K-12 reading consultant and provider of professional development from Wisconsin, had served as a leader and reviewer in numerous other alignment studies. A second external reviewer was a doctoral candidate, instructor, and pre-service teacher supervisor at the University of Wisconsin, Madison, who had participated in many other alignment studies or item analyses. A third external reviewer, currently Interim Director of Training and Strategic Partnerships and Bridge EdU, and who has roles in committees for the National Assessment of Educational Progress and for the Partnership for Assessment of Readiness for College and Careers (PARCC) had participated in one other alignment study that similarly compared ACT and SAT assessments with a third assessment. Seven Florida ELA educators from districts across the state and with English for Speakers of Other Languages (ESOL), Reading, and Exceptional Student Education (ESE) endorsements participated as reviewers who were selected from a list of highly qualified educators provided by the Florida Department of Education. Study director Norman Webb is the researcher who developed the alignment study procedures and criteria (through the National Institute for Science Education in 1997, funded by the National Science Foundation, and in cooperation with the Council of Chief State School Officers) that influenced the specification of alignment criteria by the U.S. Department of Education. The Webb alignment process has been used to analyze curriculum standards and assessments in at least 30 states to satisfy or to prepare to satisfy the Title I compliance as required by the United States Department of Education (USED). Study Technical Director Sara Christopherson has participated in and led Webb alignment studies since 2005 for over 20 states as well as for other entities.

From seven to ten panelists reviewed each assessment. For efficiency and convenience, panelists used the paper form of the assessment. The paper form was considered identical to the online version. Two forms of each assessment were analyzed. Data from each of the forms for one assessment designed to be parallel in structure provides a measure of consistency in coding for the assessment.

The Version 2 of the Web Alignment Tool (WATv2) was used to enter all of the content analysis codes during the institute. The WATv2 is a web-based tool connected to the server at the Wisconsin Center for Education Research (WCER) at the University of Wisconsin-Madison. It was designed to be used with the Webb process for analyzing the alignment between assessments and standards. Prior to the Institute, a group number was set up on the WATv2 for each of the two panels. Each panel was assigned one or more group identification numbers and the group leader was designated. Then the reporting categories and standards were entered into the WATv2 along with the information for each assessment, including the number of items, the weight (point value) given to each item, and additional comments such as the identification

number for the item to help panelists find the correct item. A sequential account of the alignment study procedures is provided below.

Training and Coding

In the morning of the first day of the alignment institute, reviewers in both the English Language Arts (ELA) group and the mathematics group received an overview of the purpose of their work, the coding process, and general training on the Depth-of-Knowledge (DOK) definitions used to describe content complexity. All reviewers had some understanding of the DOK levels prior to the institute. The general training at the alignment institute was crafted to contextualize the origins of DOK (to inform alignment studies of standards and assessments) and purpose (to differentiate between and among degrees of complexity), and to highlight common misinterpretations and misconceptions to help reviewers better understand and, therefore, consistently apply the depth of knowledge (DOK) language system. Panelists also practiced assigning DOK to sample assessment items that were selected to foster important discussions that promote improved conceptual understanding of DOK. Appropriate training of the panelists at the alignment institute is critical to the success of the project. A necessary outcome of training is for panelists to have a common, calibrated understanding of the DOK language system for describing categories of complexity.

The two groups were then separated into different rooms to receive more detailed training on the DOK levels for each content area. Through interactive and participatory training, panelists reviewed the ELA-specific definitions of the four DOK levels and worked toward a common understanding of the difference between and among each of the levels of complexity. Definitions for each DOK level for ELA are included within this report. Reviewers then worked to calibrate their use of DOK to evaluate the complexity of a subset of the standards, first assigning DOK individually and then participating in a consensus discussion. After completing coding and discussion of the subset, the panelists reviewed the DOK levels previously assigned to the LAFS (completed by other expert panels using a similar process) and flagged any standards that they wanted to discuss further, that they thought needed clarification, and/or that had a DOK assigned that they thought should be considered for adjustment because it did not accurately depict the appropriate level of content complexity. Group leaders facilitated discussions for any standards that one or more panelists flagged. If the discussion resulted in a decision to change the DOK that was assigned to a standard, then that change was made in the online data collection system, the WATv2.

The Language Arts Florida Standards are a modified version of the Common Core State Standards (CCSS). The Common Core State Standards were developed in 2010 through the coordination of the National Governors Association Center for Best Practices (NGA Center) and the Council of Chief State School Officers (CCSSO). The standards were designed to provide a clear and consistent framework to prepare pre-K through grade 12 students for college and the workforce. The standards were written to describe the knowledge and skills students should have within their K-12 education careers so that high school graduates will be able to succeed in entry-level, credit-bearing academic college courses and in workforce training programs. The CCSS have been widely used by over half of the states in the country to prepare students for college and careers. The LAFS are nearly identical to the CCSS for literacy and can be

considered as meeting the requirement of high quality standards related to college and career readiness.

This study included the 39 standards identified by Florida that defined the expectations for the grade 10 ELA course. Four language standards (L.1.1, L.1.2, L.3.4, L.3.5) occur in replicate, nested within different reporting categories to allow for differentiation of language standards within the overarching context of the different strands.

Panelists then conducted individual analyses of 3-5 assessment items from the first test form. For each item, panelists worked individually to assign a DOK level to the item and then to code each item to the standard that they judged the item to measure, i.e., what students are expected to know or do in order to respond to the question. Up to three standards could be coded as corresponding to each item.

Following individual analyses of the items, reviewers participated in a debriefing discussion in which they analyzed the degree to which they had coded particular items or types of content to the standards. This overall process was repeated at the start of each test form to maintain calibration within each group of reviewers. Reviewers then completed analysis of the remaining items individually for each test form.

Reviewers were instructed to focus primarily on the alignment between the LAFS and the assessment items on the Florida test, ACT, and SAT. However, reviewers were encouraged to offer their opinions on the standards or on the assessment tasks by writing a note about the item in the appropriate text box in the WATv2 data collection tool. Reviewers were instructed to enter a note into the WATv2 for an assessment item if the item only corresponded to a part of a standard and not the full standard. Thus, the reviewers' notes can be used to reveal if assessment items only targeted a part of the individual standards. Reviewers also could indicate whether there was a Source-of-Challenge issue with an item—i.e., a technical problem with the item that might cause the student who knows the material to give a wrong answer or enable someone who does not have the knowledge being tested to answer the item correctly.

Reviewers engaged in at least some adjudication of their results after completing the coding of each test form. After all of the reviewers completed coding an assessment form, the study director and group leader identified the assessment items that did not have a majority of reviewers in agreement on DOK or where the reviewers differed significantly on the DOK assigned (e.g., three different DOK values were assigned). When these extreme disagreements occur, it suggests that reviewers are either interpreting the DOK definitions in very different ways or are interpreting the particular assessment item in very different ways.

After discussing an item, the reviewers were given the option to make changes to their codings, but were not required to make any changes if they thought their coding was appropriate. Reviewers also discussed items for which there were great differences in coding to a standard. The adjudication process helped panelists identify and correct any errors in coding (e.g., accidentally assigning an item to the "RI" domain instead of the "RL" domain). Adjudication also helped panelists build familiarity with the standards (e.g., a reviewer might not have noticed that a particular expectation is explicit in one of the standards) as well as build common

interpretation of the standards (e.g., panelists may calibrate their understanding of the meaning of certain standards that may be interpreted in different ways due to ambiguous wording or due to differences in the way people understand the content). Overall, adjudication is intended to ensure that panelists have coded their items as they intended to; reviewers were not required to change their results after the discussion. Reviewer agreement statistics were computed after adjudication and are included in the Findings section of this report.

Reviewers were instructed to consider the full statement of expectations to consider if an assessment item should be mapped to a standard. For a reviewer to code an item to a standard, all or nearly all of the expected outcome as expressed in the standard had to be necessary for a student to perform to answer the item correctly. In some cases, reviewers could make reasonable arguments for a coding an item to different standards. For example, both LAFS.910.RL.2.4 and LAFS.910.L.3.4 include the expectation that students use context clues to identify the meaning of unknown words and phrases. If reviewers map an item to a variety of standards it may also indicate that the assessment task may be inferred to relate to more than one standard but that the item is not a close match. Reviewers may have difficulty finding where an item best fits when an assessment is coded to a set of standards that were not used in developing the assessment. If an item did not closely fit any standard, then the reviewers were instructed to code the item to a standard where there was a partial fit or to a generic standard (the cluster or domain level standard). If the item did not match any of these, then the reviewer was instructed to indicate that the item was uncodeable. No items were considered uncodable on any of the test forms in this review.

Reviewers completed the coding of one form of the Florida grade 10 ELA assessment on the first day. Reviewers continued with the second Florida ELA form on the second day. Then all reviewers moved on to the first ACT form. After the first ACT form was completed, reviewers were divided into smaller groups to maximize the number of reviewers that would be able to code each test form in the time allotted. Reviewers were working at different paces, allowing some reviewers to move on to additional test forms before others were ready. By the end of the three days allotted for coding, ten reviewers had coded both Florida grade 10 ELA forms (spring 2016 and spring 2017). Ten reviewers coded ACT form 74H and seven reviewers completed coding of ACT form 74C. Eight reviewers coded each of the two SAT test forms (April and May 2017).

Data Analysis

To derive the results from the analysis, the reviewers' responses were averaged. First, the value for each of the four alignment criteria is computed for each individual reviewer. Then the final reported value for each criterion is found by averaging the values across all reviewers. Any variance among reviewers was considered legitimate, for example, with the reported DOK level for an item falling somewhere between the two or more assigned values. Such variation could signify differences in interpretation of an item or of the assessed content and/or a DOK that falls in between two of the four defined levels. Any large variations among reviewers in the final results represented true differences in opinion among the reviewers and were not because of coding error. These differences could be due to different standards targeting the same content knowledge or may be because an item did not explicitly correspond to any standard, but could be inferred to relate to more than one standard. Standard deviations are reported in

the tables provided in **Appendix 1b.B**, which give one indication of the variance among reviewers.

The results produced from the institute pertain only to the issue of alignment between the Language Arts Florida Standards and the six assessments that were analyzed. Note that an alignment analysis of this nature does not serve as external verification of the general quality of the standards or assessments. Rather, only the degree of alignment is discussed in the results. For these results, the means of the reviewers' coding were used to determine whether the alignment criteria were met.

Alignment Criteria Used for This Analysis

This report describes the results of an alignment study of six assessments with the LAFS for grade 10 ELA. The study addressed specific criteria related to the content agreement between the standards and assessments. Four criteria received major attention:

- Categorical Concurrence,
- Depth-of-Knowledge Consistency,
- Range-of-Knowledge Correspondence, and
- Balance of Representation.

Details on the criteria and indices used for determining the degree of alignment between standards and assessments are provided below. For each alignment criterion, an acceptable level was defined by what would be required to assure that a student had reasonably met the expectations within the reporting categories for each discipline. In the descriptions below, the words "domain" and "reporting category" are used to describe reporting levels. In this analysis, the reporting categories for ELA were Key Ideas and Details (RC1); Craft and Structure (RC2); Integration of Knowledge and Ideas (RC3); Language and Editing (RC4); and Text-Based Writing (RC5). In the descriptions below, the term "standards" may be used as an umbrella term, to refer to expectations in general. In addition to judging alignment between reporting categories and assessments on the basis of the four key alignment criteria, information is also reported on the quality of items by identifying items with Source-of-Challenge and other issues.

Categorical Concurrence
An important aspect of alignment between standards and assessments is whether both address the same content categories. The Categorical-Concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of Categorical Concurrence between standard and assessments is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each reporting category. The analysis assumed that the assessment had to have at least six items for measuring content from a reporting category in order for a minimum acceptable level of Categorical Concurrence to exist between the domain and the assessment. The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Of course, many factors must be considered in determining what a reasonable number is, including the reliability of the subscale, the mean score, and cutoff

score for determining mastery. Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and that the reliability of one item is 0.1, it was estimated that six items would produce an agreement coefficient of at least 0.63. This indicates that about 63% of the group would be consistently classified as masters or non-masters if two equivalent test administrations were employed. The agreement coefficient would increase to 0.77 if the cutoff score is increased to one standard deviation from the mean and, with a cutoff score of 1.5 standard deviations from the mean, to 0.88.

Usually states do not report student results by domains or require students to achieve a specified cutoff score on expectations related to a domain. If a state did do this, then the state would seek a higher agreement coefficient than 0.63. Six items were assumed as a minimum for an assessment measuring content knowledge related to a reporting category, and as a basis for making some decisions about students' knowledge of that content under the reporting category. If the mean for six items is 3 and one standard deviation is one item, then a cutoff score set at 4 would produce an agreement coefficient of 0.77. Any fewer items with a mean of one-half of the items would require a cutoff that would only allow a student to miss one item. This would be a very stringent requirement, considering a reasonable standard error of measurement on the subscale.

Depth-of-Knowledge Consistency

Standards and assessments can be aligned not only on the category of content covered by each, but also on the basis of the complexity of knowledge required by each. *Depth-of-Knowledge Consistency between standards and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.* For consistency to exist between the assessment and the reporting categories, as judged in this analysis, at least 50% of the items corresponding to a reporting category had to be at or above the depth-of-knowledge level of the corresponding content expectation. The 50% level, a conservative minimum cutoff point, is based on the assumption that a minimal passing score for any one reporting category of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the content expectations within the corresponding reporting categories. For example, assume an assessment included six items related to one domain and students were required to answer correctly four of those items to be judged proficient—i.e., 67% of the items. If three, 50%, of the six items were at or above the depth-of-knowledge level of the corresponding expectations, then for a student to achieve a proficient score would require the student to answer correctly at least one item at or above the depth-of-knowledge level of one expectation. If a domain had between 40% and 50% of items at or above the depth-of-knowledge levels of the expectations, then it was reported that the criterion was "weakly" met.

<u>DOK Levels for Reading</u>

Interpreting and assigning depth-of-knowledge levels to both standards and assessment items is an essential requirement of alignment analysis. These descriptions help to clarify what the different levels represent in mathematics.

Level 1 (Recall)
DOK 1 involves reading text orally and with basic comprehension, decoding words, blending phonemes, receiving and reciting facts, demonstrating letter and word knowledge, and recognizing text features and common spelling patterns. DOK 1 also includes receiving or reciting facts acquired by processing text as well as reading orally without the analysis of text. Very basic comprehension of a text gained from knowledge of vocabulary and explicit structure of the text is at this category. Tasks require only a shallow understanding of the text presented and often consist of verbatim recall from text, slight paraphrasing of specific details from the text, or simple understanding of a single word or phrase. Younger students who answer direct questions about features stated explicitly in the text are performing at this category. Applying phonics and word analysis skills in decoding words are also DOK 1 tasks. Some examples that represent, but do not constitute all of, DOK 1 performance include:

- Support ideas with reference to verbatim (or only slightly paraphrased) details from the text.
- Use a dictionary to find the meanings of words.
- Recognize figurative language in a reading passage.

Level 2 (Skills and Concepts)
DOK 2 involves drawing meaning from text by using organizational structure, evidence, and context; summarizing main ideas, character traits, plots, themes, and figurative use of words; following cause-effect sequences and multiple ideas through a text; distinguishing among hypotheses and givens as well as fact from opinion; and explaining differences among genres (poetry, expository materials, fiction, etc.). DOK 2 requires the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Inter-sentence analysis or inference is required. DOK 2 tasks may require use of specific information from the text to explain given events and ideas. At this level, reading concepts (e.g., making inferences or predictions) are generally applied for purposeful reading. Multiple features of the text are processed to gain a deeper understanding of the text such as organizing in a time sequence, outlining, comparing fact from opinion, and using graphic aides. Deciphering main ideas supported by key details or drawing on details to describe a feature in a story are stressed. Younger students conveying important points from a story fit under this category. DOK 2 ideas, in general, apply the skills and concepts that constitute DOK 1. However, DOK 2 activities involve closer understanding of text, possibly through paraphrasing, such as putting in one's own words both the question and response to an assessment item. Some examples that represent, but do not constitute all of, DOK 2 performance include:

- Use context cues to identify the meaning of unfamiliar words, phrases, and expressions that could otherwise have multiple meanings.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Level 3 (Strategic Thinking)
DOK 3 involves conducting analyses of the text to make inferences about author's purpose and use of textual features (e.g., literary devices to support and convey the main message); engaging

in critical reading to attest to the credibility of the message, the internal logic, and implied values, attitudes, and biases; and going beyond the text by comparing features and meaning with other texts, considering the impact of the time period and other conditions when the text was written, and raising valid alternative hypotheses and conclusions to those presented in the text. At DOK 3, deep knowledge becomes a greater focus. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas while applying reasoning and planning. Students must be able to support their thinking. Younger students who provide some valid evidence for their breakdown of a story into meaningful parts are performing at this category.

Tasks at a Category 3 may involve abstract theme identification, inference across an entire passage with multiple paragraphs, or students' application of prior knowledge. Activities may also involve identifying more abstract connections between texts. Some examples that represent, but do not constitute all of, DOK 3 performance include:

- Explain or recognize how the author's purpose affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Level 4 (Extended Thinking)
DOK 4 involves at least as complex content as in the previous category, but also requires working on a task over an extended period of time such as when conducting a research project over a period of weeks. The extended time that accompanies this type of activity allows for creation of original work and requires metacognitive awareness that typically increases the complexity of a DOK 4 task overall, in comparison with DOK 3 activities. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking.

DOK 4 activities may have students take information from multiple passages and texts to find supporting evidence and counter points for developing an argument or reaching conclusions or could involve creating an original thesis on a topic based on information drawn from relevant references. For younger students, an extended period of time could be multiple days for reaching conclusions from reading a number of texts. Students take information from a multiple of passages and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts requiring work over an extended period of time. Some examples that represent, but do not constitute all of, DOK 4 performance include:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Range-of-Knowledge Correspondence

For reporting categories and assessments to be aligned, the breadth of knowledge required on both should be comparable. *The Range-of-Knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a reporting category is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.* The criterion for correspondence between span of knowledge for a reporting category and an assessment considers the number of standards within the reporting category with one related assessment item/activity. Fifty percent of the standards for a reporting category must have at least one related assessment item for the alignment on this criterion to be judged acceptable. This level is based on the assumption that students' knowledge should be tested on content from over half of the domain of knowledge for a reporting category. This assumes that each expectation for a reporting category should be given equal weight. Depending on the balance in the distribution of items and the need to have a low number of items related to any one expectation, the requirement that assessment items need to be related to more than 50% of the expectations for a reporting category increases the likelihood that students will have to demonstrate knowledge on more than one expectation per reporting category to achieve a minimal passing score. As with the other criteria, a state may choose to make the acceptable level on this criterion more rigorous by requiring an assessment to include items related to a greater number of the expectations. However, any restriction on the number of items included on the test will place an upper limit on the number of expectations that can be assessed. Range-of-Knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of reporting categories and a large number of expectations. If 50% or more of the objectives for a reporting category had a corresponding assessment item, then the range-of-knowledge correspondence criterion was met. If between 40% and 50% of the objectives for a reporting category had a corresponding assessment item, the criterion was "weakly" met.

Balance of Representation

In addition to comparable depth and breadth of knowledge, aligned reporting categories and assessments require that knowledge be distributed equally or proportionally in both. The Range-of-Knowledge criterion only considers the number of expectations with at least one assessment item within a reporting category; it does not take into consideration how the assessment items/activities are distributed among these expectations. *The Balance-of-Representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. This index only considers the expectations for a reporting category that has at least one related assessment item per expectation. The index is computed by considering the difference in the proportion of expectations and the proportion of items assigned to the expectation. An index value of 1 signifies perfect balance and is obtained if the corresponding items related to a reporting category are equally distributed among the expectations for the given reporting category. Index values that approach 0.0 signify that a large proportion of the items assess only one or two of all of the expectations that were measured. Depending on the number of expectations and the number of items, a unimodal distribution (most items related to one expectation and only one item related to each of the remaining expectations) has an index value of less than 0.5. A bimodal distribution has an index value of around 0.55 or 0.6. Index values of 0.7 or higher indicate that items/activities are distributed among all of the expectations at least

to some degree (e.g., nearly every expectation has at least two items) and is used as the acceptable level on this criterion. Index values between 0.6 and 0.7 indicate the Balance-of-Representation criterion has only been "weakly" met.

Source-of-Challenge Criterion
The Source-of-Challenge criterion is used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted language reporting category or expectation (i.e., construct irrelevance). Bias and sensitivity issues as well as technical issues and error could all be reasons for an item to have a Source-of-Challenge problem. Such item characteristics may result in some students not answering an assessment item, or answering an assessment item incorrectly, or at a lower level, even though they possess the understanding and skills being assessed.

Cutoffs for Alignment Criteria
For overall alignment, an assessment form is reported as *fully aligned* if no items need replacement to meet the conditions for all of the criteria described above. A test form is considered *acceptably aligned* if it needs between one and five items replaced or revised in order to meet the conditions for all alignment criteria. A test form is reported to *need slight adjustments* if six to ten items need to be replaced or revised to meet the criteria and is reported to *need major adjustments* if more than ten items need to be replaced or revised. These categories represent typically used cutoff levels.

Findings

Framework Analysis for ELA
Dr. Quast's framework analysis identified similarities and differences in the overall structures of the three assessments included in this study. The framework analysis found that the ACT College and Career Readiness Standards had a close match to only 28% of the LAFS while the SAT Skills had a close match to 56% of the LAFS. The standards/skills for both ACT and SAT were found to have at least a partial match to around 75% of the LAFS used for the grade 10 English assessment. Both the ACT and SAT also included assessment targets outside of the content within the LAFS, such as identification of subject-verb agreement, pronoun-antecedent agreement, inappropriate shifts in verb tense and other expectations.

A comparison of item counts and session times are provided in Table 1b-12. While all three assessment frameworks included standards/skills and measurements for reading and language domains, the Florida Grade 10 Statewide Standardized ELA Assessment included reading and language within one test, whereas the ACT and SAT had separate tests for each domain. Therefore, the ACT and SAT had greater total numbers of items and a higher percentage of items focused on specific language content. The Florida ELA test forms contained fewer test items than the ACT or SAT but were administered in longer sessions. Additionally, whereas the ACT and SAT tests consisted of passage-based multiple-choice items, between a quarter and a half of the Florida Grade 10 Statewide Standardized ELA Assessment used technology-enhanced items (TEIs). In terms of passages, the Florida Grade 10 Statewide Standardized ELA Assessment, ACT, and SAT were comparable in the percentage of literary and informational texts included. Text complexity was described differently for each assessment, preventing a

direct comparison. The Florida Grade 10 Statewide Standardized ELA Assessment set the complexity of texts as "accessible for tenth grade." Complexity of texts at benchmark score for the ACT was "somewhat challenging," and for the SAT was "complex."

As shown in Table 1b-12, the ACT had the most items (115), SAT had slightly fewer (96), and the Florida grade 10 ELA assessment had the fewest number of items (60-64).

Table 1b-12. Comparison of Florida Grade 10 Statewide Standardized ELA Assessment, ACT, & SAT Item Counts, Types, and Session Times

	Florida Reading, Listening, and Language Test	**ACT English and Reading Tests**	**SAT Reading and Writing and Language Tests**
Assessment Time	180 total minutes (over 2 days)	80 total minutes Reading: 35 min English: 45 min	100 total minutes Reading: 65 min Writing/Language: 35 min
Number of Items	54-58 operational + 6 field test	115 items combined Reading: 40 English: 75	96 items combined Reading: 52 Writing/Language: 44
Type of Items	Passage-based multiple choice & TEIs (25-50%)	Passage-based multiple choice	Passage-based multiple choice

Because the Florida ELA test forms contained fewer test items than the ACT or SAT but were administered in longer sessions, students would have significantly more time per item (2.8 minutes/item versus between 0.8 and 1.3 min/item) to read passages and answer items than the corresponding ACT and SAT tests. Average time per item/task is shown in Table 1b-13 below.

Table 1b-13. Time per assessment item/task for Florida Grade 10 Statewide Standardized ELA Assessment, ACT, and SAT ELA tests

Test	**Number of Items**	**Assessment Time**	**Average Time per Item**
Florida	54-58 (operational) + 6 field test	180 min	2.8 min
ACT Reading	40	35 min	0.9 min
ACT English	75	45 min	0.6 min
SAT Reading	52	65 min	1.3 min
SAT Writing/Language	44	35 min	0.8 min

Table 1b-14 provides an overview of the content and context between the Florida Grade 10 Statewide Standardized ELA Assessment essay, ACT essay, and the SAT essay. A significant difference was the amount of time allocated to student writing, with the Florida Grade 10 Statewide Standardized ELA Assessment allowing over three times more time than the ACT allowed and over twice the time that the SAT allowed. Additionally, the number of passages

provided in the stimulus varied across the three assessments. The nature of the essays was also slightly different: The Florida Grade 10 Statewide Standardized ELA Assessment task required either an explanatory or argumentative essay, the ACT task required an argumentative essay, and the SAT task required a written analysis of a source text.

Table 1b-14. Comparison of Florida Grade 10 Statewide Standardized ELA Assessment, ACT, & SAT Essays

	Florida Writing Essay	**ACT Writing Essay (optional)**	**SAT Writing Essay (optional)**
Time	120 minutes	35 minutes	50 minutes
Item Format	Stimulus (up to 4 passages) & 1 prompt	Stimulus (issue description & two texts providing two different viewpoints) & 1 prompt	Single source text & 1 prompt
Essay Type	Explanatory or Argumentative	Argumentative	Written Analysis of Source text
Domains	Purpose, Focus, & Organization Evidence & Elaboration Conventions of Standard English	Ideas & Analysis Development & Support Organization Language Use	Reading Analysis Writing

There were no field test items on any of the ELA assessments and no items were excluded from the ELA analysis. On all test forms, all items except for the writing prompt were weighted as one point. The Florida writing prompt was weighted at 10 points, distributed across a three-part rubric. The ACT essay was weighted at 12 points, total, distributed evenly across a four-part rubric. The SAT was weighted at 24 points, reflective of a 12-point maximum score from each of two raters, distributed evenly across a three-part rubric.

Standards

Study results are reported according to the five reporting categories (RCs) that included a total of 39 LAFS, as shown in Table 1b-14. Four language standards (L.1.1, L.1.2, L.3.4, L.3.5) occur in replicate, nested within different reporting categories to allow for differentiation of language standards within the overarching context of the different strands. For all but three of the standards, the DOK levels for the LAFS assigned by the state were used as the DOK levels in this study (http://www.fldoe.org/core/fileparse.php/5390/urlt/0081014-lafs.pdf). Reviewers flagged, discussed, and decided to change the assigned DOK levels for the following three standards: LAFS.910.RI.1.2, LAFS.910.RI.3.7, and LAFS.910.RI.3.9. Each of these standards was assigned a DOK 2 within state documents but reviewers assigned a DOK 3.

A summary of the levels of complexity within the LAFS is also given in Table 1b-15. None of the standards included in the study were considered DOK 1. In other words, assessment targets did not include recall or reproduction of memorized information or routine tasks. Twenty-one percent of the standards were considered a DOK level 2, emphasizing work that involves both comprehension and subsequent processing of text, as well as making basic inferences from text

and using specific information from text to explain events and ideas. The majority of standards (72 percent) were considered to be DOK 3, emphasizing expectations for deep analysis of text and abstract thinking, including making holistic inferences based on text, and engaging in critical reading to consider aspects of author's purpose and use of textual features. Three standards (8 percent) were considered DOK 4. A DOK 4 expectation is one that is both at least as complex as a DOK 3 but also requires extended time—days, weeks, or months—to complete. Although some components of these DOK 4 standards may be reasonably assessed by on-demand assessments, DOK 4 standards should not be expected to be fully assessed by an on-demand test.

Table 1b-15. Expectations by Depth-of-Knowledge (DOK) Levels for Language Arts Florida Standards, October, 2017

ELA	Total Number of Expectations	DOK Level	Number of Standards by Level	Percent within RC by Level
RC1 Key Ideas and Details	6	2 3	2 4	33.33 66.67
RC2 Craft and Structure	10	2 3	2 8	20.00 80.00
RC3 Integration of Knowledge and Ideas	8	3	8	100.00
RC4 Language and Editing	2	2 3	1 1	50.00 50.00
RC5 Text-Based Writing	13	2 3 4	3 7 3	23.08 53.85 23.08
Total	39	2 3 4	8 28 3	21 72 8

Mapping of Items to Standards

If no particular grade-level standard is targeted by a given assessment item, reviewers were instructed to code the item at the cluster, strand, or domain level. This coding to a "generic standard" sometimes indicates that the item is inappropriate for a particular grade level (for example, the item might better match a standard from another grade level). If the item is grade-appropriate and an appropriate standard was not found, then this situation may instead indicate that there is a part of the content within the standards that is being interpreted differently by different parties. These items may highlight areas in the standards that state representatives and test developers need to discuss to ensure common interpretation. These items may then be revised to ensure that they target specific on-grade standards. Generic coding may also occur when mapping a test to a set of standards that is different from the set used to develop the test. In this case, some items on an assessment may simply target a different set of learning expectations.

Table 1b-16 shows the items for each assessment that more than one reviewer coded to a generic standard. This table shows the generic standard to which the item was coded, the number of reviewers who coded the item to the generic standard, and a summary of the reasons for the coding. No items from the Florida grade 10 ELA test were mapped to generic standards, suggesting that the Florida test is, overall, a close match with the corresponding standards. On the ACT, across both forms analyzed, reviewers were unable to find a precise standard match for 11 items (~10% of total items). On the SAT across both forms analyzed, reviewers were unable to find a precise standard match for 10 items (~11% of total items) suggesting that some of the content on both the ACT and SAT was not represented in the corresponding standards. Reviewers were required to write an explanation in the case of assigning an item to a generic standard. These notes can be found in **Appendix 1b.C**. Items assigned to generic standards by more than one reviewer should be reviewed. There were two main reasons that reviewers coded items to generic standards. [Information subject to nondisclosure agreements has been omitted for public release.]

Table 1b-16 Items Assigned to Generic Content Expectations by Assessment by More than One Reviewer for the Florida Alignment Analysis, October 2017

ELA Test/Form	Generic Content Expectation	Item Number (# of Reviewers)	Reason
ACT 47H	3.2	83(8)	[Information subject to nondisclosure agreements has been omitted for public release.]
	3.2	85(9)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	119(6)	[Information subject to nondisclosure agreements has been omitted for public release.]
	5.1	119(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
ACT 47C	3.1	83(5)	[Information subject to nondisclosure agreements has been omitted for public release.]
	3.1	84(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	3.1	85(4)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	28(2)	[Information subject to nondisclosure agreements has been omitted for public release.]

	4.1	56(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	58(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	119(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
SAT April 2017	3.2	19(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	3.2	39(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	3.2	40(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	63(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	76(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
SAT May 2017	4.1	67(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	77(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	81(3)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	94(2)	[Information subject to nondisclosure agreements has been omitted for public release.]
	4.1	95(3)	[Information subject to nondisclosure agreements has been omitted for public release.]

Although total item counts differed significantly (ACT – 115 items, SAT – 96 items, Florida grade 10 ELA – 60-64 items) the total *percentage* of LAFS that each test targeted was about the same for each assessment (see Table 1b-17).

Table 1b-17. Number and Percent of Language Arts Florida Standards with at least One Item Found by a Majority of Reviewers as Corresponding

Assessment	Number of Items (including writing prompt)	Number of LAFS Standards Targeted	Percentage of Total LAFS Standards with at least One Corresponding Assessment Item
Florida Spring 2016	65	21	54%
Florida Spring 2017	65	23	59%
ACT Form 74H	116	19	49%
ACT Form 74C	116	20	51%
SAT Apr 2017	97	16	41%
SAT May 2017	97	23	59%

Comparison of Overall DOK Distribution

A comparison of the overall DOK distribution for each assessment, averaged across the two test forms is shown in Table 1b-18. The SAT test forms had zero items that reviewers considered a DOK 1. The Florida assessment had the greatest percentage of DOK 3 items, followed by the SAT, and the ACT had the fewest DOK 3 items. Thus, the Florida Grade 10 ELA assessment was found to have the most items with the highest content complexity assessable in an on-demand assessment. If both DOK 2 and DOK 3 items are considered, the SAT has the greatest number of higher-DOK items (100%) followed by the Florida Grade 10 ELA assessment (88%) and then the ACT (76%). Both the Florida Grade 10 ELA assessment and the ACT included DOK 1 items. Because none of the LAFS were considered DOK 1, it may be argued that there should be no DOK 1 items on an exam that assesses the grade 10 set of LAFS.

On the ACT, the distribution of the items by DOK levels varies some from the intended distribution of DOK levels for the ACT forms, according to the ACT Technical Manual. The English test blueprint specifies 33-41% DOK 3 items and the Reading test blueprint specifies 25-50% DOK 3 items while reviewers coded only 14% of all items as DOK 3. This suggests that the DOK definitions used by ACT are different from the original Webb definitions, which have been updated and used in this study. The intended DOK definitions for SAT were not provided.

Table 1b-18. DOK Distribution, averaged on two test forms for FL Grade 10 ELA, ACT, and SAT

Test	DOK 1	DOK 2	DOK 3
FL Grade 10 ELA	12%	54%	34%
ACT	24%	62%	14%
SAT	0%	80%	20%

Alignment of Florida Grade 10 ELA Statewide, ACT, and SAT Assessments Forms with Language Arts Florida Standards

Overall alignment results are summarized in Table 1b-19 below and then detailed for each test form in the pages that follow. For Reporting Categories 1-4, the FL test forms were found to have acceptable alignment with the Language Arts Florida Standards, the ACT forms were found to need major adjustments, and one SAT form was found be acceptably aligned while the other needed slight adjustments. For full alignment, each Florida test form would need five

items revised or replaced, the ACT test forms would need 12 or 17 items replaced, and the SAT test forms would need five or seven items replaced. These findings are shown in Table 1b-19.

Table 1b-19. Overall Alignment Findings for Two Forms Each of Florida Grade 10 ELA, ACT, and SAT Assessments

Test Form	Alignment Findings	Number of Items that Need Revision/Replacement for Full Alignment
Florida Grade 10 Spring 2016	Acceptably Aligned	5
Florida Grade 10 Spring 2017	Acceptably Aligned	5
ACT 74H	Needs Major Adjustments	17
ACT 74C	Needs Major Adjustments	12
SAT April 2017	Acceptably Aligned	5
SAT May 2017	Needs Slight Adjustments	7

All test forms had an alignment issue with the DOK Consistency criterion for Reporting Category 4, Language and Editing.

Results by Test Form

The results of the analysis for each of the four alignment criteria are provided in Tables 1b-20.1 to 1b-20.3 for each ELA test form for Reporting Categories 1-4. More detailed data on each of the criteria are given in **Appendix 1b.B**, in the first three tables for each test form. The reviewers' notes and debriefing comments (**Appendices 1b.C** and **1b.D**) provide further detail about the individual reviewers' impressions of the alignment. Some reviewer comments are summarized in the results reported below.

In Tables 1b-20.1 to 1b-20.3, "YES," indicates that an acceptable level was attained between the assessment and the reporting category on the criterion. "WEAK" indicates that the criterion was nearly met, within a margin that could simply be due to error or reasonable variation in reviewer coding. "NO" indicates that the criterion was not met by a noticeable margin—10% under an acceptable level for Depth-of-Knowledge Consistency, 10% under an acceptable level for Range-of-Knowledge Correspondence, and 0.1 under an index value of 0.7 for Balance of Representation.

Florida Statewide Standardized Grade 10 ELA Assessment Alignment Study Results

Both Florida grade 10 ELA assessments for Reporting Categories (RC) 1-4 were found to be acceptably aligned with the LAFS. The only alignment issue for both test forms was a lack of DOK consistency for RC4 (Language and Editing). For full alignment, both test forms would need five items revised or replaced to meet DOK consistency for RC4. Although both test forms yielded similar alignment results, four reviewers noted in their debriefing comments that they thought the Spring 2017 form was better aligned than the Spring 2016 form. This impression is reflected to some extent in the alignment statistics for RC2 and RC3 in Tables 1b-20.1a and 1b-20.1b but may also reflect other qualitative aspects of the assessment content and prompts. For

example, one reviewer commented that the items on the Spring 2017 form had a "cleaner match" to the standards than did the items on the 2016 form.

Table 1b-20.1a and **1b**-20**.1b** *Summary of Alignment Statistics and Findings for Florida Statewide Standardized Grade 10 ELA Assessment Forms and LAFS*

Table 1b-20.1a. Florida Grade 10 ELA Spring 2016 Form

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	23	53%	100%	0.75	YES	YES	YES	YES
RC2 CS	19	50%	64%	0.83	YES	YES	YES	YES
RC3 IKI	7	66%	54%	0.85	YES	YES	YES	YES
RC4 LE	15	24%	100%	0.87	YES	NO	YES	YES

Table 1b-20.1b. Florida Grade 10 ELA Spring 2017 Form

	Alignment Statistics				**Alignment Findings**			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	21	54%	96%	0.77	YES	YES	YES	YES
RC2 CS	21	61%	71%	0.76	YES	YES	YES	YES
RC3 IKI	12	80%	63%	0.81	YES	YES	YES	YES
RC4 LE	11	14%	100%	0.89	YES	NO	YES	YES

*Number of items

ACT ELA Results

Both ACT test forms were considered to need major adjustments for RCs 1-4 to be aligned with the LAFS. Across both forms, the main alignment issue for RC2 (Craft and Structure) and RC4 (Language and Editing) was unmet or weak DOK consistency. This means that although the test forms included items that related to the standards within RC2 and RC4, the items were, overall, considered lower complexity than the level of complexity that is expected in the LAFS. Only 30% or 46% of the items that targeted RC2 content and only 15% or 27% of the items that targeted RC4 content were found to match the complexity expected by the corresponding LAFS. Across both forms, the main alignment issue for RC3 (Integration of Knowledge and Ideas) was unmet Categorical Concurrence and unmet Range-of-Knowledge. Unmet Categorical Concurrence means that there were not enough items that targeted standards from within RC3 to make a reliable inference about student mastery of the content based on test results. The breadth of knowledge expected within RC3 was also not represented on either test form.

For full alignment, Form 74H would need 17 items revised or replaced and Form 74C would need 12 items revised or replaced. In their debriefing notes, a number of reviewer comments suggest that reviewers had a harder time matching the ACT items to the standards compared with the Florida items.

Table 1b-20.2a and **1b-20.2b** *Summary of Alignment Statistics and Findings for ACT Test Forms and LAFS*

Table 1b-20.2a. ACT Form 74H December 2016 – ELA

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	26	59%	51%	0.68	YES	YES	YES	WEAK
RC2 CS	13	30%	42%	0.74	YES	NO	WEAK	YES
RC3 IKI	4	73%	22%	0.92	NO	YES	NO	YES
RC4 LE	34	15%	93%	0.90	YES	NO	YES	YES

For ACT Form 74H, RC2, three items would need to be revised or replaced to meet the DOK Consistency criterion. If one of these items targeted a standard that was not yet targeted, the weakness in Range of Knowledge could also be resolved. For RC3, two items would need to be revised or replaced to meet Categorical Concurrence. If both of these items targeted standards that are not currently assessed, the Range-of-Knowledge criterion could also be met. For RC4, 12 items would need to be revised or replaced to meet DOK Consistency.

Table 1b-20.2b. ACT Form 74C June 2017 – ELA

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	25	64%	85%	0.68	YES	YES	YES	WEAK
RC2 CS	14	46%	66%	0.74	YES	WEAK	YES	YES
RC3 IKI	2	75%	16%	0.88	NO	YES	NO	YES
RC4 LE	25	27%	83%	0.90	YES	NO	YES	YES

*Number of items

For ACT Form 74C, RC2, one item would need to be revised or replaced to meet the DOK Consistency criterion. For RC3, four items would need to be revised or replaced to meet Categorical Concurrence. If each of these four items targeted a separate standard that is not currently assessed, the Range-of-Knowledge criterion could be met. For RC4, seven items would need to be revised or replaced to meet DOK Consistency.

SAT ELA Results

One SAT form was found be acceptably aligned while the other needed slight adjustments for RCs 1-4 to be fully aligned with the LAFS. For full alignment, Form April 2017 would need 5 items revised or replaced and Form May 2017 would need 7 items revised or replaced. Across both forms, the main alignment issue for RC3 (Integration of Knowledge and Ideas) was unmet Range of Knowledge. This means that the breadth of knowledge expected within RC3 was not represented on either test form. Across both forms, the main alignment issue for RC4 (Language and Editing) was unmet DOK Consistency for RC4. This means that although the test forms included items that related to the standards within RC4, the items were, overall, considered lower complexity than the level of complexity that is expected in the LAFS. Only 21% or 25% of the items that targeted RC4 content were found to match the complexity expected by the corresponding LAFS.

Reviewer debriefing notes suggest that reviewers found it relatively easy to match the SAT items to the standards. Reviewers commented that there were more informational texts on the SAT compared with the Florida ELA assessment. According to the assessment frameworks, the FSA test is intended to have 30% informational passages compared with 20% on the SAT.

Table 1b-20.3a and **1b-20.3b** *Summary of Alignment Stats and Findings - SAT Test Forms and LAFS*

Table 1b-20.3a. SAT Form April 2017 – ELA

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	34	75%	85%	0.65	YES	YES	YES	WEAK
RC2 CS	15	46%	50%	0.73	YES	WEAK	YES	YES
RC3 IKI	14	91%	37%	0.65	YES	YES	NO	WEAK
RC4 LE	10	25%	50%	0.88	YES	NO	YES	YES

For SAT Form April 2017, RC2, one item would need to be revised or replaced to meet the DOK Consistency criterion. For RC3, one item would need to be revised or replaced to target a standard that is not currently assessed in order to meet the Range-of-Knowledge criterion. For RC4, three items would need to be revised or replaced to meet DOK Consistency.

Table 1b-20.3b. SAT Form May 2017 – ELA

	Alignment Statistics				Alignment Findings			
	CC*	DOK %	Range	Balance	CC	DOK	Range	Balance
RC1 KID	29	72%	93%	0.60	YES	YES	YES	WEAK
RC2 CS	23	56%	62%	0.71	YES	YES	YES	YES
RC3 IKI	13	89%	22%	0.79	YES	YES	NO	YES
RC4 LE	14	21%	75%	0.81	YES	NO	YES	YES

*Number of items

For SAT Form May 2017, RC3, three items would need to be revised or replaced to target standards that are not currently assessed in order to meet the Range-of-Knowledge criterion. For RC4, four items would need to be revised or replaced to meet DOK Consistency.

Text-Based Writing Reporting Category 5

Each assessment included a single weighted writing prompt that was evaluated according to a three-part or four-part rubric. The writing prompts for all test forms, along with the corresponding rubrics, were considered to target corresponding writing standards at an appropriate level of complexity. Although all writing prompts were considered reasonably aligned, there are some differences between the different tests. The Florida essay may be argumentative or explanatory, depending on the prompt, so may target W.1.1, a complex standard that expects students to write a robust argument, or W.1.2, a similarly complex standard that expects students to write informative/explanatory texts. The Florida essay was also coded to W.2.4 and W.2.5, which emphasize production and revision of clear and focused writing. The ACT essay is argumentative, addressing W.1.1 as well as W.2.4 and W.2.5. The SAT essay was coded to W.1.2 and 2.4, similar to the other two assessments, but also was coded to

W.3.8 and W.3.9, which relate to gathering information or evidence from other texts, integrating the information, and using it to support an analysis. This reflects the structure of the SAT essay, which is a written analysis of source text. Reviewers commented on the significant difference in the allotted task time on each assessment: 120 minutes on the FSA essay compared with just 35 minutes on the ACT and 50 minutes on the SAT. Reviewers noted that limited time affords less of an opportunity to meet the full depth of some of the expectations within the Text-Based Writing reporting category.

Source of Challenge Issues and Reviewers' Comments
Reviewers were instructed to document any Source-of-Challenge issue and to provide any other comments they may have about an item. A Source-of-Challenge is a technical issue with an item that can result in a student answering the item correctly or incorrectly for the wrong reason. There were no items on the Florida test forms for which more than one reviewer left a Source-of-Challenge comment. There was one item (#60) on ACT Form 74C for which two reviewers commented that there could be two plausible answers. There were no items on the SAT forms for which more than one reviewer left a Source-of-Challenge comment.

Reviewers also wrote notes about many items on each form. Some notes indicate when only part of a particular standard was targeted by an assessment task. These notes also include general comments as well as indicate concerns with items. Some notes include suggestions for resolutions to issues identified. After coding each assessment form, reviewers were asked to respond to four debriefing questions. Reviewers Comments and Source-of-Challenge notes can be found in **Appendix 1b.C**.

Reliability among Reviewers
Reviewers engaged in some adjudication of their data after all reviewers finished their coding for an assessment. These discussions were used to identify any mistakes in coding. Reviewers were not required to change their coding after discussion unless they found a compelling reason. The agreement statistics shown in Table 1b-21, on the following page, were computed after adjudication. If the intraclass correlation and pairwise agreements are low after adjudication, then this could be an indication of a misfit between the standards and the assessment items. Reviewers will vary in their codings the more they have difficulty in finding a precise match between an assessment item and the standards. The overall intraclass correlation among the seven-to-ten ELA reviewers' assignment of DOK levels to items was high (0.86 or higher) for all six analyses (Table 1b-21). An intraclass correlation value greater than 0.8 generally indicates a high level of agreement among the reviewers.

A pairwise comparison was used to determine the degree of reliability of reviewers coding at the reporting category level and the standard level. The pairwise comparison was computed by considering for each item the coding assigned by each reviewer compared to the coding by each of the other six-to-nine reviewers. For example, for 10 reviewers a total of 45 comparisons were computed for each item. For most alignment studies, the standards pairwise agreement is higher than 0.6. The pairwise agreement for assigning standards to items was greater than 0.6 for both Florida test forms but lower (between 0.36 and 0.51) for ACT and SAT test forms. This low pairwise agreement is much lower than what is normally observed in an alignment analysis

and suggests that the items on the ACT and SAT test forms were harder to match to the LAFS standards than were the items on the Florida assessment.

Another factor that contributed to the low pairwise agreement was that there are some standards that have overlap in content as well as some standards that were included in multiple reporting categories. For example, within RC2 (Craft and Structure), both LAFS.910.RL.2.4 and LAFS.910.L.3.4 (and also for the corresponding informational text standards) include the expectation that students use context clues to identify the meaning of unknown words and phrases. Reviewers were divided between these standards for five items on each Florida test form (#2, 3, 8, 27, 42 for Spring 2016 and #1, 17, 34, 53, 59 for Spring 2017). This same divide occurs three times (#82, 92, 113) for ACT Form 74H, twice (# 94, 103) on ACT Form 74C, eight times (#3, 10, 15, 25, 28, 36, 48, 51) on SAT April 2017 and nine times (#3, 14, 16, 26, 28, 37, 38, 42, 45) on SAT May 2017. The SAT Reading test specifically identifies "Words in Context" as an assessment target, which explains the greater number of items on the SAT for which reviewers were split on the standard relating to understanding words in context.

For coding to the level of reporting category, a pairwise agreement of 0.90 is desired. For all test forms, pairwise agreement for reporting category is lower than what is normally observed in an alignment analysis. Agreement was reasonably high for the Florida test forms (0.85 for both test forms) but lower for both the ACT and SAT (between 0.71 and 0.76). The lower reviewer agreement for reporting category coding for the ACT and SAT suggests that reviewers had a harder time deciding where the test items best fit. There was also some struggle with the structure of the standards. For example, standards L1.1 and L1.2 and WL1.1 and WL 1.2 are in different reporting categories (RC4: Language and Editing and RC5: Text-Based Writing) and caused some coding disagreement, centering around the words "when writing" in the Language standards. Reviewers had a hard time differentiating between these standards, interpreting components of standards from both reporting categories to relate to similar writing tasks.

Table 1b-21. Intraclass and Pairwise Comparisons, Florida Alignment Analysis, FL ELA, ACT, and SAT with LAFS

Test Form	Intraclass Correlation (DOK)	Pairwise Comparison (DOK)	Pairwise Comparison (Reporting Category)	Pairwise Comparison (Standards)
Florida Grade 10 Spring 2016	0.94	0.64	0.85	0.63
Florida Grade 10 Spring 2017	0.94	0.62	0.85	0.62
ACT 74H	0.95	0.64	0.76	0.51
ACT 74C	0.88	0.5	0.71	0.36
SAT April 2017	0.87	0.64	0.72	0.38
SAT May 2017	0.86	0.66	0.71	0.42

Summary of Comparisons of the Three Assessments

A summary of alignment results by test form (excluding the writing prompt) is provided in Table 1b-22. The two Florida test forms were found to be acceptably aligned with the LAFS. The

two ACT test forms were found to need major adjustments in order to be aligned with the LAFS, One SAT test form was found to be acceptably aligned, while the other was found to need slight adjustment. If considering an average across both test forms, the average Florida test form would need five items revised or replaced for full alignment while the average SAT test form would need six items revised or replaced for full alignment. The SAT, therefore, is considered conditionally aligned with the LAFS, depending on the test form. The writing prompts for all test forms were considered to target appropriate corresponding writing standards within RC5 (Text-Based Writing) at an appropriate level of complexity.

Table 1b-22. Percent of LAFS Reporting Categories 1-4 for ELA Grade 10 with Acceptable Level on Each Alignment Criteria when Compared to Six Assessments

Assessment Form	Categorical Concurrence (Percent of RCs with over six items)	Depth-of-Knowledge Consistency (50% at/above)	Range-of-Knowledge (50% of standards)	Balance of Representation (without possible weakness)
FL Spr. 2016	100%	75%	100%	100%
FL Spr. 2017	100%	75%	100%	100%
ACT 74H	75%	50%	50%	75%
ACT 74C	75%	50%	75%	75%
SAT Apr 2017	100%	50%	75%	50%
SAT May 2017	100%	75%	75%	75%

The three assessments also varied in structure. The Florida grade 10 ELA test had the fewest number of items, followed by the SAT, and then by the ACT, which had the most items. The Florida grade 10 test allowed the most time for an assessment session, followed by the SAT, and then by the ACT, which allowed the least time for an assessment session. Similarly, for the essay component, the Florida grade 10 test allowed the most time for the essay task, followed by the SAT, and then by the ACT, which allowed the least time for the essay task. Both the ACT and SAT included multiple choice items only while the Florida grade 10 ELA test included 25-50% technology-enhanced items.

Conclusion

The central research question for the alignment analysis was to what degree the ACT or SAT can be used in lieu of the Florida Grade 10 ELA assessment designed to assess student proficiency on the corresponding LAFS to meet federal requirements. The two Florida Grade 10 ELA test forms analyzed were both found to be acceptably aligned with the corresponding LAFS, although neither test form was found to be fully aligned. Each test form would need five items revised or replaced for full alignment.

The SAT test forms were found to have conditional alignment, depending on the test form: One SAT test form was found to need five items revised or replaced for full alignment while the

other SAT test form was found to need slight adjustments—seven items revised or replaced—for full alignment. Thus, the SAT test and the LAFS were found to be conditionally aligned, depending on the test form considered. Although the overall alignment for the Florida ELA test forms and one SAT ELA test form was similar in terms of the number of items that would require revision or replacement for full alignment, the two assessments differed in their alignment issues. Both the Florida test forms and the SAT had unmet DOK Consistency for RC4 (Language and Editing) while the SAT test forms additionally had unmet Range of Knowledge for RC3 (Integration of Knowledge and Ideas) on both test forms as well as weaknesses in DOK Consistency and Balance of Representation for one or more reporting categories on each test form.

Study results show that the ACT would need major adjustments—defined as needing 10 or more items revised or replaced—to be fully aligned with the Florida Grade 10 LAFS. Although all three assessments had different numbers of items, 60-64 (FL), 115 (ACT), or 96 (SAT), the assessments all covered a similar percentage of the LAFS.

Because neither the ACT nor the SAT had enough items that corresponded to a sufficient number of standards for RC3 (Integration of Knowledge and Ideas), neither assessment fully addressed the breadth of the LAFS, as measured by the Range of Knowledge criterion. The assessments also differed in the overall distribution of content complexity of the items. Both the Florida Grade 10 ELA assessment and the ACT included DOK 1 items while the SAT did not. The Florida Grade 10 ELA assessment included the greatest percentage of DOK 3 items, followed by the SAT, and then by the ACT, which had the least percentage of DOK 3 items. The assessed LAFS, in contrast, included no DOK 1 standards and were primarily (72%) DOK 3.

The analysis indicated that 12 or 17 items would need to be added or revised on the ACT and that five or seven items would need to be added or revised on the SAT to attain full alignment according to the criteria used in this study. The Florida test forms would also need five items added or revised for full alignment. While augmenting the ACT or SAT to gain an acceptable level of alignment is certainly possible, it should be noted that augmentation adds costs, additional testing time, and complexity to the assessment administration process.

References

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

Valencia, S. W., & Wixson, K. K. (2000). Policy-oriented research on literary standards and assessment. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Vol. III.* Mahwah, NJ: Lawrence Erlbaum.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and mathematics education*. Council of Chief State School Officers and National Institute for Mathematics Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Section 2
Comparability Studies (Criterion 3)

Executive Summary

In determining whether to use the SAT or ACT in lieu of the Florida Standards Assessment (FSA), the comparability of the three exams must be examined. Comparability involves examining multiple aspects of the different assessments to determine how similar they are in content assessed, form structure, including psychometric properties of the form, administration requirements, and linking results.

This chapter summarizes the more qualitative analyses from other chapters and provides statistical analyses of the comparability based on students taking the FSA English Language Arts (ELA) and Algebra 1 End-of-Course (EOC) tests and either the SAT or ACT. Because many of the same students took two of the tests of interest, their results can be compared to see if it would be possible to link the two tests and create a table showing the equivalent score of one test based on another.

The primary method for linking the two tests was an equipercentile linking, which assumes that a known distribution of performance on one test can be applied to another. That is, if 50% of students pass one test, a cut score can be found on the second tests that would also result in 50% of students passing.

The Florida Department of Education (FDOE) had collected a rich data set of students who had taken either the ACT or SAT as well as the FSA ELA at Grade 10 and the Algebra 1 EOC tests. There were three years of data that included the ACT. Although there were four years of data for the SAT, because the test had changed dramatically in 2015, only two years of data were used for this study. Although this dataset had hundreds of thousands of students, it did not contain students who had taken the Algebra 1 EOC test in grade 8 or lower, which meant the mathematics analyses did not contain the most able students. Also, because the FSA ELA was given in grade 10 and the ACT/SAT was typically taken in grade 11, there was a year's difference in results during which students may have improved their ELA knowledge and skills.

Several analyses were conducted to account for the gaps in the data. Within each analysis, cut scores were calculated, and comparisons of the results were made. Because of the time differences, a prediction equation was also developed, using a linear regression to predict performance on the FSA based only on the score of either the ACT or SAT.

Ultimately, the results showed that the linking is acceptable for ELA but not for mathematics. Given the difference in content and in the time the test is administered, the results for the mathematics linking are not reliable. In an analysis of the validity of the results, the performance level the student actually scored in (based on their FSA result) was compared to the performance level the student was calculated to have performed in given their ACT or SAT score. The consistency of classification in the same performance level was quite low. It varied by subject and test, but in no condition were more than 52% of students classified in the same level.

Because of the differences in content and the time of the administration, which resulted in low classification consistency, the data do not support treating the FSA and ACT/SAT scores as interchangeable. The results show that neither the ACT nor SAT produce results that are fully comparable to the FSA and should not be considered alternatives to the ELA Grade 10 or Algebra 1 EOC tests.

Introduction

In this section, the comparability of the three assessments being studied is examined. The following topics are addressed:

- Content comparability
- Comparability of items, forms, and test statistics
- Comparability of test administration
- Statistical comparability of the three assessments
- Data files and samples of students in Florida
- Linking methodology and results
- Issues with state data and approaches to linking
- Changing the samples for improved linking
- An alternative approach to linking
- Classification consistency
- Conclusions

In determining whether to use the SAT or ACT in lieu of the Florida Standards Assessment (FSA), the comparability of the three exams must be examined. Comparability involves examining multiple aspects of the three assessments to determine if they are similar enough in content assessed, form structure, including psychometric properties of the form, administration requirements, and linking results.

This chapter summarizes the more qualitative analyses from other chapters and provides statistical analyses of the comparability based on students taking the FSA English Language Arts (ELA) and Algebra 1 End-of-Course (EOC) tests and either the SAT or ACT. Because many of the same students took two of the tests of interest, their results can be compared to see if it would be possible to link the two tests and create a table showing the equivalent score of one test based on another.

Students take the FSAs, ACT, and SAT at different grades. Most students take the ACT and SAT in their junior year of high school while Florida students take the Florida ELA assessment during grade 10 and the Florida Algebra 1 EOC whenever they take the course (although course enrollment is not required). In 2017, students took the Algebra 1 EOC in grades ranging from 5 to 12. These and other differences resulted in the need to perform several different statistical analyses in order to determine whether the tests can be used interchangeably in Florida's education system.

A strong linking study has three requirements of the data:

1. Either the same students take two tests or some of the same items appear on the two tests. In this case, the same students took the FSA and either the ACT or SAT.
2. The students who took both tests must be representative of the full population the results will cover. In this case, the students who took both tests did not have results that looked similar to what would be expected of all students in Florida. However, the data could be sampled to select a subgroup of students that better matched the full population of all students in the state.
3. The two tests must be given close in time and under the same administration requirements. Although the administration requirements are similar (but not exact), the bigger issue is that there are often large gaps in time between when the student took the FSA and when the student took the ACT or SAT. Furthermore, the gaps in time were often correlated with the ability level of the student. Data could be selected to reduce the gap in time, but that would increase the difference of this sample compared to the full population of students in Florida.

Several analyses were done to determine the best approach to linking the ACT and SAT to the FSA. The methodology and steps in the analysis process were as follows:

1 – Conduct Linking Studies. Equipercentile linking assumes that the same percentage of students should score at each performance level on one test as the other. The issue in using this method is the difference in time that students take the FSAs and either the ACT and/or the SAT.

A. Examine narrower groups of students to better link the results of the tests
    a. Entire student population
    b. Those taking the FSAs and ACT or SAT within 360 days of each other
    c. Those taking the FSAs and ACT or SAT within 120 days of each other
B. Perform statistical analyses on each group
    a. Test intercorrelations
    b. Equivalent cut points
    c. Effect sizes

Each time-period of students examined had results that were skewed away from what would be expected of the full population of students in Florida. Thus, a second set of analyses was conducted.

2 – Check Representativeness of Sample. These studies started with a new student sample by forcing a normal distribution onto the total sample and performing the same analyses. The issue with this approach is that there could be up to three years' difference in the time taking the two tests, during which learning will have occurred.

A. Calculate new mean scores vs. expected target mean scores
B. Revise test intercorrelations for new sample
C. Revise cut points for new sample

3 – Develop a Prediction Equation. Neither of these approaches fully met the requirements of a strong linking study. Another method for linking two tests does not require the times the tests were taken to be the same. A prediction equation would not assume the results are interchangeable but would use the results of one test to predict another. In this case, because all results need to be reported on the FSA scale, the ACT or SAT would be used to predict performance on the FSA.

A. Develop a prediction equation using the linear regression model
B. Calculate new cut scores
C. Compare the results to the results from the equipercentile linking

4 – Determine the Accuracy of Linking. Finally, further analyses were run to check the accuracy and validity of the various methods.

A. Classification consistency: Comparison of how students are classified based on data from the two different tests, and whether they are classified consistently
B. Exact match consistency: Only report the degree of exact match and the direction of the differing classifications

The results of all the analyses and conclusions from the comparability analyses are presented on the following pages.

Content Comparability

Criteria 1 and 2 examined the content comparability, also known as the alignment between the ACT and FSA and SAT and FSA. That is, the earlier chapters examined whether or not the three assessments measure the same subject material. In ELA, both ACT and SAT have large areas of overlap with the grade 10 content of Florida standards, but there are some deficiencies with both ACT and SAT that require additional or other items being used to complete the alignment. The ACT includes more items with a lower depth of knowledge and the categorical concurrence and range of knowledge of the reporting category *Integration of Knowledge and Ideas* is out of alignment with the Florida ELA standards. The SAT has a more limited range of knowledge for the same reporting category than the Florida ELA exam, but similar categorical concurrence.

It was more difficult to ascertain the content comparability for the mathematics assessments. The FL EOC focused only on Algebra 1, which includes algebra and modeling, functions, and statistics. The SAT includes items on linear equations, systems, problem solving, data analysis, complex equations, geometry and some trigonometry. The ACT assesses pre-algebra, elementary algebra, intermediate algebra, plane geometry, and coordinate geometry. Therefore, students are assessed at a greater depth in a more narrow content area on the Florida test and a larger content area at less depth on the ACT and SAT. When the alignment study focused solely on the items coded by the ACT and College Board as representing Algebra 1 content, the SAT was fully aligned in the algebra and modeling category, but weaker in the range of knowledge of the items in the functions and modeling and statistics and number system categories. The ACT performed less well, showing the test was not fully aligned with any of the three Florida reporting categories. The range of knowledge was weak or out of alignment for all three

reporting categories, and the categorical concurrence was not aligned for the third reporting category of statistics and number systems.

In the alignment study (Section 1 of the report), a recommendation was made in both content areas for both college entrance tests to add or modify a number of items to improve the alignment of the ACT and SAT with the Florida ELA and mathematics standards.

Comparability of Forms
When examining the comparability of the forms (Criterion 3), it is important to examine the number and types of items, how difficult they are, and the reliability of the forms themselves. Reliability refers to the degree of confidence that a student taking the same form a second time would receive the same score.

Number and Types of Items and Test Statistics – Table 2-1 shows the form construction of the three assessments in ELA and mathematics based on information from their technical reports. Although the FSAs have a slightly higher reliability coefficient than the ACT or SAT, all tests are sufficiently reliable. In ELA, both ACT and SAT have more items as they divide ELA into reading comprehension and language. The numbers of items in mathematics are very similar.

Table 2-1. Form Characteristics

	ELA				**Mathematics**		
Criterion	FSA: ELA 10	ACT	SAT		FL ALG 1 EOC	ACT	SAT
Form reliability	0.91	0.89	0.89		0.92	0.91	0.90
Form length	53 items + writing prompt	115 items + writing prompt	96 items + writing prompt		58 items	60 items	58 items
Distribution of item types	58% MC[1]; 23% Editing text choice; remaining is multi-select, hot text, and evidence-based Selected Response	MC + essay	MC + essay		Vast majority of item types are MC and SCR[2] (SCR = grid-in or equation editor). Other =table and matching	MC	MC + grid-in

[1] MC=multiple-choice item, with four options and one correct answer
[2] SCR=short constructed-response item

Item difficulty[3]							
--Mean	0.65	0.58	0.58		0.21	0.58	0.58
--Min	0.12	0.20	0.03		0.00	0.20	0.03
--Max	0.92	0.89	0.98		0.75	0.89	0.98

Florida has more variability in the types of items used. In ELA, both ACT and SAT rely on passage-based multiple-choice (MC) items. Although approximately 58% of the items on the Florida ELA grade 10 assessment are MC, approximately 23% of the items involve editing text choice, a type of technology-enhanced item. The remaining items are also technology enhanced and include multi-select, which involves more than four response options with more than one correct option; hot text items, which require students to highlight the correct response in a text; and evidence-based selected-response items, which contain multiple parts, typically an answer-explain type of question. All three ELA tests have an essay component.

For mathematics, the ACT has only MC items; the SAT includes MC and grid-in items; and the Florida Algebra 1 EOC exam contains primarily MC and short constructed response (SCR) that could involve either gridding in the correct answer or using an equation editor to write out the correct equation. The FL exam may also include items that require a student to complete a table or match equivalent equations.

Item Difficulty – Another area of comparability involves the difficulty of the items, and how the test is constructed to produce the most precision. The level of difficulty of the items is similar in ELA, but in mathematics, the Florida Algebra 1 EOC appears to have much more difficult items as a whole. Florida has a greater proportion of items that are not multiple-choice, meaning the student does not have a 25% chance of answering correctly by guessing. In this case, the differences in item types appear to have a much larger effect on mathematics.

Test Precision – Tests are constructed to provide more reliable results in certain areas. That is, if a test is designed to determine whether or not a student is college ready, more items are written around the difficulty level that truly distinguishes those who have the knowledge and skills necessary to succeed in college from those who do not. Test developers can examine the level of precision at each score point on a score scale using something called an information curve. That curve plots the reliability of each score point. The FL exams are reported using four cut scores to divide scores into five performance categories; therefore, the tests are designed to have the greatest precision across those cut scores. Both ACT and SAT have only one college-ready benchmark, and thus the precision of the assessment is focused on that cut score.

Nonetheless, the test information curves are similar for five of the assessments, with the most information in the middle of the scale and less precision at the low and high ends. However, the ACT math assessment has a bi-modal distribution, meaning that the greatest precision is actually in two areas of the scale: around scale scores 15-17 and 25-27, with higher errors between 0-12 and 30-32, which is expected, but also around scores 20–22, which is surprising

[3] Item difficulty is shown as the percentage of students answering an item correctly. The minimum and maximum show the percentage of students answering the hardest and easiest item on a form correctly. The mean gives an indication of the overall difficulty of the form, by summarizing the percentage of items answered correctly.

given that is the area of the college-ready benchmark (and where some colleges make admittance decisions).

Scale Scores – Finally, it is important to examine how the student results are converted into scale scores for reporting. The FSA exams use item response theory, which includes an analysis of item difficulty, discrimination, and guessing factor, and places both the items and students on the same scale. Both the ACT and SAT use a more classical approach that involves a normative component to ensure the students are spread across the scale in a typical bell curve format. It should be noted that both ACT and SAT were designed several decades ago when classical test theory was the primary method used for tests. Although the purpose of this study was not to evaluate different scaling methodologies, it is important to note that different methodologies can produce different results.

Comparability of Test Administration

The FSAs are not intended to be timed tests. Students have a half of a school day in which to complete them, but those needing additional time may have up to an additional half of a school day to do so. The ACT and SAT are both timed tests, with specific times set for each section so that every student must begin and end each section at the same time. Less able students often do not finish all items in those time frames. This difference in time allotted per item raises questions about the comparability of the assessments. When tests are designed to force a student to work quickly, they are testing the student's fluency with the subject matter, and not simply their knowledge and skills.

Criterion 4 addresses the comparability of accommodations for the three different assessments. Although there is an approval process for the ACT and SAT that differs from the FSA, which only requires accommodations to be listed in students' IEPs or language development plans, the majority of accommodations for students with disabilities allowed on the FSA are also allowed on the ACT and SAT. As part of this study, the panels evaluating the accommodations for ACT and SAT expressed concern about only two groups: English learners and students who are deaf/hard-of-hearing. If these two groups receive different accommodations on the FSA than on either the ACT or SAT, that will affect the comparability of the scores. One score would demonstrate what a student knows and can do when given a specific accommodation (on the FSA) while the other would show what that student knows and can do without that same accommodation (on the ACT or SAT).

Statistical Comparability

In this phase of the analysis, the statistical comparability among the ACT, SAT, and FSA tests were measured. The goal of the analyses was to determine if the SAT and ACT could be linked to the FSA scales to provide comparable, valid, and reliable scores. If the scores could be used interchangeably, then the SAT and ACT can possibly be used in lieu of the current Florida state assessments for Grade 10 ELA and Algebra 1 EOC.

Samples of Students Studied

Over the past two years, thousands of students have taken the FSAs and either the ACT or SAT. In some cases, students took the ACT or SAT more than once, and, in that case, the test administration closest in time to the FSA of interest was used. This trimming of the file resulted

in a set of 157,287 paired scores for the ACT and the FL Grade 10 ELA test, 32,617 paired scores for the ACT and FL Algebra 1 test, 210,676 paired scores for the SAT and Grade 10 ELA test, and 51,036 for the SAT and FL Algebra 1 test. These sample sizes are more than sufficient to conduct linking studies. However, it is worth noting the smaller sample sizes in mathematics, were likely due to larger differences between test administration times for some students. In addition, because the SAT changed dramatically in 2016, there are only two years of SAT data. Thus, a student who took the Algebra 1 EOC test in eighth grade and the SAT in eleventh grade would not be included in this sample. There are three years of ACT data, but if a student took the Algebra 1 test in seventh grade and the ACT in eleventh grade, those data would not be included.

Two concerns with the dataset were analyzed and addressed in the following sections. First, Florida data show that only about half of the students who are enrolled as tenth or eleventh graders actually take either the ACT or the SAT before they graduate from high school. This means that matched samples of students who participate in the FSA assessments who took one or both of the college entrance tests seriously underrepresent the full sample of FSA test-takers (who presumably are the lower-scoring students that take only the FSAs).

Second, the data provided by the FDOE indicates that 31% of first time Algebra 1 test takers were in eighth grade and 52% were in ninth grade. The ACT or SAT is typically taken in spring of tenth grade, fall of eleventh grade, and/or spring of eleventh grade. This implies that appropriate matching of the data will need to focus on the grade level and semester of students when taking ACT or SAT.

Large distances between time of testing in the two tests will increase measurement error as learning continues to occur between those test administration times. In short, the data provided for this study are neither representative of the full population of students nor were the tests taken close enough in time to assume no learning occurred. If we try to correct for one of the problems, we increase the amount the second problem would affect the scores.

Linking Methodology

The Buros Center for Assessment had conducted a linking study in April 2017 to place ACT and SAT scores on the Florida scale. This study resulted in a concordance table that shows for every possible score on the ACT or SAT what the corresponding score on the Florida scale would be. ACT and the College Board have used similar tables to show the equivalent ACT score for each SAT score and vice versa.

The same procedures used for that study were applied to this more recent data. A full equating requires equal constructs, similar administrations, and either the same items or the same students. In this case, the constructs are similar but not equal, different administration conditions exist, and the same students are tested across different time periods. Thus, the scores can only be linked, not equated. The single-group equipercentile linking method (Kolen & Brennan, 2014) was used to build the concordance tables between the FSA assessments and alternative test scores.

The objective of equipercentile linking is to find alternative test scores on the SAT or ACT that correspond to the FSA Grade 10 ELA and Florida Algebra 1 EOC scores at the same percentile rank using the observed score distributions. Equipercentile linking is a successful method that has been used extensively for comparing different test scores obtained from a single sample (Sawyer, 2007). This method has been used for various applications including the linking of SAT scores to ACT scores (Dorans, 1999). Prior to applying the linking procedure, the data were smoothed using a loglinear approach. For more details on the smoothing and linking, please see **Appendix 2.A**.

One issue raised in the previous set of analyses is related to the differences in time between when students took the FSA and either the ACT or SAT. This problem is particularly prominent for Algebra 1, as that subject is not connected to any specific grade level. Students take that test anywhere from Grade 5 through Grade 12 with a modal grade of Grade 9. Conversely, the ELA assessment is given to all students at Grade 10. Students most commonly take the ACT or SAT in the spring of eleventh grade. Thus, an "average" student could take the Algebra 1 exam in grade 9, the ELA exam in grade 10, and either the ACT or SAT in grade 11. Therefore, concording a score from the ACT/SAT to replace the FSA could be based on scores that are, on average, one year apart for ELA and two years apart for Algebra 1. For high performing students, this time differential could be even higher in mathematics.

As mentioned earlier, a choice needed to be made regarding which ACT or SAT score to use when a student has taken the test more than once. At this point, students taking the ACT or SAT at all are likely to be higher achieving students who are pursuing college after high school graduation. Yet, the policy of using ACT or SAT will be applied across the district, so the linking must be accurate for students scoring across the achievement scale, including those not planning to attend college. Using the highest ACT/SAT score would exacerbate the problem. Using the score closest in time to the FSA administration would minimize the problem of the time differential and be more representative. For the vast majority of the cases in mathematics, the administration closest in time to the FSA is also the first administration. However, for ELA, there were approximately 13,000 students whose first SAT administration was not the one closest to the ELA administration or about 7% of the scores. The number was 2,000 for ACT, which represented about 1% of the sample.

One solution is to analyze the concordance of a narrower population of students. For instance, scores could be linked only for students who took the two exams within a semester or a year to examine which group was most similar to the full population of Florida students. Thus, the results show several conditions for time constraints using the ACT/SAT scores from the closest point in time to the FSA score.

Initial Linking Results

The first analysis run was to evaluate whether there were sufficient numbers of students for each time period selected. Table 2-2 shows that every cell has at least 1,000 students, which is sufficient for the linking study. Interestingly, the largest number of students taking tests in the same semester was for students taking the ACT and Florida Algebra 1 EOC. That means this sample contains students who took the EOC exam in either tenth or eleventh grade, which

would tend to be lower performing students. However, the fact that they took a college-entrance test points to them being higher performing.

Table 2-2. Sample Sizes for Various Points in Time

	ELA		Math	
Time lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	157,287	210,676	32,617	51,036
Students taking test within 360 days (one year) of the FSA	109,099	172,794	6,907	8,140
Students taking test within 120 days (one semester) of the FSA	15,254	2,490	1,363	1,122

Test Intercorrelations

The next step was to examine the correlations between the Florida exams and the ACT and SAT to see if they improved when looking at narrower points in time. Table 2-3 shows that using a narrower time differential has very little effect in ELA. Surprisingly, within one semester, the correlation between the FSA and the SAT decreases slightly. For mathematics, the correlation increases significantly for mathematics when narrowing the time differential to one year for SAT and to one semester for ACT. Overall, a correlation of 0.80 is fairly strong for two ELA assessments; however, even the best correlation of 0.67 between the Florida Algebra 1 exam and either the SAT or ACT is not strong enough to endorse using the results interchangeably.

It is important to note that in 1997, when the ACT and SAT scores were first concorded, the correlation between the two composite scores was 0.92. The Verbal section of the SAT correlated with the reading and English sections of the ACT at 0.83. The Math portions of the two tests correlated at 0.89. All of the correlations in Table 2-3 were lower than the initial calculated correlations. Studies are still ongoing between the current versions of the ACT and SAT, so the current correlations are unknown.

College Board has released concordance tables between the ACT and the new SAT, but they were calculated by first linking the old SAT to the ACT and then linking the new SAT to the old SAT. This method introduced considerable error, and ACT has refused to accept the results. They are conducting a new study that is expected to be released in Summer 2018.

Table 2-3. Correlations between ACT or SAT and the FSA for ELA and Mathematics for Various Points in Time

	ELA		Math	
Time lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	0.80	0.80	0.58	0.58
Students taking test within 360 days (one year) of the FSA	0.80	0.80	0.61	0.67
Students taking test within 120 days (one semester) of the FSA	0.81	0.75	0.67	0.67

Equivalent Cut Points

Next, the equivalent cut scores after linking were examined to see if there were differences based on the time differential between test administrations. The results for ELA do not differ much across the time periods, as shown in Table 2-4. For mathematics, however, the results provide additional evidence that students taking the two tests within a semester are lower performing students.

Table 2-4. Equivalent Cut Points for ACT and SAT for Cut Score 3 (Proficiency) of the FSA

	ELA Gr 10 (FL cut = 350)		**Algebra 1 (FL cut = 497)**	
Time lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	18	490	17	450
Students taking test within 360 days (one year) of the FSA	18	490	16	450
Students taking test within 120 days (one semester) of the FSA	18	480	16	430

Interestingly, Table 2-5 shows that while the population who took two tests is skewed, it is skewed in opposite directions for ELA and mathematics. Students in the sample did better on the ELA test than the full population of Florida students; however, students in the sample did significantly worse on the Algebra 1 EOC test than the full population. Again, the latter result is most likely to do with the fact that the higher performing students have longer gaps in time between their Algebra 1 EOC scores and their ACT/SAT scores and thus are not included in this dataset.

Table 2-5. Comparing Percentage of Students Scoring at each Florida Performance Level, by Subject and Population

	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
FL Grade 10 ELA State results	25	25	21	19	9
Students who also took ACT	20	26	20	23	12
Students who also took SAT	18	24	23	24	12
FL Algebra 1 EOC State results	27	11	30	17	15
Students who also took ACT	40	17	30	9	5
Students who also took SAT	41	17	29	9	4

Effect Sizes

With the differences shown in Table 2-5, an effect size was computed by comparing the mean of the current sample in all the variant forms with that of the total population of Florida state test takers. An effect size is a calculation that quantifies the advantage that one group may have over the other. A larger effect size shows that one group has a larger advantage over the other. For instance, when a student's ACT score was used to determine the comparable FSA ELA

score, the scores were higher, on average, than the actual score. However, it is not sufficient to simply say that the averages differed by 4 points. The standard deviation must be considered to determine how much of that difference could be due to random effects.

Even small effect sizes can have large implications when comparing two samples. For example, an effect size of 0.2 means that the 50th percentile of one group is equivalent to the 56th percentile of the second group.[4] As shown in Table 2-6, effect sizes were all greater than 0.16 and deemed to be discrepant enough to warrant a subsample comparison. It is worth noting, however, that the effect sizes for ELA are considered "small" statistically, and the effect sizes for mathematics are considered "medium" statistically.

Table 2.6 Effect sizes for the different populations

ELA	**Mean**	**SD**	**Effect Size**
Florida total	348	23	0.167
ACT Linked	352	22	
Florida total	348	23	0.216
SAT Linked	353	22	
MATH	Mean	**SD**	**Effect Size**
Florida total	498	30	-0.294
ACT Linked	490	28	
Florida total	498	30	-0.349
SAT Linked	488	28	

Changing the Sample

Given the differences in the distribution and the different mean scores of the various populations, a new sample of students was drawn that was a subsample of all students who took both the Florida exams and either the SAT and ACT. The sample was drawn around the desired mean of each of the four datasets: ACT ELA, ACT math, SAT ELA, and SAT math.

A subsample was drawn from each of the four categories by forcing a normal distribution onto the total sample, i.e., the desired mean (the actual Florida mean) was established as the center point of sampling, and 5 intervals were sampled in both directions around the center point, representing ±.5, ±1, ±1.5, ±2, and ±3 standard deviations, respectively, creating 10 total range-specific sampling points. The standard deviation used was that of the total sample, but the true Florida standard deviation was nearly identical. A sample size of 10,000 was drawn for each of the ELA conditions, and a sample size of 7,000 was drawn for each of the mathematics conditions.

The new means were very good approximations of the target means:

- ACT ELA: Target mean: 348. Obtained sample mean: 347.98
- SAT ELA: Target mean: 348. Obtained sample mean: 347.97

[4] See Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

- ACT MATH: Target mean: 498. Obtained sample mean: 497.63
- SAT MATH: Target mean: 498. Obtained sample mean: 497.62

Then, the analyses presented in Tables 2-2 through 2-4 were rerun using this new sample. The results are shown in Tables 2-7 through 2-9.

Revising the Samples

Table 2-7 shows drastically reduced sample sizes, as expected, but the number of cases that took the two tests within 120 days has dropped below 1,000 in all four scenarios and became inadequate for use due to the increase in uncertainty. Again, the number of students who took the tests within a year from each other is much smaller in mathematics than in ELA.

Table 2-7. Revised Sample Sizes for Various Points in Time Using the Reduced Sample

	ELA		Math	
Time Lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	9,900	9,900	7,000	7,000
Students taking test within 360 days (one year) of the FSA	6,811	8,095	1,521	1,125
Students taking test within 120 days (one semester) of the FSA	949	104	341	208

Revised Intercorrelations

Table 2-8 shows similar correlations with the subsample for ELA but better correlations for mathematics. These results indicate that the scores are better aligned for the higher performing students, which were oversampled in the ELA group and under-sampled in the mathematics group. The correlations for students taking the two tests within 120 days in ELA also over-represent the higher performing population, and the correlation is strongest there. Given the small sample sizes for the group of students taking the test within 120 days in the sampled population; however, these numbers are less reliable than the other two time conditions.

Table 2-8. Revised Correlations Between ACT or SAT and the FSA for ELA and Mathematics for Various Points in Time Using the Reduced Sample

	ELA		Math	
Time Lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	0.80	0.79	0.69	0.71
Students taking test within 360 days (one year) of the FSA	0.81	0.80	0.73	0.75
Students taking test within 120 days (one semester) of the FSA	0.81	0.82	0.77	0.77

Revised Cut Points

Finally, Table 2-9 shows small differences in the cut scores for the reduced sample that better match the full sample of Florida students.

Table 2-9. Equivalent Cut Points for ACT and SAT for Cut Score 3 (Proficiency) of the FSA Using Linked Data of the Reduced Sample

	ELA Gr 10 (FL cut = 350)		Algebra 1 (FL cut = 497)	
Time Lag	**ACT**	**SAT**	**ACT**	**SAT**
All students	19	500	17	450
Students taking test within 360 days (one year) of the FSA	19	500	17	440
Students taking test within 120 days (one semester) of the FSA	19	480	15	420

The full concordance tables for each condition represented in Table 2-9 can be found in **Appendix 2**; the results are in **Appendix 2.B** and **Appendix 2.C** for ACT and SAT, respectively.

Given the data from these analyses, constraining the sample by time appears to skew the results, albeit in different directions for the two subjects. For ELA, constraining the results by less than one year would result in only including students who took the ACT or SAT in the spring of their sophomore year, when they took the FSA grade 10 test. This subsample represents high performers with family incomes sufficient to pay to take the ACT or SAT multiple times. The most frequent pattern of test taking is for students to take the FSA ELA grade 10 test in Grade 10 and the ACT or SAT in Grade 11. Restricting the sample to students taking the two tests within 360 days is reasonable once the data are sampled to include a distribution of scores that best represent the state of Florida. These concordance tables are shown in **Appendix 2** (**2.B** and **2.C**, Tables 2.4 and 2.5 for ACT and SAT, respectively).

In mathematics, constraining the sample by time would result in only including students who took Algebra I in eleventh grade, which is later than the typical student in Florida and is indicative of lower mathematics performance. Even constraining the sample to those taking the exam within 360 days would only include students who took Algebra 1 in tenth grade. The grade in which most students take Algebra 1 in Florida is Grade 9, and students who are advanced in mathematics tend to take the course in middle school. Therefore, the preferred approach is to use the sample that best represents all students in Florida and not constrain the results by time, so as to better include the higher performing students who take the Algebra 1 test in earlier grades. **Appendix 2** (**2.B** and **2.C** Tables 2.6 and 2.7) will be used to determine the effect on state accountability reporting.

Discussion of Application of Linked Scores

As described earlier in this chapter, the purpose of this study is to determine the fairness of using three different assessments in Florida's accountability system. If the policy is fully enacted, districts choosing to use the Florida test will be evaluated by their students' performance in ELA at grade 10 and in mathematics at whatever grade level the student takes Algebra 1. Districts choosing to use the ACT or SAT will be evaluated by student performance in both subjects at grade 11. There is an inherent unfairness in that approach, as districts using the ACT or SAT will have an additional year (or more) to teach the students, and it is expected

that students will have learned more in the content area being assessed. Thus, adjustments in the analytical approach needed to be made to address that inherent inequity.

Given the differences in content covered by the Florida Algebra 1 EOC exam and the ACT and SAT mathematics exams, caution should be used in interpreting the equated scores between those tests. The correlations between the ELA tests appear to substantiate the alignment findings that the tests are similar in content with only some necessary augmentation. In addition, the findings from the alignment study show that the three tests are fairly comparable and measure the state ELA standards moderately well. However, the mathematics portions of the ACT and SAT differ from each other and from the Florida State Sunshine Standards to a larger degree, as found in the alignment study, and also confirmed by a lower statistical correlation in results of students taking both tests. These two tests also have larger differences in item types and item difficulties, and the alignment of them is less certain.

An alternative approach to linking

As a result of the issues found in attempting to fully link the tests, an alternative approach to linking the three tests was tried. These analyses are based on an equipercentile method of linking. That method assumes that the same percentage of students should score at each performance level on one test as the other. In this case, that means that the same percentage of students scoring at level 3 on the FSA ELA test in Grade 10 should score at level 3 on either the ACT or SAT in Grade 11. Likewise, regardless when students took the Algebra 1 EOC exam, the assumption is made that the score distribution will be the same as on the ACT or SAT mathematics in eleventh grade. Based on the data that were analyzed, these assumptions are questionable, at best.

Linking requires that the following conditions be true:

- Either the students or the items are the same across both tests being linked;
- The conditions under which the tests are given are the same;
- The populations being linked are equivalent; and
- No learning occurs between the administration of the two tests.

In these analyses, only two of the four conditions are true. The same students took the two assessments, but the time differences between the two were extreme enough that learning surely occurred. If the sample is restricted such that only students who took the two tests within the same semester are included, the sample is no longer representative. In ELA, the sample becomes students who took the FSA and the ACT/SAT exams in Grade 10, which tends to be the higher performing students. In mathematics, the sample becomes students who took the Algebra 1 and ACT/SAT exams in Grade 11, which tends to be the lower performing students. In neither case are the conditions under which the students took the tests the same.

Given that all the requirements for linking cannot be met, a less restrictive approach would be to develop a prediction equation. That is, a regression equation could predict the score of one test given the other. The issue with a prediction equation is that it assumes linearity. That is, the same formula will be applied to scores at the lower end of the scale as on the upper end, which

may not be supported. It is worth providing those results, however. The formula that would be applied to the four conversions as follows:

FL Algebra 1 Scale Score = 394.149 + 0.217 * SAT Math Score
FL ELA Scale Score = 259.54 + 0.182 * SAT ELA Score
FL Algebra 1 Scale Score = 402.66 + 5.338 * ACT Math Score
FL ELA Scale Score = 292.45 + 3.034 * ACT ELA Score

To see how static the formulas would be with a subpopulation, the regression was re-run for mathematics, using only the population who took the test within 120 days, which should contain students at the lower part of the scale. Those two prediction equations are as follows:

FL Algebra 1 Scale Score = 375.261 + 0.273 * SAT Math Score
FL Algebra 1 Scale Score = 392.131 + 6.244 * ACT Math Score

Using these prediction equations, an SAT math score of 350 would be the equivalent of a 470 using either equation. However, if the SAT math score was a 650, the equation built from the full population would predict a Florida Algebra 1 score of 535, while the equation built from the lower performing population predicts the Algebra 1 score to be 552. This implies that using the time constrained sample severely over-predicts performance at the high end of the scale. It could also mean that the regression equation might need to add more variables to predict better at different points on the scale.

For ACT, the FSA scores associated with a score of 17 are 493 and 498 for the full population and the time-limited population, respectively. At the high end, an ACT score of 30 converts to a 562 and a 579, respectively. Again, these results show that the population chosen has a strong effect on the results. Using the lower-performing sample over-predicts performance on the high end. It is not possible to build a prediction equation for the higher performing students, because there are no data from students who took Algebra 1 before ninth grade, but these results seem to indicate that a non-linear prediction equation is needed. These results are also quite different from the results obtained through equipercentile linking, as they spread the concordance linearly, while the equipercentile approach matches the distribution of performance on the Florida score scale.

The sample is more representative for ELA as the time between administrations is primarily limited to one year. The prediction equations are much more similar to each other for the different populations within ELA. For the SAT, a score of 350 corresponds to an FSA score of 323 for the full population and a 331 for the time-limited sample. At the upper end, an SAT score of 650 corresponds to scores of 377 and 381 for the full population and time-limited sample, respectively. For the ACT, a score of 17 corresponds to an FSA score of 344 using the full population and 346 using the time-limited population. At the upper end, a score of 30 converts to 383 and 386, for the full population and time-limited samples, respectively. The ACT conversions are close to each other and close to the results obtained using the equipercentile method, perhaps due to the more truncated ACT scale. The SAT conversions are similar at the upper end of the scale, but have wide divergence at the lower end.

Implications

To truly determine the effect of using one test or another, it is important to compare actual scores to concorded scores. This will allow a comparison of how students are classified based on data from the two different tests, and whether they are classified consistently. That is, for students who scored a 350 on the ELA test in the full sample, what percentage of them would have been considered proficient using their actual ACT or SAT score and the linked cut scores of X and Y, respectively? Similarly, for students who scored a 349 on the ELA test and would not have been deemed proficient on the Florida test, what percentage had an actual ACT or SAT score above the linked cut score? Does the classification consistency differ for mathematics?

Classification Consistency

Tables 2-10 through 2-13 show the classification consistency when applying the linked cut scores to students' actual performance. To read this table, start with the first row (FSA Level 1) and read across.

The first row indicates that 65% of students who scored at Level 1 on the FSA would also score at Level 1 if they had only taken the ACT and the concordance table was used to give them a score on the FSA; 23% would have been categorized in Level 2, 3% in Level 3, and 1% in Level 4. For both ELA and math, only at performance levels 1 and 5 are more than half of the students consistently placed in the same performance level by both assessments.

Table 2-10. Classification Consistency Between FSA ELA Grade 10 and ACT for Sampled Population Taking the Two Tests Within One Year

	ACT linked Level 1	**ACT linked Level 2**	**ACT linked Level 3**	**ACT linked Level 4**	**ACT linked Level 5**
FSA Level 1	65%	23%	3%	1%	0%
FSA Level 2	27%	52%	16%	5%	0%
FSA Level 3	7%	33%	32%	26%	2%
FSA Level 4	1%	9%	21%	52%	17%
FSA Level 5	0%	1%	4%	35%	59%

Table 2-11. Classification Consistency Between FSA ELA Grade 10 and SAT for Sampled Population Taking the Two Tests Within One Year

	SAT linked Level 1	**SAT linked Level 2**	**SAT linked Level 3**	**SAT linked Level 4**	**SAT linked Level 5**
FSA Level 1	66%	29%	4%	1%	0%
FSA Level 2	27%	48%	20%	4%	0%
FSA Level 3	6%	31%	39%	22%	2%
FSA Level 4	1%	8%	26%	48%	17%
FSA Level 5	0%	1%	5%	33%	61%

Table 2-12. Classification Consistency Between FL Algebra 1 EOC and ACT for Sampled Population

	ACT linked Level 1	ACT linked Level 2	ACT linked Level 3	ACT linked Level 4	ACT linked Level 5
FSA Level 1	73%	18%	8%	0%	1%
FSA Level 2	55%	24%	19%	2%	1%
FSA Level 3	30%	24%	35%	6%	5%
FSA Level 4	6%	9%	38%	19%	28%
FSA Level 5	2%	2%	16%	14%	66%

Table 2-13. Classification Consistency Between FL Algebra 1 EOC and SAT for Sampled Population

	SAT linked Level 1	SAT linked Level 2	SAT linked Level 3	SAT linked Level 4	SAT linked Level 5
FSA Level 1	50%	19%	24%	5%	2%
FSA Level 2	29%	20%	38%	10%	4%
FSA Level 3	16%	14%	40%	18%	11%
FSA Level 4	4%	6%	30%	27%	33%
FSA Level 5	2%	3%	15%	22%	59%

Exact Match Consistency

Another way to examine the data is to collapse the categories and only report the degree of exact match and the direction of the differing classifications. Table 2-14 shows the percentage of students placed in the same category, a lower category, or a higher category by the alternative test compared to the Florida test.

The typically used standard is 90% or greater on-diagonal agreement between the tests being compared. None of the comparisons came close to this. The ACT and SAT put students in both lower and higher levels. Unfortunately, the results show that the two tests could categorize a few students in as much as *four levels* off from their actual level.

The ACT is slightly more likely to under-predict performance compared to the FSA, while the SAT is more likely to over-predict performance. Regardless, the consistency between either the ACT or SAT and the FSA scores is very low. Again, the results call into doubt the comparability of the assessments and the feasibility of using the scores interchangeably for purposes of accountability.

Table 2-14 collapses the results to show simply whether the ACT and SAT placed students in the same, higher, or lower category. Again, the results call into doubt the comparability of the assessments and the feasibility of using the scores interchangeably for purposes of accountability.

Table 2-14. Decision Consistency Across the Four Linked Assessments

	ELA			Math		
ACT	**ACT placed lower**	**ACT matched**	**ACT placed higher**	**ACT placed lower**	**ACT matched**	**ACT placed higher**
FSA Level 1	0%	65%	35%	0%	73%	27%
FSA Level 2	27%	52%	21%	55%	24%	21%
FSA Level 3	40%	32%	28%	54%	35%	11%
FSA Level 4	31%	52%	17%	53%	19%	28%
FSA Level 5	41%	59%	0%	34%	66%	0%
	ELA			**Math**		
SAT	**SAT placed lower**	**SAT matched**	**SAT placed higher**	**SAT placed lower**	**SAT matched**	**SAT placed higher**
FSA Level 1	0%	66%	34%	0%	50%	50%
FSA Level 2	27%	48%	25%	29%	20%	51%
FSA Level 3	37%	39%	24%	30%	40%	30%
FSA Level 4	35%	48%	17%	40%	27%	33%
FSA Level 5	39%	61%	0%	41%	59%	0%

Classification Consistency Overall Summary

Because the classification consistency is quite low, especially for mathematics, students could meet the high school graduation requirement who are not truly prepared, and there will be others who are truly prepared but are placed in a lower level.

Aggregating only the scores that were placed in the same level across the two administrations shows that only about half the time will the ACT or SAT place the students in the same performance level as the FSA. Table 2-15 focuses only on the percentage of students classified at the same level by both assessments, regardless of performance level. The results show that the linking works better for ELA than for mathematics and slightly better for the ACT than the SAT.

Table 2-15. Percentage of Students Placed in the Same Performance Level by the Alternate Test as by the FSA

Test	ACT	SAT
ELA	51.3%	50.8%
Math	48.4%	40.2%

Conclusion

If FDOE chooses to use the equipercentile linking method, the recommendation is to use the results from the sampled population that best represents the full student population in the state. Constricting the sample by time would only make sense if the students would be similarly restricted in the time frame in which they took the three assessments. That is, if Florida changed the policy such that the state assessment was given in grade 11, and a new mathematics assessment, not linked to a specific course, was also given in grade 11, then an equipercentile

linking method constrained by time differences would be the most appropriate method for linking all of the tests. Or if the policy would be that students needed to take the ACT or SAT after the semester they would have normally taken the FSA tests, then the time-restricted population would be appropriate.

However, state policy currently assesses students in grade 10 for ELA and whenever they complete Algebra 1, which included students in all grades between 5–12 in 2017. The intended policy is to now compare those scores to scores of students taking the ACT and SAT in grade 11; therefore, it would not be appropriate to restrict the linking sample beyond taking the ACT/SAT within one year of the FSA ELA Grade 10 test.

Because the population of students taking both tests was skewed towards higher performing students in ELA and lower performing students in mathematics, it was necessary to draw a more representative sample from the available students. The result was a more representative sample, particularly in mathematics. The correlations between the two tests increased for both the ACT and SAT across all time variables. And the differences in sample sizes showed the majority of students take the ELA exams within a year of each other, but the majority of students take the mathematics exams more than a year apart. Given the intended use of the linking, using the results from the sampled population taking the two ELA tests within a year is preferred.

Because of the multiple complexities associated with the mathematics tests, policymakers might question using the SAT or ACT in place of the Algebra 1 test and calling them equivalent. The alternative tests may, in fact, be better representations of high school learning in mathematics, but they are not highly comparable to Algebra 1, specifically. Because of the great variation in time between the two test administrations, it is recommended to use the linked tables that are based on a representative (and larger) population with no time restrictions.

However, the results shown in the previous section raise serious questions regarding the wisdom of making accountability decisions based on the use of the three different assessments. Districts using the FSA option may have very different results than districts using either the ACT or SAT options.

In Section 4 (Criterion 5), a series of accountability analyses are described in order to demonstrate the effect of using these concordance tables on student score reports and Florida school report cards. For the schools, both the performance measures and the gain scores will be calculated for similar schools under the three testing conditions: using the ACT, the SAT, or the FSA. Again, the decision consistency will be compared, but at the school level.

At this point, however, it appears the ACT and SAT do not produce results comparable to the FSA and should not be considered alternatives to the ELA grade 10 or Algebra 1 EOC assessments.

Section 3
Accommodations Studies (Criterion 4)

Executive Summary

This report presents the findings of Accommodations Studies that were conducted to evaluate the degree to which the ACT and SAT provide accommodations that permit students with disabilities and English learners (ELs) the opportunity to participate in the assessment and receive comparable benefits to the Florida grade 10 statewide, standardized ELA assessment and the Algebra 1 end-of-course (EOC) assessment. The studies were conducted by the National Center on Educational Outcomes (NCEO) as part of a larger effort by the Assessment Solutions Group (ASG) for the Florida Department of Education to determine whether Florida high schools can choose which high school assessment they wish to use. NCEO is an independent non-profit organization at the University of Minnesota that focuses on the inclusion of students with disabilities, ELs, and ELs with disabilities in comprehensive assessment systems.

The Accommodations Studies were designed to obtain input from Florida educators on accommodations for the ACT, SAT, and Florida Standards Assessments (FSA). Toward this end, NCEO organized a math educator panel and an ELA educator panel to review the assessments using materials and processes developed by NCEO. A 1½-day meeting of the two educator panels was held on October 26-27, 2017, in Orlando, Florida.

Panel members identified four considerations when evaluating whether the ACT and SAT offered comparable benefits: (1) uses (e.g., accountability, graduation, programmatic changes); (2) accommodations allowed; (3) process (e.g., requests, college reportability); and (4) cultural sensitivity. The findings of the panel members and NCEO for each consideration are in this section.

Uses (e.g., School Accountability, Graduation, Programmatic Changes)
NCEO and panel members found that ACT and SAT offer comparable benefits to the FSA for the identified purposes of school accountability and graduation. The results of the FSA are used for accountability and graduation purposes. Similarly, Florida may allow the ACT and SAT results to be used for accountability and graduation purposes.

Educator panel members indicated that FSA scores are sometimes used to make programmatic changes (e.g., instructional decision making). Timely receipt of results are important when assessment data are used for programmatic changes. Based on the information provided by ACT and SAT, the panel members wondered whether there would be a quicker turn around for the ACT and SAT compared to the FSA turn-around time. NCEO concluded that there was no information available to address the use for programmatic changes.

Accommodations Allowed
There is a slightly higher number of accommodations for students with disabilities for the ACT and SAT compared to the FSA. For ELs, there were differences in the number of accommodations across the three tests, with the ACT providing more, then FSA, followed by SAT providing a much smaller number. One reason for the smaller number of accommodations

for ELs for the SAT was because the College Board assumed that the student was EL only (i.e., EL without a disability), while ACT included accommodations for ELs that typically would be used only by ELs with a disability. NCEO and panel members thought that the differences in numbers of accommodations for students with disabilities and ELs may indicate a fairness issue, but this issue goes across all three assessments.

Process (e.g., Requests, College Reportability)

For the process aspect of comparability, NCEO and panel members found there is not complete comparability between the FSA, and ACT and SAT. The accommodations request process is a little more cumbersome for the ACT and SAT than for the FSA, given that accommodations for the FSA have already been determined through the IEP process. However, for school-day administrations both ACT and College Board will require each school that is established as a test site to have a Test Accommodations Coordinator (TAC), who will facilitate the accommodations request and appeals process for the students in the school. Many schools already experience the accommodations request process for students with disabilities and ELs when they request accommodations for the national test day administration. NCEO believes the request process becomes an issue when the individuals assigned to process requests for the school do not have the experience or time to provide convincing documentation of the need for accommodations. This may occur in resource-poor schools.

NCEO and most members of the educator panels believe that both ACT and the College Board provided testing accommodations that would give students with disabilities the opportunity to participate in these assessments. However, they also found that the use of some accommodations could result in students with disabilities and ELs receiving scores that are non-college reportable and cannot be used for college admissions purposes.

The accommodations decision-making process used by ACT and College Board (for the SAT) are not fully transparent. School staff can request accommodations for students with accessibility needs and complete required documentation, yet the accommodations are not always approved as college-reportable accommodations. Neither ACT nor College Board makes the criteria they use to approve or reject accommodations requests publicly available. This is different from the FSA process, where 504/IEP/EL teams or other school personnel who know the student are the final decision makers for all accommodations except those that are special requests, which is a small minority of the accommodations. The use of some accommodations on the ACT and SAT always result in non-college reportable scores for students with disabilities. For example, the use of the American Sign Language (ASL) accommodation to deliver test content will result in non-college reportable scores.

Cultural Sensitivity

Although cultural sensitivity was identified as an important aspect of comparability, there was not a systematic discussion of this aspect. Some educators noted concerns about the cultural sensitivity of the FSA, but there was no way to judge any differences across tests. Given typical cultural sensitivity approaches of test developers, NCEO concluded that the three tests are likely comparable, although FSA may be more Florida-centric.

Overall Conclusion

Based on its knowledge and 27 years of experience with accommodations policies, and the input from Florida educators, NCEO concludes that in many ways, in terms of the provision of accommodations, the ACT and SAT could provide comparable benefit to the FSA for purposes of school accountability and graduation, although this was less evident for ELs for the SAT (and in general across the assessments). In general, both ACT and College Board indicated that they would provide greater numbers of the accommodations in the NCEO list of accommodations than were provided for the FSA. Whether these differences were appropriate for the Florida standards was not addressed in these studies.

Comparability in the process for accommodations requests was less clear and often relevant more to the use of the tests for college entrance; comparability to the FSA cannot be judged here because the FSA does not provide a score that can be used for college entrance. Still, if a district is basing a decision to use one of these tests in lieu of the Florida assessments on the possibility of having college entrance scores for all of its students, this goal is unlikely to be realized for some students with disabilities and ELs. The lack of transparency in the decision-making process about which specific accommodations would result in a college reportable score for which specific students is likely to result in non-comparability for some student groups compared to other student groups, which could be a concern when making the decision about whether to allow Florida districts to use either the ACT or the SAT in lieu of the Florida assessments.

Introduction

Some students with disabilities and English learners (ELs) use accommodations so that their knowledge and skills can be appropriately measured during assessments; these accommodations produce scores that support valid interpretations about the students' knowledge and skills.

The National Center on Educational Outcomes (NCEO) conducted Accommodations Studies for the State of Florida. NCEO is an independent non-profit organization at the University of Minnesota that focuses on the inclusion of students with disabilities, ELs, and ELs with disabilities in comprehensive assessment systems. This work was part of a larger effort by the Assessment Solutions Group (ASG) for the Florida Department of Education to determine whether it should allow high schools in the state to choose to use the ACT or the SAT in lieu of the Florida Standards Assessments.

For these studies, NCEO conducted research and analyses to evaluate the degree to which the ACT or the SAT "provides accommodations that permit students with disabilities and English learners the opportunity to participate in the assessment and receive comparable benefits" (Bureau of Contracts, Grants, and Procurement Management Services, 2017, p. 7). The charge was further described in the Florida Department of Education request for proposals (RFP):

> "C. Criteria 4 – Accommodations:
> The Contractor shall organize and conduct in-person studies in an agreed-upon Florida location in which Florida educators and other relevant experts evaluate the degree to which the ACT and the SAT meet Criteria 4 for their suitability for use in lieu of the grade 10 statewide, standardized ELA assessment and the Algebra 1 EOC assessment."
> (Bureau of Contracts, Grants, and Procurement Management Services, 2017, p. 16)

Federal Legislations Considerations

The need to examine accommodations for ACT and SAT emerges from new language in the 2015 reauthorization of the Elementary and Secondary Education Act (ESEA) as the Every Student Succeeds Act (ESSA). ESSA included the opportunity for states to allow districts to use locally-selected, nationally-recognized high school academic assessments in place of the statewide assessment. ESSA required that these assessments provide comparable, valid, and reliable data on academic achievement for all students and for each subgroup of students. Before approving any of these assessments, states must ensure that the use of appropriate accommodations does not deny students with disabilities or ELs the benefits of participation in the assessments that are provided to students without disabilities or who are not ELs.

The participation of students with disabilities in all state- and district-administered assessments is required by the Individuals with Disabilities Education Act (IDEA), which also requires that they be provided accommodations as appropriate. IDEA requires that the Individualized Education Program (IEP) include "any individual appropriate accommodations that are necessary to measure the academic achievement and functional performance of the child on State and districtwide assessments."

ESSA requires the participation of all students in assessments used for accountability purposes. It also requires that students with disabilities have appropriate accommodations as required by IDEA, and that ELs be assessed in a valid and reliable manner and provided appropriate accommodations.

The Americans with Disabilities Act (ADA), which applies to both K-12 and post-secondary public education institutions, addresses accommodations for individuals with disabilities who are seeking entrance to an institution of higher education and who must earn a reportable score on a college entrance examination. In 1990, when ADA was first enacted, it required responses to four questions: (a) Does the individual have an impairment that affects one or more major life activities as identified under ADA? (b) Does the impairment rise to the level of a disability? (c) What is the impact of the impairment on performance in the area for which accommodations are being considered? and (d) What are the appropriate accommodations, given the specific tasks required of the individual? ADA applies to individuals with impairments that substantially limit major life activities. Disabilities that it covers include a substantial hearing or visual impairment, intellectual disability, or a specific learning disability. It does not cover individuals with minor conditions of short duration (e.g., broken arm, flu, etc.).

Technical assistance guidance from the U.S. Department of Justice (2015) defined what was intended for the provision of testing accommodations. It clarified that the tests covered included "exams administered by any private, state, or local government entity related to applications, licensing, certification, or credentialing for secondary or postsecondary education, professional, or trade purposes…" (p. 2). It also clarified that the following were sufficient documentation of the need for accommodations:

- Past testing accommodations on similar standardized exams or high-stakes tests
- Formal public school accommodations (e.g., IEP- or Section 504-documented accommodations)
- Documentation from a qualified professional

Although federal laws continue to refer only to "accommodations," state assessments, including Florida's high school assessments, use a broader accessibility framework to ensure that all students, including students with disabilities and ELs have appropriate access to assessments. Table 3-1 presents accessibility terms and definitions that are used in states and in this report. It also presents terms that are used to address whether accommodations from locally-selected, nationally-recognized tests produce scores that can be used for college entrance.

Table 3-1. Definitions

Accessibility Terms and Definitions
Accessibility Features – Test features that may be used by any students with accessibility needs. It includes both universally available features, and those that must be turned on by an adult.

Accommodations/test accommodations – Adjustments that do not alter the assessed construct that are applied to test presentation, environment, content, format (including response format, or administration conditions for particular test takers, and that are embedded within assessments or applied after the assessment is designed. Tests or assessments with such accommodations, and their scores, are said to be *accommodated*. Accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together. (AERA, APA, NCME, 2014, p. 215)

Modification/test modification – A change in test content, format (including response formats), and/or administration conditions that is made to increase accessibility for some individuals but that also affects the construct measured, and consequently, results in scores that differ in meaning from scores from the unmodified assessment. (AERA, APA, NCME, 2014, p. 221)

Tiered system of supports – A type of accessibility framework that includes accessibility features and accommodations.

Universally available features – Accessibility tools that any student may use.

Terms Used to Indicate Whether Accommodations from Locally-Selected, Nationally-Recognized Tests Can be used for College Entrance

College reportable accommodations – Accommodations approved by ACT/College Board that will provide college reportable scores.

Non-college reportable accommodations (State-allowed accommodations) – Accommodations not approved by ACT/College Board. If the accommodation is used, the scores might be used for state purposes (accountability, graduation), but the score is not college reportable. A state-allowed accommodation can be an accommodation that ACT/College Board does not approve for a specific student, or it can be a specific accommodation that Florida and ACT/College Board decide will always be considered state allowed.

Table 3-2 indicates the number of students who had the opportunity to use certain accommodations on the Florida grade 10 reading and writing ELA assessments and the Algebra 1 EOC assessment. The table includes only numbers for some accommodations, especially those that are (a) documented and enabled in advance so that they are available on the online platform or (b) alternate formats that are ordered in advance (e.g., braille form). It does not include data about the number of students who use many of the most common accommodations (e.g., extended time, breaks, multiple days, small group administration, individual administration, etc.).

Table 3-2. Number of Students with Certain Accommodations on the FSA Grade 10 ELA and Algebra 1 EOC Assessments, Spring 2017 Administration

Accommodation	Grade 10 ELA (reading)	Grade 10 ELA (writing)	Algebra 1 EOC
Text to speech (TTS)[1]	4,090	3,711	3,760
American Sign Language (ASL) enabled reading[1]	127	--	--
Closed captioning (CC) enabled reading[1]	148	--	--
Both ASL and CC[1]	*88*	--	--
TTS, ASL, and CC[1]	29	--	--
Alternate Formats			
Paper test/answer book (regular print)[2]	1,496	1,461	1,741
Paper test/ answer book (large print) [2]	54	49	108
EBAE Contracted Braille[2]	7	8	8
EBAE Uncontracted Braille[2]	2	3	3
UEB Contracted Braille[2]	0	0	2
UEB Uncontracted Braille[2]	[2]0	0	14
Paper test/ answer book (one item per page) [2]	4	4	7
Reading/writing passage booklet (regular print) [2]	470	410	--
Reading/writing passage booklet (large print) [2]	5	8	--
Reading audio passage transcript & animation stimuli book[2]	166	--	--

Source: Data provided by the Florida Department of Education.

[1] Accommodation requested and enabled. Does not indicate whether the accommodation was actually used.

[2] Alternate format materials ordered. Does not indicate whether the accommodation was actually used.

Process

The Accommodations Studies were designed to obtain input from Florida educators on accommodations for the ACT, SAT, and Florida Standards Assessments. Toward this end, NCEO developed materials for educators to review and a process to obtain their input. The 1½-day educator panel meetings were held on October 26-27, 2017, in Orlando, Florida. This section describes the roles of participants attending the meeting, the materials provided, and the meeting agenda and procedures.

Participant Roles. Sixteen educators participated in the meetings. Two panels were assembled, one for each content area (ELA, math). Each panel had eight members. Participants were selected to represent five roles for each content area (ELA, math): (a) special educator (Exceptional Student Education – ESE); English learner educator (English for Speakers of Other Languages – ESOL); blind/low-vision educator; deaf/hard of hearing educator; and content educator. The name and roles of the participating educators are shown in Table 3-3. Additional details about the educators, including their schools, are included in **Appendix 3A**.

Table 3-3. Participants in Accommodations Studies Educator Panels

Grade 10 ELA		Algebra 1 End of Course	
Educator	**Role**	**Educator**	**Role**
Carden, Rene	VI	Albritton, Keisha	Math
Conover, Zoe	ESE	Clark, Sue	DHH
Conrad, Sherry	DHH	Elliot, Marion	Math
Hodges, Kenny	ESE	Haines, Kathy	VI
Holley, Mary	ESOL	Munn, Paley	ESE
Sanabria, Izzy	ELA	Salamone, Chris	ESE
Sherlock, Jean	ELA	Valentine, Joshua	ESOL
Zuaro, Elisha	VI	Wilson, Katherine	ESOL

ESE – Exceptional Student Education; ESOL – English as a Second or Other Language; VI – Vision; DHH – Deaf/Hard of Hearing; Content – ELA or Math

In addition to the educators, representatives of the Florida Department of Education, ACT, and College Board (for SAT) participated in part of the meeting. The individuals attending from each organization are included in Table 3-4.

Table 3-4. Other Representatives Participating in Parts of the Accommodations Studies Meeting

Florida Department of Education	ACT	College Board
Victoria Gaitanis	Katie Featherson	Sharon Cowley
Kath Visconti	Gaye Fedorchak	John Fallon

The meeting was facilitated by NCEO staff: Sheryl Lazarus, Martha Thurlow, Linda Goldstone, and Chris Rogers. It was also attended by an ASG representative: Ed Roeber.

Materials. Meeting materials included summaries of each of the tests under consideration for use in lieu of the Florida high school assessments. For each assessment (ACT, SAT, Florida Standards Assessment – Grade 10 ELA and Algebra 1 End of Course), NCEO developed a summary that included a general description of the assessment and answers to eight questions:

a. Is a tiered accessibility framework used?
b. Does the use of some accessibility features or accommodations result in scores that are not college reportable?
c. Which student groups can receive accommodations?
d. What is the accommodations-request process?
e. What is the process to appeal a denied accommodation?
f. How likely is it that an accommodations request will be approved?
g. Is a list of accommodations provided or are examples given?
h. What accessibility features and accommodations are available?

The NCEO-developed summaries are included in **Appendix 3.B**, Documents B-7, B-8 and B-10) with the other meeting materials. Additionally, a supplemental document containing additional information provided by ACT was included in the meeting materials (Document B-9). NCEO also provided sample crosswalks of the accommodations request process and the universal

design features and accommodations provided for each assessment (see Documents B-11 and B-12). The sample crosswalks of features and accommodations had been developed for earlier work in which NCEO compared assessments for ACT and SAT with those provided by the Partnership for Assessments for College and Career Readiness (PARCC) and by the Smarter Balanced Assessment Consortium (Smarter Balanced); as a result, the crosswalks were limited to the 74 features and accommodations for students with disabilities and 32 features and accommodations for ELs included in those assessments and that were included in the three assessments (i.e., FSA, ACT, SAT) included in this analysis. The crosswalk tables should **not** be considered as defining the universe of features and accommodations for these populations.

Two sets of discussion questions were provided to participants for content-specific discussions (see **Appendix 3.B**, Documents B-13 and B-14, for the sets of questions). The first set of content-specific questions focused on the accommodations request and approval process, and included the following six questions:

1. How streamlined is it to submit an accommodation request for this assessment? (- for students with disabilities? – for ELs?)
2. How does it differ from the request/identification process for the current Florida assessment? (- for students with disabilities? – for ELs?)
3. How streamlined is the accommodations approval process?
4. How streamlined is the process to appeal a denied accommodation?
5. How streamlined is the unique or other accommodations special request process (Assessment being reviewed? Current Florida assessment?)
6. Other comments

The second set of content-specific questions focused on the specific accommodations available, and included the following five questions:

1. How do the accommodations compare across the assessments being reviewed (i.e., ACT, SAT) and current Florida assessment? (- for students with disabilities? –for ELs)
2. How accessible are alternate forms of the assessment (e.g., braille, large print, fewer items per page, etc.?) (Assessment being reviewed [i.e., ACT, SAT]? Current Florida assessment?)
3. Are some accommodations not college reportable? When does this occur?
4. How do the benefits that students with disabilities and ELs would receive from this test compare with benefits other students (i.e., those without identified disabilities or EL status) receive? (Assessment being reviewed – students with disabilities? ELs?) (Current Florida Assessment – students with disabilities? ELs?)
5. Other comments

In addition to these questions used with the content-specific groups, a set of questions was used for an entire group discussion, with the Grade 10 ELA and Algebra 1 End of Course groups together. The following questions were used for this whole group discussion (see **Appendix 3.B**, Document B-15, for the set of questions):

1. How does the accommodations request/identification process differ across the ACT, SAT, and current FDOE assessments for (- students with disabilities? – ELs?)
2. How well do the accommodations across ACT, SAT, and current FDOE tests enable student participation in these assessments for (- students with disabilities? – ELs?
3. What are the implications or consequences, if any, if test scores are not-reportable?
4. Do students with disabilities receive benefits comparable to other students for (-ACT; - SAT; - Florida Assessments)?
5. Do ELs receive benefits comparable to other students for (- ACT; - SAT; - Florida Assessments)?
6. Other comments

Finally, a questionnaire was used at the conclusion of the meeting to obtain (a) educator opinions about the three assessments, and (b) an evaluation of the meeting process. This questionnaire is included in **Appendix 3.B** (Document B-16).

Meeting Agenda and Procedures. The full agenda for the meeting is included in **Appendix 3.B** (Document B-1). Prior to beginning their participation in the meeting, all participants signed Florida non-disclosure forms. Non-disclosure forms were not required for the ACT and SAT because no live items were presented by ACT or SAT. All participants also signed in each day.

During the morning of Day 1, all participants were together, including representatives of the Florida Department of Education, ACT, and College Board. During the morning hours, participants were informed of the purpose of the Accommodations Studies; they were also provided overviews of each assessment by representatives of the Florida Department of Education, ACT, and College Board.

Prior to lunch, the test representatives were excused from the room. During this time, educators were provided with an overview of the afternoon tasks, including a reminder of the purpose of the work to be completed and an introduction to all materials that were available to review. Educators also were informed of the processes that would be used during their content-specific work.

Most of the afternoon work involved the content-specific groups reviewing summaries of the tests and test materials and discussing the content-specific questions. For each topic (accommodations process; accommodations available), each content-specific group first divided into two smaller groups, with each group focused on one of the assessments under consideration to be used in lieu of the Florida Standards Assessments (ACT or SAT). Discussions were guided by the questions provided. At the end of each topic discussion, the two groups came back together to discuss overall responses to the questions, building on the observations of both groups. During their discussions, and at the end of the final cross-test group discussion, educators developed questions for the representatives from the Florida Department of Education, ACT, and College Board.

All groups came back together at the end of the first day to pose their questions to Florida, ACT, and College Board. This took place for approximately 60 minutes. ACT and SAT were excused

from any further participation in the meeting, although they remained available during Day 2 in case additional questions arose.

The first day of the meeting concluded with a quick wrap-up of the day's activities and a reminder of the purpose of the educator panels. A quick overview of the work to be done on Day 2 was also presented.

On Day 2, educators were given the opportunity to reflect on the first day's activities and to share any "aha" moments they had since Day 1. The educators were informed of the decision to summarize the questions that had been posed the previous afternoon and submit those for ACT and College Board to provide written responses. They were also informed that ACT and College Board would be asked to use the crosswalk that contained the approved Florida accommodations to indicate which would and would not be likely to result in college reportable scores. Then, the purpose of the Accommodations Studies was reviewed. A discussion of "comparable benefit" followed, with the educators providing their perspectives on key aspects of comparable benefit.

For the remainder of the Day 2 meeting, educator panel members shared information by responding to discussion questions. They also were asked to discuss possible positive and negative consequences (e.g., pros and cons) of using the ACT or SAT in lieu of the Florida Standards Assessments.

The meeting concluded with a brief discussion of the format of the report. Panelists then completed a questionnaire about their individual opinions about the assessments, followed by a set of meeting evaluation questions (see **Appendix 3.B**, Document 3B-16).

Following the Accommodations Studies meeting, NCEO staff and Ed Roeber from ASG summarized the questions and crosswalk grid (see **Appendix 3.C**, Documents C-1 and C-2) for ACT and College Board, and sent those to them on October 30 for response by November 6. In addition, NCEO staff assembled notes from all the discussions and tabulated the questionnaire and evaluation results.

Educator Panel Findings

The results of the educator panel discussions are presented here. First are the results of the panel discussions about the meaning of comparable benefits. Next are results of the discussions of the two content-specific groups – math (Algebra 1 EOC) and ELA (grade 10 ELA); these are divided into the accommodations request **process discussion** (for students with disabilities and ELs) and the **accommodations allowed discussion** (for students with disabilities and ELs). After that, results of discussions across the content groups are presented, including written information obtained from ACT and SAT in response to questions posed by the educator panels.

Comparable Benefits Discussion

The panels were tasked with evaluating whether ACT and SAT have accommodations that provide students with disabilities and ELs the opportunity to participate in the assessment and

receive comparable benefits, and the degree to which ACT and SAT are suitable for use in lieu of the Florida grade 10 statewide, standardized ELA assessment and the Algebra 1 EOC assessment. To do this, the panelists needed to reach a common understanding of what comparable benefits meant.

Following a discussion, the panel members identified four considerations when evaluating whether the ACT and SAT offered comparable benefits: First, they noted that the uses of assessment results from ACT and SAT should be the same as those for which FSA results were used. These uses included school accountability, graduation determination, and programmatic changes. Second, panel members thought that it was important to have appropriate accommodations allowed for use when taking the ACT and SAT. They also noted that these might be accommodations in addition to those allowed for the FSA. Third, educators were interested in exploring the process that was used to obtain accommodations for all assessments, with the hope that the processes would be equally easy for all assessments and that they would result in scores that could be reported (specifically for college entrance). Finally, educator panel members talked about the importance of cultural sensitivity for comparability of benefit. During the discussions, not all of these considerations were addressed to the same extent. Specifically, most of the panels' discussions focused on comparable benefit in terms of allowed accommodations and the request process.

Math Panel Process Discussion
When reviewing and discussing the accommodations request and appeal process, math educator panel members noted that for the FSA, students are entered into the system once. Entry of information does not need to be repeated annually. The panel members also noted that both ACT and SAT used an online approval process. For both ACT and SAT, time for approval was a limiting factor and a concern. For the ACT, it took seven to ten days to receive notification of whether an accommodation was approved. For SAT, it typically took up to seven weeks to receive notification about the approval of an accommodation, which panel members thought was excessively long and might have an impact on the timing of IEP team meetings. The approval process requires entering one student at a time into the system.

Panel members expressed concern about the documentation required for the appeal process when an accommodation was not initially approved. Some educators on the panel had prior experience requesting accommodations for the national administrations of the ACT or SAT, and they brought those experiences into the panel discussion. Educator panel members noted that they would find it helpful to know the reason why a request is denied, something that they had not experienced with the national administrations. They also thought it would be helpful to receive detailed information on the additional information needed for an appeal.

Students with Disabilities. Educator panel members identified two areas where the accommodation request processes of ACT and SAT for students with disabilities differed from the FSA process. The first difference was that ACT and SAT have a request process where the final decision about which accommodations are allowed is made by the vendor rather than the 504 plan or individualized education program (IEP) team. For the FSA, the state publishes a list of accommodations and IEP teams use the list to make informed accommodations decisions for

the assessment for individual students. The IEP team is the final decision maker for most FSA accommodations.

For both the ACT and the SAT, an accommodations request must be submitted to the test vendor and the vendor either approves or does not approve the request. Both the ACT and SAT ask for student disability category as part of the request process. The educators noted that the disability categories listed in ACT and SAT materials did not align with the federal disability categories used by the Florida Department of Education. The ACT drop-down menu made it relatively easy to identify which accommodations might be requested for students with visual impairment (VI), but it was more difficult to determine the options from a drop-down menu for other types of disabilities because there was a need to guess which disability category to select. For requests for accommodations not listed in the ACT policy (i.e., special or unique accommodations), educators thought that ACT did not provide sufficient information about the request process for these accommodations.

For some students who had previously approved accommodations, SAT requires that multiple types of documentation be included with a request each year, including frequent updates of a current medical diagnosis, a current IEP, documents indicating current use of the accommodation during instruction, and some teacher input about whether the accommodation request aligned with what was in the student's IEP. For many students with disabilities, once a student's request for accommodations for the SAT is approved, the approval remains valid throughout their high school career. College Board provided detailed information about how special or unique accommodations could be requested for the SAT.

The second difference between the Florida request process, and the ACT and SAT request processes, was the amount of time required for accommodations decisions. Educator panel members noted that it took less time to select accommodations for the FSA than for the ACT and SAT. Even though the ACT request process was considered relatively easy, educators noted that it could be a time-consuming process to make separate applications for each student because there was a need to identify the accommodations for each student using an online tool, and the process needs to be repeated for each content area (i.e., reading, writing, math). This potentially would be very time-consuming, especially in large school districts. For the SAT, teachers found it "interesting" that for each of the disability categories, evaluations were required by people with professional credentials who often would not be school staff. Again, this process was a concern because it could be a very time-consuming process. In comparison, FSA was more streamlined and automatic; FSA accommodations are provided by the school and documented in the testing paperwork but no 504 or IEP plan is ever uploaded into the FSA Portal Tests Information Distribution Center (TIDE). The IEP or 504 plan team is the final decision maker for which accommodations students with disabilities receive on the FSA.

English Learners (ELs). Math educator panel members noted that for the FSA, local teams make accommodations decisions for ELs. An application was not required for each student, and only a parent needed to indicate yes/no on a consent letter. In comparison to the FSA, for the ACT and SAT, an application needs to be made for each student. Educators noted that the ACT and SAT request process for ELs seemed relatively simple. Still, although the College Board provided criteria to be used when making requests, there was little information on how student

eligibility was determined, what accessibility features and accommodations were available, and what would be approved.

Concern was expressed that for the SAT, accommodations for ELs were available only on school test dates, and were not available on Saturday test dates. It was noted that information provided by the College Board indicated that only 87% of accommodations requests were approved. According to information provided by ACT, 98% of requests were approved in 2016-17. Although both ACT and College Board provided data on approval rates, it was unclear whether these rates were realized only after multiple appeals processes. It was also noted during the discussion of the approval process that ELs who had extended time would not receive a college reportable score. Accommodations do not need to be approved for the FSA. The State provides a list of accommodations, and then the EL planning team, or other educators who know the student make the accommodations decisions.

Math Panel Accommodations Discussion
Educator panel members noted that a broader range of accessibility features and accommodations was available for the FSA than for the ACT or SAT. ACT provided a list of accommodations. The College Board did not provide a comprehensive list of accommodations, but rather provided examples, so it was difficult to determine whether the SAT would provide fewer accommodations without additional information. On the ACT, there were fewer universally available features and designated supports that any student could use than there were for the FSA.

Students with Disabilities. There was discussion about variations across assessments in accommodations for students with visual impairments (VI). Educators expressed some concerns for students who use braille because manipulatives are not allowed on the ACT and SAT, whereas they are permitted for the FSA. (Written responses from ACT and College Board following the educator panel meetings indicated that manipulatives are reviewed on a case-by-case basis. College Board provided the following example: "A student who needs geometric shapes to assist in interpreting math and science diagrams could be approved for college reportable scores"). (See **Appendix 3.D**, Document D-2 for details.)

English Learners. Educator panel members described accommodations for ELs on ACT compared to FSA. Still, the provided accommodations chart had blanks that were unclear. FSA allowed a proctor to be present who speaks the heritage language; nothing similar seemed to be provided by ACT or SAT.

For the ACT and SAT, the use of some accommodations (those not approved by the vendor, but allowed by the state of Florida) would result in a score that was not college reportable. Some panel members did not think that it mattered whether students' scores were college reportable if the primary purpose was to obtain data for accountability purposes. But, if a student plans to attend college, the student would need to retake the test without using state-allowed accommodations, and this might be at cost to the student. Compared to students who are not ELs, ELs did not receive a reportable score if they used some accommodations, so they would need to retake the test. For example, ELs using extended time on the SAT would not receive a college reportable score. Also, breaks were not an allowed accommodation. For ACT, the score

would also be non-college reportable if breaks were provided. Though a word-to-word glossary was allowed, a word-to-word dictionary was not on the SAT (and would result in a non-college reportable score on the ACT), whereas the FSA allowed both. SAT did provide written directions in the heritage language, but FSA does not.

Concern was expressed that if a purpose of school administrations of the ACT and SAT was to level the playing field for students who may not otherwise take a college entrance exam, this purpose would not be achieved because some students who used state-allowed accommodations would have a non-reportable score for the ACT or SAT.

English Language Arts Panel Process Discussion

Students with Disabilities. Educator panel members identified points for consideration for the use of either ACT or SAT in lieu of the current FSA system. They indicated that if the ACT or SAT is used, there would be a need to prepare all teachers, especially general educators, to gather needed information as evidence for requesting accommodations for students with disabilities. Special educators noted that they have used online IEP tools such as the Florida Department of Education's online IEP-writing software, "PEER," and additional online IEP-writing software such as, "Frontline" (formerly "Enrich"), and perceived that these might be useful tools for this purpose.

Some educator panel members indicated that the accommodations application process for ACT and SAT can be seen as a difficulty or barrier, yet they also noted that with a clearly established routine and schedule, issues could be minimized. They voiced uncertainty about what would demonstrate proof of use of accommodations during classroom instruction, which was required by the College Board. Panel members who examined the ACT request process commented that the digital format sounded sufficiently streamlined based on the information provided, yet they indicated that they would like to see the portal to understand how it actually works. Panel members who reviewed the College Board request process indicated that it seemed user friendly according to the screen shots that they viewed, for accommodations for both students with disabilities and ELs, but that it was difficult to tell without actually having access to the portal.

English Learners. Educator panel members noted that there is not a formal process for seeking accommodations for individual students on the FSA, other than reporting EL status on a few documents associated with student rosters. They voiced uncertainty about the documentation that would be needed for the ACT and College Board request processes. Panel members indicated that they wanted to know more about how parent involvement and consent would work, given parents' use of heritage languages. Educator panel members noted that the online process for the ACT and SAT tests provides some transparency and clarity, yet that reasons for denials need to be clearly communicated to all parties. They also indicated that they had varying degrees of familiarity and personal experience with the appeal process when seeking unique accommodations for students with disabilities for the FSA.

Teachers with experience requesting unique accommodations offered their own experiences. They noted that some information has been offered about the general steps and timelines for the appeal process, yet also indicated that further information was needed for the ACT and the

College Board, such as the range of time frames (rather than only average length of time for appeal process decisions). Educator panel members also expressed uncertainty about documentation required for the appeal process for the ACT and College Board. Other concerns were voiced about parents having enough information to make informed decisions about whether to have their child take tests using accommodations that would render scores not reportable to higher education institutions, but that would still result in meeting the assessment requirements for graduation.

ELA Panel Accommodations Discussion

Educator panel members noted that the opportunity for students to obtain a test score for higher education admission is a welcome benefit of taking the ACT or SAT in place of the FSA. Still, they recognized that some students with disabilities and ELs will not receive college-reportable scores. They noted that it would be very important for IEP teams to be made aware of which state-approved accommodations could render students' ACT or SAT scores not reportable for college admissions purposes. They noted that having this additional consideration would complicate unusual test-day events – such as a student inadvertently using an accommodation (e.g., a magnifier) on test day that was not pre-approved.

Students with Disabilities. Educator panel members examining the accommodations for the ACT and FSA indicated that in addition to the accommodations on the list, there was the possibility of gaining approval for other accommodations. Panel members examining the accommodations for the SAT expressed uncertainty about which accommodations can be provided for the SAT because only examples were provided, not a comprehensive list.

Educator panel members had concerns about the availability of alternate forms of testing for students with sensory impairments. Based on information provided, they thought that ACT and SAT might better meet the individual needs of students with sensory impairments, especially visually impaired students, than the FSA. Panel members noted that presenting the reading passages through sign language appeared to be permitted for the ACT and SAT on the ELA test, an accommodation that is not allowed on the FSA; still, only Signed Exact English and American Sign Language (ASL) were provided on the ACT and SAT tests. Additionally, the use of ASL results in a non-college reportable score on both assessments. The preference of the ACT and SAT for Signed Exact English over ASL was discussed at length and concerns over equitable student access were noted by educators.

English Learners. Educator panel members wanted to know for both the ACT and SAT whether information about accommodations was provided in native languages to parents seeking accommodations for their EL children. This was not clear.

Findings Across Content Areas

After the educator panels for each content area completed their separate analyses, the groups re-convened as one large group. Representatives of the Florida Department of Education, ACT, and the College Board (for SAT) also participated in this session. The combined group of educator panel members raised many questions. ACT and College Board representatives responded verbally to the questions. To ensure that accurate responses are reported here, the

educator questions were compiled into a list that was submitted to each organization for written responses. The questions were:

1. What accommodations are permitted for ELs?
2. What documentation is required to document the need for accommodations for ELs?
3. Does the district need to re-submit the accommodations request annually?
4. How can parents who do not speak English access information on the accommodations process and give consent to those requested and to be used?
5. When sensitive data (e.g., medical records, etc.) are being sent to support the accommodations request, what assurances are there that these data will remain secure?
6. What percent of accommodations requests are approved initially? What percent of additional accommodation requests are eventually approved (after one or more additional submissions of information)? What percent of accommodations requests are ultimately not approved?
7. What is the turn-around time for accommodation requests to be approved?
8. What is the turn-around time for unique/other/special accommodations to be reviewed and decided on?
9. How many unique/other/special accommodations are requested per year and what percentage of these is approved?
10. What is the appeal process for an accommodation request that was initially not approved?
11. If an accommodation is not approved, will schools be given a reason why it was not approved (and how they might appeal the decision)?
12. Why are there different criteria and different timelines for review and decisions for different disability categories?
13. Can educators or parents request any accommodations?
14. What is the process for parents to ask for an accommodation when the school is not asking for it?
15. If parents can request accommodations separately from the school, are schools required to provide accommodations that parents requested during school day testing, if approved by your organization?
16. If an accommodation is approved for individual administration, but a small group of students all require this same accommodation, can the assessment be administered to the small group and will the scores from each student be college reportable?
17. If a case goes to due process, there may be a need to prove that accommodations were provided. What proof will be available that an accommodation was actually provided?
18. Are there manipulatives that may be used with the braille tests without invalidating the scores from them?
19. Are unique seating arrangements allowed for students who are visually impaired?
20. Which sign languages result in college reportable scores for the students who are deaf and which do not?
21. If a requested accommodation is deemed to be a State Approved Accommodation that would not result in college-reportable scores for the student, is the IEP team (educators and parents/guardians) given the option of either proceeding to administer the test with the accommodation that does not yield college-reportable scores, OR having the student not use the accommodation and be eligible for college-reportable scores?

22. If a student first takes the assessment with an accommodation that would not result in a college reportable score, and then later takes the same test without that accommodation (so as to receive a college reportable scores), what is included in the school report that is sent to colleges regarding prior participation with non-reportable score, etc.?

See **Appendix 3.D**, Documents D-1 and D-2 for the full written responses of ACT and College Board.

Some of the questions raised by educator panel members were policy decisions that Florida will need to make:

1. What is the accommodations request process for students who move from one Florida county to another where a different assessment (e.g., the FSA or a college entrance test) is used for graduation?
2. Do ELs have to take the math assessment if they have taken three-fourths or more of the Algebra 1 class?

ACT and College Board were also asked to complete tables that indicated which accessibility features and accommodations are allowed on their respective assessments, and whether those accessibility features and accommodations were considered college reportable or non-college reportable for students with disabilities and ELs who use those accommodations. Because the exact terminology used to describe accommodations varied across assessments, NCEO used the accommodations nomenclature used in Lazarus and Thurlow (2016) when making the request. See **Appendix 3.D**, Documents D-1 and D-2 for ACT and College Board responses.

Tables 3-5 and 3-6 show the number of accommodations offered by the FSA, ACT, and SAT. These numbers were derived by summing the number of universally available features, other accessibility features, and accommodations for each assessment (see definitions in Table 3-1); this approach was used because each assessment had a somewhat different categorization schema. For the ACT and SAT, the numbers are based on the tables in the ACT and College Board responses in **Appendix 3.D**, Documents D-1 and D-2 (combining college-reportable and non-college reportable accommodations). For the FSA, the numbers are based on information in the *2017-18 FSA Accommodations Guide* (Florida Department of Education, 2017). NCEO compiled summary tables and verified these tables with the Florida Department of Education (see **Appendix 3.B**, Document B-7).

Table 3-5. Summary of the Number of Accommodations and Accessibility Features Offered to Students with Disabilities on the FSA, ACT, and SAT[1, 2]

	FSA ELA	FSA Algebra 1	ACT Reading & English	ACT Math	SAT
Total available accommodations [3]	74	74	74	74	74
Number of accommodations for students with disabilities on assessment	52	56	69	74	69
Number of accommodations not offered for assessment	22	18	5	0	5

[1] Numbers in table include both college reportable accommodations and non-college reportable accommodations.
[2] Numbers in table include universally available features (features any student can use), other accessibility features (those identified for any student by an adult), and accommodations (available only to certain test takers, generally students with disabilities and ELs). See definitions in Table 3-1.
[3] This is based on the total body of accommodations in the assessments being analyzed (FSA, ACT, SAT), using the accommodations typology in Lazarus and Thurlow (2016). The total available accommodations should **not** be considered as defining the universe of features and accommodations for students with disabilities.

Based on the Lazarus and Thurlow (2016) typology, there were 74 possible accommodations for students with disabilities on high school assessments that were included in at least one of the three assessments included in this analysis. The FSA Grade 10 ELA assessment has 52 accommodations and the Algebra 1 End of Course assessment has 56. Based on ACT and College Board self-reports of accommodations that are available, both ACT and SAT offer more accommodations than FSA. ACT has 69 accommodations for Reading and English, and 74 for Math. SAT has 69 accommodations. (College Board did not disaggregate by content area.)

Again, based on the Lazarus and Thurlow (2016) typology, there were 32 possible accommodations for ELs on high school assessments that were included in at least one of the three assessments included in this analysis (see Table 3-6). The FSA Grade 10 ELA assessment has 14 accommodations and the Algebra 1 End of Course assessment has 15. Based on ACT and College Board self-reports of accommodations that are available, ACT offers more accommodations than FSA. ACT has 30 accommodations for Reading and English, and 31 for Math. In contrast, SAT offers eight accommodations. (College Board did not disaggregate by content area.). For ELs, direct comparisons cannot be made across assessments. College Board indicated that the accommodations that it identified for the SAT were for students who were EL only (i.e., not ELs with disabilities). The ACT list included some accommodations (e.g., food/medication for individuals with medical need, magnification, sign language directions only – American Sign Language) that would typically only be used for ELs with disabilities.

Table 3-6. Summary of the Number of Accommodations and Accessibility Features Offered to ELLs on the FSA, ACT, and SAT[1, 2]

	FSA ELA	FSA Algebra 1	ACT Reading & English	ACT Math	SAT
Total available accommodations [3]	32	32	32	32	32
Number of accommodations for ELs on assessment	14	15	30	31	8[4]
Number of accommodations not offered for assessment	18	17	2	1	24

[1] Numbers in table include both college reportable accommodations and non-college reportable accommodations.
[2] Numbers in table include universally available features (features any student can use), other accessibility features (those identified for any student by an adult), and accommodations (available only to certain test takers, generally students with disabilities and ELs). See definitions in Table 3-1.
[3] This is based on total body of accommodations in the assessments being analyzed (FSA, ACT, SAT) based on the accommodations typology in Lazarus and Thurlow (2016). The total available accommodations should **not** be considered as defining the universe of features and accommodations for students with disabilities.
[4] The list for SAT is for students who are EL only (i.e., not ELs with disabilities).

Educator Panel Recommendations

Pros and Cons

Educator panel members identified pros and cons of using the ACT or SAT in lieu of the FSA assessments of grade 10 ELA and Algebra 1 EOC. The identified pros and cons were sorted into three topic areas for presentation here: accommodations request and approval process; availability of accommodations; and testing context considerations.

Table 3-7 summarizes the educator-identified pros and cons of the three tests in relation to the **accommodations request and approval process**. As evident in this table, the educator panels tended to identify more cons for ACT and SAT compared to FSA.

Table 3-7. Pros and Cons of Request and Approval Process

	FSA	ACT	SAT
Pros	• Easy / no application and approval process for accommodations • Accommodation manual (including "how to")	• Quicker appeal decision (than for the SAT)	• No repeat (yearly) applications for most static / same accommodations • School statement regarding IEP/504 accepted (no upload)
	FSA	**ACT**	**SAT**
Cons	• "Delays" in getting scores for instruction, etc. • Limited practice opportunities with accommodations	• Need to plan ahead to apply for accommodations (in order to meet deadlines) • Must apply for accommodations yearly • Parents can apply for accommodations independent of school (compliance) • Parental accommodation requests – who responds? • Potential IEP compliance problems because IEP team is no longer the final decision maker • Earlier planning needed for instruction and transition • Additional teacher training needed (but less than SAT) • Confidentiality of student information uncertain	• Need to plan ahead to apply for accommodations (in order to meet deadlines) • Accommodation requests require parent signatures again and approval wait • Parents can apply for accommodations independent of school (compliance) • Potential IEP compliance problems since IEP team is no longer the final decision maker • Earlier planning needed for instruction and transition • Additional teacher training needed • Confidentiality of student information uncertain

As shown in the table, educator panel members identified several pros about the accommodations request and approval process for the FSA, ACT, and SAT. The FSA requires no additional process for accommodations already approved by the state. Accommodations decisions are made by IEP or EL teams. These procedures have already been established and the

procedures are presented in the state's accommodations manual. For the SAT, the additional request and approval process would not be required for most accommodations that students had already received in previous years (after the first year); that is, only requests for changes in the accommodations for students would be required. For the ACT, the appeal process decisions seem to be issued more quickly than for SAT. Another distinction between the ACT and SAT is that, for the SAT, evidence for requesting accommodations could simply be a school statement that the student has an IEP or 504 plan, rather than submitting the entire plan itself (which is required for the ACT).

Educator panel members also identified cons about the request and approval process. They were concerned about the limitations of each of the three tests, and expressed the need for these considerations to be addressed before decisions about tests that will be made available to districts.

They observed that the approval process for accommodations for both the ACT and SAT were less familiar to teachers who were more accustomed to decision making for the FSA, and that teachers will need training if the ACT or SAT is offered. They noted that recurrent annual approvals are required for the ACT, and that the parent needs to sign request documents for the SAT. Although the educators noted that these procedures could be perceived as difficult, they also suggested that these new processes could become routine and less onerous with practice.

Another concern expressed was that student information, particularly medical diagnosis information, would need to be submitted as evidence for accommodations requests for the ACT and SAT. Educators were concerned about data security and privacy issues.

Educator panel members also observed that the FSA has limitations, such as few opportunities for practice with realistic test items using accommodations; ACT and SAT were perceived to have more practice resources.

Table 3-8 summarizes the educator-identified pros and cons of the three tests in relation to the **accommodations allowed**. As evident in this table, educator panels reported many pros that were common across the ACT and SAT.

Table 3-8. Availability of Accommodations

	FSA	ACT	SAT
Pros	• Universal design features • Full-day / multiple day tests	• Variety of formats including oral presentation • More universally available features (than for the SAT) • Triple time (extended time) allowed	• Many accessible formats • Flexible accommodations "list" • Standard test – one day • Can separate component administrations
	FSA	**ACT**	**SAT**
Cons	• Limited accommodations for some disabilities (e.g., visual impairments) • Computer based challenging for some	• No ASL (except directions) • Limited accommodations for ELs • Uncertain which accommodations will be approved.	• No ASL • Limited extended time • Uncertain which accommodations will be approved.

According to the pros identified by educator panel members (shown in Table 3-8), ACT has a list of accommodations, while the College Board only provides examples of accommodations rather than a comprehensive list. ACT permits "triple time" (extended time that is three times the standard time limit), which panel members saw as helpful. Educator panel members also noted that the FSA can permit tests to be taken across the full day, and even across multiple days.

Panel members also noted that the FSA, as well as the ACT and the College Board, have incorporated universal design features. They noted that ACT seemed to have more universal features than the SAT. Educators observed that there are limitations on the availability of specific accommodations on each of the tests. For example, the FSA lacks some accommodations for students with visual impairments, ACT has limited accommodations for ELs, and both ACT and SAT do not allow American Sign Language (ASL). If ASL is used for SAT, the score is not college reportable; similarly, for ACT, ASL may not be used (except for directions).

Some general remarks about pros and cons reflected **testing context considerations** as indicated in Table 3-9. Several educator panel members observed that the use of a nationally recognized test (ACT or SAT) would make students' achievement scores comparable to those of students in other states across the nation.

Table 3-9. Testing Context

	FSA	ACT	SAT
Pros	• Designed to align to state standards and on-grade level • Consistency / familiarity across grades • No registration (systems communicate) • Testing capacity exists (apply accommodations; understand data) • Feedback from state regarding performance on standards • Programmatic changes can be made based on results • Florida test is from Florida educators • Florida-relevant (language, etc.) • Mostly computer based • Options – available on paper • Can get concordance scores on other tests	• Nationally recognized and scores same across nation • Continual test development • More consistency across time • Faster score reports (than FSA) • Multiple opportunities for higher scores regardless of score (low/high pass/fail) • Possibly higher motivation • Possibly provides college-reportable score • Practice opportunities • Multiple administrations • Commercially-available practice materials online and in print	• Nationally recognized and scores same across nation • Continual test development • More consistency across time • Faster score reports (than FSA) • Multiple opportunities for higher scores regardless of score (low/high pass/fail) • Possibly higher motivation • Possibly provides college-reportable score • More practice and supports (Khan Academy) • Multiple administrations • Commercially-available practice materials online and in print
	FSA	**ACT**	**SAT**
Cons	• Time consuming (administration) • Transparency of score values (item weight?) • "Delays" in getting scores for instruction, etc. • Less culturally responsive • Technology infrastructure limits	• Accountability (state and local) concerns • Not sure whether aligned to state standards (course) • Potential additional costs for certification • Test development disclosure	• Accountability (state and local) concerns • Not sure whether aligned to state standards (course) • Potential additional costs for certification

	• Limited practice opportunities with tests • No opportunity to get higher score unless you get level 1/2 only		

Educators identified pros for both the ACT and SAT relative to the FSA; they included, for example, greater access to test preparation resources and quicker score reporting. They also identified many pros for the FSA, including that the results provide feedback on students' performance on state standards and information for making programmatic changes. Additionally, they noted that the "Florida test is from Florida educators" and that it is more relevant to Florida, including language considerations, than either the ACT or SAT.

Educator panel members also identified some contextual cons though they recognized that the contextual issues were general concerns, and were not specific to students with disabilities or ELs. Educators voiced concern that districts have worked hard to arrange computer access for students on test days in order to implement the online FSA, but would have to revert to paper-pencil versions of ACT and SAT. On the other hand, they noted that an example of a con for the FSA in comparison to the ACT and SAT is that the technology infrastructure can break down. (The State of Florida is considering offering the paper and paper versions of the ACT and SAT so there would be no technology infrastructure to break down.) Educator panel members also noted a concern about whether and to what degree ACT and SAT align with the Florida content standards, as well as the potential additional costs for certifying administrators for administration of the ACT.

Educator Panel Conclusions

Educator panel members concluded their participation in the Accommodations Studies by completing a questionnaire about their conclusions. Summaries of their conclusions are presented in Table 3-10 for ACT. Disaggregated results for the ELA and math educator panels are presented in **Appendix 3.E**. The appendix also contains disaggregated data for "strongly disagree & disagree" and "agree & strongly agree." Participant comments also are also included in that appendix.

Table 3-10. Participant Questionnaire Responses for ACT *(N=16)*

Question	No Response (%)	Strongly Disagree & Disagree (%)	Neutral (%)	Agree & Strongly Agree (%)
1. The ACT provides testing accommodations that provide **students with disabilities** with the opportunity to participate in the ACT.		6	6	88
2. The ACT provides testing accommodations that provide **ELs** with the opportunity to participate in the ACT.		38	25	38
3. The ACT uses an appropriate approval process for accommodations for **students with disabilities**.		6	13	82
4. The ACT uses an appropriate approval process for accommodations for **ELs**.		25	13	62
5. **Students with disabilities** taking the ACT receive comparable benefits to participation in the Florida Standards Assessment.		25	19	57
6. ELs taking the ACT receive comparable benefits to participation in the Florida Standards Assessment.	6	75	6	13
7. It would be acceptable if the ACT was administered in lieu of Florida's grade 10 statewide, standardized English Language Arts (ELA) assessment and the Algebra 1 end-of-course (EOC) assessment.	12	25	13	50

Note: % may not total exactly 100% due to rounding.

As shown in Table 3-10, most educator panel members agreed or strongly agreed that ACT provides testing accommodations that provide students with disabilities with the opportunity to participate in the ACT, whereas almost two-thirds either disagreed with or were neutral to the statement that the ACT provides testing accommodations that provide ELs with the opportunity to participate in the ACT.

More than 80% of the participants either agreed or strongly agreed that ACT uses an appropriate approval process for accommodations for students with disabilities. For ELs, more

than 60% agreed or strongly agreed at ACT uses an appropriate approval process for accommodations; however, one-quarter of the participants disagreed.

More than 50% of the participants agreed or strongly agreed that students with disabilities taking the ACT receive comparable benefits to participation in the FSA; for ELs, 75% of the participants either disagreed or strongly disagreed with the statement that ELs taking the ACT receive comparable benefits to participation in the FSA.

Panelists were extremely divided on whether it would be acceptable to have the ACT was administered in lieu of Florida's grade 10 statewide, standardized English Language Arts (ELA) assessment and the Algebra I end-of-course (EOC) assessment. Half either agreed or strongly agreed with the statement that it would be acceptable, but a quarter disagreed or strongly disagreed. Additionally, almost one third of the panel members were neutral (13%) or did not answer this question (19%).

Table 3-11 shows the summary of panelists' conclusions for the SAT. As for ACT, disaggregated results are presented in **Appendix 3.E**.

Table 3-11. Participant Questionnaire Responses for SAT *(N=16)*

Question	No Response (%)	Strongly Disagree & Disagree (%)	Neutral (%)	Agree & Strongly Agree (%)
1. The SAT provides testing accommodations that provide **students with disabilities** with the opportunity to participate in the SAT.		6	13	82
2. The SAT provides testing accommodations that provide **ELs** with the opportunity to participate in the SAT.		44	25	31
3. The SAT uses an appropriate approval process for accommodations for **students with disabilities**.	6	13	6	82
4. The SAT uses an appropriate approval process for accommodations for **ELs**.		19	38	37
5. **Students with disabilities** taking the SAT receive comparable benefits to participation in the Florida Standards Assessment.		25	19	56
6. ELs taking the SAT receive comparable benefits to participation in the Florida Standards Assessment.		56	19	25
7. It would be acceptable if the SAT was administered in lieu of Florida's grade	19	31	13	37

Question	No Response (%)	Strongly Disagree & Disagree (%)	Neutral (%)	Agree & Strongly Agree (%)
10 statewide, standardized English Language Arts (ELA) assessment and the Algebra 1 end-of-course (EOC) assessment.				

Note: % may not total exactly 100% due to rounding.

As shown in Table 3-11, more than 80% of educator panel members agreed or strongly agreed that the SAT provides testing accommodations that give students with disabilities the opportunity to participate in the SAT, whereas more than two-thirds either disagreed with or were neutral to the statement that the SAT provides testing accommodations that provide ELs with the opportunity to participate in the SAT. More than 80% of the educators either agreed or strongly agreed that SAT uses an appropriate approval process for accommodations for students with disabilities. In contrast, for ELs, more than 50% of educators were neutral or disagreed with the statement that SAT uses an appropriate approval process for accommodations requests.

More than 80% of the educator panel members agreed or strongly agreed that students with disabilities taking the SAT receive comparable benefits to participation in the FSA; for ELs, 75% of the educator panel members disagreed or were neutral with the statement that ELs taking the SAT receive comparable benefits to participation in the FSA.

Panelists were divided on whether it would be acceptable for the SAT to be administered in lieu of Florida's grade 10 statewide, standardized English Language Arts (ELA) assessment and the Algebra I end-of-course (EOC) assessment. More than one-third either agreed or strongly agreed with the statement that it would be acceptable; still, more than 40% were either neutral or disagreed. Almost 20% of the panelists did not respond to this question.

Panelists also completed a meeting evaluation form. As shown in the summary of the evaluation results that is in **Appendix 3.F**, most panel members agreed or strongly agreed that appropriate processes were used, that they had enough information to participate in the discussions, that they had adequate time, that their questions were answered, that they felt free to provide their input and recommendations, and that they were comfortable with their conclusions and recommendations.

Limitations of Studies

The accommodations studies had several limitations. One limitation is that the educators who participated in the matheatics and ELA panels probably were more interested in accommodations and accommodations decision making than typical educators in Florida. This suggests that the panel may have had different findings if "typical" teachers participated. For example, the educators who participated in the panels believed that it might be time consuming to complete accommodations request and documentation process – though they generally seemed confident that they could complete the application process in a way that their students'

accommodations would be approved. Less experienced teachers may struggle to complete the application request appropriately and end up with many more students with non-approved accommodations.

Another limitation is that the panels may have lacked educators with a deep knowledge of students in some disability categories. Each panel contained a: a) special educator (Exceptional Student Education – ESE); English learner educator (English for Speakers of Other Languages – ESOL); blind/low-vision educator; deaf/hard of hearing educator; and content educator. However, the selection process did not have selection criteria that specifically addressed disabilities other than vision impairment and deaf/hard of hearing. Therefore, the panels may have been unable to fully evaluate the appropriateness of the ACT and SAT request process for students in some disability categories that may be more likely to have their accommodations requests denied (e.g., emotional disabilities, learning disabilities, other health impaired).

A third limitation is that the analysis of the number of accommodations offered on the ACT and SAT is based on self-reported information provided by ACT and College Board specifically in response to questions generated during the educator panel meetings about the number of accommodations (see **Appendix 3.D**). Prior to the accommodations studies meetings, ACT provided NCEO with a list of accommodations, and College Board provided examples (but not a comprehensive list of accommodations). Panel members wanted a better understanding of which accommodations were available (particularly for the SAT), so they asked both ACT and College Board to indicate on a table provided by NCEO which accommodations were available on their respective assessments. The findings would be stronger for the number of accommodations if the number was derived from publicly available sources. For the FSA the numbers are based on publicly available information in the *2017-18 FSA Accommodations Guide* (Florida Department of Education, 2017).

A fourth limitation is that direct comparisons of the number of accommodations for ELs on ACT and SAT **cannot be made** because ACT and College Board reported information on accommodations for ELs in different ways. SAT provided information for students who were EL only (i.e., ELs who did not have a disability), whereas the ACT list for ELs included accommodations typically used only by ELs with disabilities (e.g., food/medication for individuals with medical need, magnification, sign language directions only – American sign language).

Conclusions

This report presents the findings of the research and analyses NCEO conducted to evaluate the degree to which the ACT or the SAT provides accommodations that permit students with disabilities and ELs the opportunity to participate in the assessment and receive comparable benefits. To do this, NCEO conducted two Accommodations Studies with panels of Florida educators; the panels were designed to obtain input from Florida educators on accommodations for the ACT, SAT, and FSA. As previously described, the panel members identified four considerations when evaluating whether the ACT and SAT offered comparable benefits: (1) uses (e.g., accountability, graduation, programmatic changes); (2) accommodations allowed; (3) process (e.g., requests, college reportability); and (4) cultural sensitivity. The findings of the

NCEO and the panel members for each consideration are in this section. This is followed by the overall conclusions.

Uses (e.g., School Accountability, Graduation, Programmatic Changes)
NCEO and panel members found that ACT and SAT offer comparable benefits to the FSA for the identified purposes of school accountability and graduation. The results of the FSA are used for accountability and graduation purposes. Similarly, Florida may allow the ACT and SAT results to be used for accountability and graduation purposes.

Educator panel members indicated that FSA scores are sometimes used to make programmatic changes (e.g., instructional decision making). Timely receipt of results is important when assessment data are used for programmatic changes. Based on the information provided by ACT and SAT, the panel members wondered whether there would be a quicker turn around for the ACT and SAT compared to the FSA turn-around time. NCEO concluded that there was no information available to address the use for programmatic changes.

Accommodations Allowed
There is a slightly higher number of accommodations for students with disabilities for the ACT and SAT compared to the FSA. For ELs, there were differences in the number of accommodations across the three tests, with the ACT providing more, then FSA, followed by SAT providing a significantly smaller number. One reason SAT allowed the fewest accommodations for ELs was because the College Board assumed that the student was EL only (i.e., an EL without a disability), while ACT included accommodations for ELs that typically would be used only by ELs with a disability. NCEO and panel members thought that the differences in numbers of accommodations for students with disabilities and ELs may indicate a fairness issue, an issue that goes across all three assessments.

Process (e.g., Requests, College Reportability)
For the process aspect of comparability, NCEO and panel members found there is not complete comparability between the FSA, and ACT and SAT. The accommodations request process is a little more cumbersome for the ACT and SAT than for the FSA, given that accommodations for the FSA have already been determined through the IEP process. However, for school-day administrations, both ACT and College Board will require each school that is established as a test site to have a Test Accommodations Coordinator (TAC), who will facilitate the accommodations request and appeals process for the students in the school. Many schools already experience the accommodations request process for students with disabilities and ELs when they request accommodations for the national test day administration. NCEO believes the request process becomes an issue when the individuals assigned to process requests for the school do not have the experience or time to provide convincing documentation of the need for accommodations. This may occur in resource-poor schools.

NCEO and most members of the educator panels believe that both ACT and the College Board provided testing accommodations that would give students with disabilities the opportunity to participate in these assessments. However, they also found that the use of some accommodations could result in students with disabilities and ELs receiving scores that are non-college reportable and cannot be used for college admissions purposes.

The accommodations decision-making process used by ACT and College Board (for the SAT) are not fully transparent. School staff can request accommodations for students with accessibility needs and complete required documentation, yet the accommodations are not always approved as college-reportable accommodations. Neither ACT nor College Board makes the criteria they use to approve or reject accommodations requests publicly available. This is different from the FSA process, where 504/IEP/EL teams or other school personnel who know the student are the final decision maker for all accommodations except those that are special requests, which is a small minority of the accommodations. The use of some accommodations on the ACT and SAT always result in non-college reportable scores for students with disabilities. For example, the use of the American Sign Language (ASL) accommodation to deliver test content will result in non-college reportable scores.

Cultural Sensitivity

Although cultural sensitivity was identified as an important aspect of comparability, there was not a systematic discussion of this aspect. Some educators noted concerns about the cultural sensitivity of the FSA, but there was no way to judge any differences across tests. Given typical cultural sensitivity approaches of test developers, NCEO concluded that the three tests are likely comparable, although FSA may be more Florida-centric.

Overall Conclusion

Based on its knowledge and 27 years of experience with accommodations policies, and the input from Florida educators, NCEO concludes that in many ways, in terms of the provision of accommodations, the ACT and SAT could provide comparable benefit to the FSA for purposes of school accountability and graduation, although this was less evident for ELs for the SAT (and in general across the assessments). In general, both ACT and College Board indicated that they would provide greater numbers of the accommodations in the NCEO list of accommodations than were provided for the FSA. Whether these differences were appropriate for the Florida standards was not addressed in these studies.

Comparability in the process for accommodations requests was less clear and often relevant more to the use of the tests for college entrance; comparability to the FSA cannot be judged here because the FSA does not provide a score that can be used for college entrance. Still, if a district is basing a decision to use one of these tests in lieu of the Florida assessments on the possibility of having college entrance scores for all of its students, this goal is unlikely to be realized for some students with disabilities and ELs. The lack of transparency in the decision-making process about which specific accommodations would result in a college reportable score for which specific students is likely to result in non-comparability for some student groups compared to other student groups, which could be a concern when making the decision about whether to allow Florida districts to use either the ACT or the SAT in lieu of the Florida assessments.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Bureau of Contracts, Grants, and Procurement Management Services (2017). *Request for proposals: Feasibility of the use of the ACT and SAT in lieu of statewide assessments (RFP2018-48)*. Orlando FL: Florida Department of Education.

Florida Department of Education. *2017-18 FSA Accommodations Guide*. Retrieved from: http://fsassessments.org/wp-content/uploads/2017/08/2017-2018_FL_FSA_Accomm_Guide_082917_Final.pdf

Lazarus, S.S. & Thurlow, M.L. (2016). 2015-16 high school assessment accommodations policies: An analysis of ACT, SAT, PARCC, and Smarter Balanced (NCEO Report 403). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Justice. (2015). *Testing accommodations.* Available at: http://www.ada.gov/regs2014/testing_accommodations.html

Section 4
Accountability Studies (Criterion 5)

Executive Summary

Moving the discussion to the effect of the using the alternate tests (ACT/SAT) in place of the Florida tests for high school accountability focuses the analysis on aggregate school results rather than individual results. Overall, the question to be answered is about fairness: if Florida evaluates all schools using the same types of indicators, but those indicators are derived from scores on different tests, will the cross-school comparisons be fair?

Florida accountability systems use the state testing results in three ways: (1) calculating performance of students across the five achievement levels; (2) calculating the number of students making gains in test scores from one year to the next; (3) calculating growth using a value-added model (VAM). Using the sample of students with two years of data from the FSAs and an ACT or SAT score, simulated schools were created to examine the effects of calculating school-level indicators using the different tests.

Overall, differences are shown across all three indicators. The results show that the numbers going into the accountability determination would differ for many schools by the test selected. Richer calculations can be done for ELA as state test data is available for grades 8, 9, and 10. For mathematics, the time at which the Algebra 1 EOC test is taken varies by student and for many, there is no prior year mathematics score upon which to base a growth calculation.

There are two important findings to consider from this accountability study: one data-based and one more theoretical. First, the differences shown for ELA vary by type of school. Larger schools with a greater number of lower performing students are advantaged by using the alternate tests (ACT/SAT). This finding has implications for policy, as districts could use these results to select a test, rather than making a more holistic determination about its students and what test best fits the population.

Second, there will often be very different students being compared in the growth models. For example, in mathematics, the learning gain using the FSA will be calculated based on Grade 8 math and Grade 9 Algebra 1. However, using the alternate test, a similar high school would be evaluated based on the learning gain between Algebra 1 EOC in Grade 10 and the ACT or SAT in Grade 11. Likewise, for the value-added model, only schools using the FSAs will have a VAM score for ELA and only some of those for mathematics. For districts that elect to use the ACT or SAT in lieu of the FSA Grade 10 ELA and Algebra 1 EOC tests, two years of prior data will not exist for students taking the ACT or SAT in grade 11.

Both of these findings indicate that the answer to the question on fairness is: "no – it is not fair to compare schools that use the state tests in their accountability system to those that use the alternate tests."

Introduction

Ultimately, the decision to allow districts to use the ACT or SAT in lieu of the Florida State Assessments will impact the Florida school accountability system. In Section 2 (Criterion 3), the comparability study showed that the differences in content, particularly in mathematics, resulted in scores that gave different results in terms of achievement levels depending on the test taken. This study (Criterion 5) provides information on how those differences manifest themselves in accountability calculations. In the comparability study, comparisons of classification consistency demonstrated that only about 50% of students would be placed in the same achievement category for ELA if they took the ACT or SAT in lieu of the FSA Grade 10 ELA test. The numbers were even lower for mathematics. However, for accountability purposes, these results are aggregated across schools or teachers.

If all students in Florida had both an ACT/SAT and an FSA score, it would be possible to rerun the accountability calculations using the ACT/SAT data to see if the school received the same grade as it did with the FSA score. However, only about half of the students in Florida took the ACT or SAT and they are scattered across the state. Therefore, students who did have both scores were reorganized into mock schools and the indicators that contribute to the accountability model were calculated using first one test and then the other. The analyses in this chapter simulate the various indicators that are used in accountability measures to show the impact on using different tests to make decisions. Overall, the question to be answered is about fairness: if Florida evaluates all schools using the same types of indicators, but those indicators are derived from scores on different tests, will the cross-school comparisons be comparable or fair?

This section of the report is organized as follows:

- Describe the various accountability systems in Florida,
- Provide the results for the simulations, and
- Discuss the implications of the proposed policy to allow districts to administer the ACT or SAT in Grade 11 in lieu of the Algebra I EOC and ELA 10 tests.

A series of accountability analyses are described in order to demonstrate the effect of using the concordance tables created in Section 2 on student score reports and Florida school report cards. For the schools, both the achievement measures and the learning gain scores are calculated for similar schools under the three testing conditions: using the ACT, the SAT, or the FSA. Decision consistency is compared at the school level. In addition, because a value-added model of growth is used in Florida for evaluation of teacher preparation programs, the impact on the different assessments is described for this measure as well.

Florida Accountability Systems

Florida has several accountability systems with different purposes. One of the better-known components is the school grades given to every school every year. The intent of school grades is to provide an easily understandable metric to describe the performance of a school that can be used by parents and the general public to analyze how well each school is serving its students.

The school grades calculation was revised substantially for the 2014-15 school year to implement statutory changes made by the 2014 Legislature and incorporate the new Florida Standards Assessments (FSA). The simulations in this chapter are based on those revisions.

The school grading system focuses the school grading formula on student success measures, which include indicators for:

- Achievement
- Learning gains
- Graduation
- Acceleration success
- Maintaining a focus on students who need the most support

In terms of comparing the Florida assessments to the ACT and SAT, only the achievement and learning gains could differ. The learning gains are calculated two ways: for the school as a whole and then for the 25% of students who scored the lowest on the achievement tests.

At the high school level, the two tests used for achievement scores are FSA Algebra 1 EOC and FSA ELA Grade 10. These are the two tests that would be replaced by the ACT or SAT for districts exercising that option. Currently, students must score at Level 3 or higher on these two FSA tests in order to graduate. The score on the FSA Algebra 1 EOC test also counts for at least 30% of the student grade if the student is enrolled in the course when taking the EOC test.

In addition, there is a separate accountability system being developed for schools run by the Department of Juvenile Justice (DJJ), but it has not yet been adopted by the State Board of Education. Currently, there are only two proposed measures that would use FSA results, and both are based on the learning gains calculations, with one important modification. In the DJJ accountability system, students will need to have been in the school at the time of testing, and for at least 40 days prior. Although this is an important distinction for school accountability, it has little impact on simulations of the effect of whether the students take the Florida assessments or the ACT or SAT.

The final component of the Florida accountability system involves the use of Value-Added Models (VAM). Value-added analysis is a statistical method that estimates the effectiveness of a teacher by seeking to isolate the contribution of the teacher to student learning. Conceptually, a portion of the difference between a student's actual score on an assessment and the score the student was expected to achieve is the estimated "value" that the teacher added during the year. These models became optional for districts to use under HB 7069 as a measure of student growth.

Additionally, the annual evaluation of teacher preparation programs requires the use of VAM data. The "Annual Program Performance Report" or "APPR" is the yearly public report card issued by the FDOE for a state-approved teacher preparation program that includes results of outcome-based performance metrics specified in Sections 1004.04(4)(a), 1004.85(4)(b) and 1012.56(8)(c)2, F.S., and SBE Rule 6A-5.066, Florida Administrative Code (FAC). Points awarded on the APPR are based on progress on six performance measures:

1. Placement of program completers in instructional positions in Florida public schools and private schools, if available;
2. Retention of completers employed in instructional positions in Florida public schools;
3. Performance of students in preK-12 who are assigned to in-field program completers on statewide assessments using the results of the student learning growth formula adopted under section 1012.34, F.S.;
4. Performance of students in preK-12 who are assigned to in-field program completers aggregated by selected student subgroups;
5. Results of program completers' annual evaluations; and
6. Production of program completers in Critical Teacher Shortage Areas, as defined in section 1012.07, F.S.

The third and fourth of these measures are calculated through VAMs.

Thus, to compare the impact of using ACT or SAT in lieu of the FSAs, it will be important to calculate the following indicators:

1. Achievement
2. Learning gains
3. VAM growth measure

Depending on the impact of the different assessments on those indicators, it will be possible to predict if a school or program will obtain a higher or lower score on the overall accountability system based on the assessment chosen.

Simulations
In order to determine the effects of using different assessments on the school accountability measures, various school types were simulated. Because only about 50% of students took the ACT or SAT, it was not possible to use real schools. Instead, the students who had both FSA and either ACT or SAT test scores were grouped into mock schools to examine the effect of using different tests on school accountability. The analyses varied both the achievement level and size of the school. Then, the three indicators were calculated: Achievement, learning gains, and VAM growth.

These simulated schools allowed for analyses to answer these questions:

- Are there differences in indicators that would contribute to the school grades based on the overall achievement level of the school?
- Are there differences in indicators that would contribute to the school grades based on the size of the school?

These are important questions and were a frequent criticism of the previous accountability system under the *No Child Left Behind* Act of 2001 where larger, more diverse schools were disproportionately penalized.

In order to best simulate performance as it would be under the proposed policy, the sample was limited to students who took the ACT or SAT in eleventh grade. Of those, the number of students who also had an FSA test score in ELA or Algebra 1 *and* a prior year score are shown in Table 4-1.

Table 4-1. Number of Students who Could be Included in State A–F Accountability System

Test	ACT	SAT
ELA	59,935	12,010
Mathematics	1,056	136

ELA is the easiest to simulate as the students would have taken FSA ELA in Grade 10 and the ACT or SAT in Grade 11. Because there is also an FSA ELA test in grade 9, the vast majority of these students will have a prior year score. As with the comparability analyses, the population of students who took both the ACT or SAT and the Algebra 1 EOC test was small, as there were only three years of data for ACT and two years of data for SAT. Further reducing the sample to those who also had a prior year's score made the sample quite small.

Next, simulated schools were created to match school characteristics of typical "A" schools, "B" schools, "C" schools, "D" schools, and "F" schools, as assigned by the Florida School Grade policy and applied in 2017. Using FDOE data, average percentages of students in each level were calculated using the average of schools with letter grade 'F', 'D', etc. These averages were used in the simulations. That is, a school simulated to have an A is a school that was sampled to have the same percentage of students in each level as the average A school in the state. All simulations were conducted using prior year achievement so that learning gains were computable.

The other variable that was manipulated was school size, which was defined as the number of students per grade. Because school size is not a variable that is distributed evenly like a bell curve, samples were drawn using various percentiles at irregular intervals: 10%, 25%, 40%, 60%, 75% and 90%. These corresponded to school sizes shown in Table 4-2.

Table 4-2. School Sizes Corresponding to Key Percentiles

Percentile	10%	25%	40%	60%	75%	90%
Class Size	39	117	262	411	508	614

In addition, there were 13 schools with more than 800 test takers, so the decision was made to also simulate a school with a school size of 800.

For ELA, it was possible to generate a matrix of 7 (size) x 5 (accountability grade level) or 35 schools. Due to rounding, actual counts of students fluctuated by one or two students depending on the condition. Then, students were chosen at random to populate each school. Twenty versions of each simulated school were created to determine how much the random selection might influence the results. The results reported include the average of those 20 versions of the school.

For mathematics, the number of schools that could be generated was much smaller and focused more on size than achievement level. Specific conditions will be described with each table. For each type of school, analyses were run that examined how the school would fare using the FSAs, the ACT, or the SAT. The next three parts in this section show the results for

A) The achievement indicators,
B) The learning gains indicator (both overall and for the lowest scoring 25%), and
C) A discussion of the data going into the value-added model.

A. Achievement
It is important to determine how consistent the ratings would be for the percentage of students scoring in each achievement category.

a.1. ELA
First, school results were analyzed for the ACT. The analyses answered the question of whether there would be differences in the achievement indicator if the ACT was used instead of the FSA ELA Grade 10 test. The analyses examined the question for schools of different achievement levels and size.

The factor that seems to most determine whether there will be significant differences in the achievement indicator is the size of the school. None of the simulated schools with school sizes of 39 showed any significant differences in the percentages of students scoring at each Achievement Level as calculated with either the Florida or ACT score. Likewise, no B, C, or D school with school sizes of 117 or C schools with school sizes of 262 showed any differences by test. However, for the larger schools, there were some differences. Schools with a typical performance distribution of an A school tended to have their scores move towards the tails of the distribution when using ACT scores. That is, more students were likely to score at Levels 1 and 5 and fewer at Level 3.

Table 4-3 shows an example of an A school with 411 students per grade. The mean represents the average number of students across the twenty simulated schools whose scores fell within the achievement level range for the given test (i.e., the FSA or the ACT). The standard error of measurement (SEM) is shown to demonstrate the degree in variance in the 20 schools simulated. To determine if the difference between the two sets of scores is significant, add the two SEMs within a particular row and then compare that sum to the difference in the two means in that same row. If the differences between the means is greater than the sum of the two SEMs, then it is a statistically significant difference. For example, for Achievement Level 1,

ACT mean – FSA mean = 49 – 42 = 7
FSA SEM + ACT SEM = 1.42 + 1.22 = 2.64

The mean difference of 7 is larger than the sum of the SEMs (2.64), so that difference is statistically significant.

Table 4-3. Distribution of Scores Across Achievement Levels for an "A" School with 411 Students per Grade

Achievement Level	FSA Mean	FSA SEM	ACT Mean	ACT SEM
1	42	1.42	49	1.22
2	100	1.64	96	1.00
3	84	1.83	72	1.62
4	119	1.43	120	1.49
5	64	1.21	71	1.22

Although the differences are small, the ACT scores place more students in Levels 1 and 5 and fewer in Levels 2 and 3. Overall, however, there are only three fewer students scoring at the graduation requirement of Level 3 or higher using the ACT.

A similar pattern of very small differences is shown for a C school of 411 students, as shown in Table 4-4.

Table 4-4. Distribution of Scores Across Achievement Levels for a "C" School with 411 Students per Grade

Achievement Level	FSA Mean	FSA SEM	ACT Mean	ACT SEM
1	128	0.96	131	1.78
2	140	2.05	136	2.65
3	65	2.12	58	1.56
4	55	1.12	58	0.93
5	22	0.82	25	0.83

Contrast those patterns with a B school with only 117 students. Table 4-5 shows that there are virtually no significant differences between the FSA and ACT at any achievement level. There might be one more student scoring at Level 5 using the ACT and one fewer at Level 3 depending on the school, but overall, the differences are not statistically significant.

Table 4-5. Distribution of Scores Across Achievement Levels for a "B" School with 117 Students per Grade

Achievement Level	FSA mean	FSA SEM	ACT mean	ACT SEM
1	26	0.56	27	0.77
2	36	0.72	36	0.78
3	21	0.75	19	0.95
4	22	0.85	22	0.69
5	9	0.62	11	0.46

Finally, F schools had an interesting pattern. Again, no differences were seen between FSA and ACT score distributions for the smallest schools. However, schools with 117 or more students per grade showed the pattern of fewer students scoring at Level 1 using the ACT and more scoring at Levels 2, 3, and 4. This implies that F schools are advantaged by using the ACT. For the majority of the other scenarios, the larger proportion in the highest level is negated by the larger proportion in the smaller level, or vice versa.

Table 4-6. Distribution of Scores Across Achievement Levels for an "F" School with 614 Students per Grade

Achievement Level	FSA Mean	FSA SEM	ACT Mean	ACT SEM
1	403	2.21	368	2.08
2	189	2.39	213	2.29
3	18	0.69	25	1.18
4	2	0.34	6	0.57
5	0	0	0	0.15

Some of the same patterns were found with SAT, but they were not as consistent. For A schools, there were no significant differences in the percentage of students scoring at Achievement Level 5, regardless of the test. However, patterns across the other four achievement levels were less clear. For instance, A schools with a class size of 411 showed a pattern of the SAT placing more students in Level 1 and the FSA placing more students in Level 4.

Differences in the other levels were not significant (see Table 4-7). All other school sizes showed more students in Level 3 with SAT than with FSA scores. School sizes of 508 and higher showed the pattern of more students placed in Levels 1 and 3 by the SAT and in Levels 2 and 4 by the FSA. For school sizes 117 and 262, more students were placed in Levels 1 and 3 by SAT and in Level 2 by FSA. The implication is that schools will perform better if they choose the FSA Grade 11 ELA over the SAT.

Table 4-7. Distribution of Scores Across Achievement Levels for an "A" School with 411 Students per Grade

Achievement Level	FSA Mean	FSA SEM	SAT Mean	SAT SEM
1	40	1.11	44	0.96
2	93	1.63	93	1.66
3	82	1.97	83	1.43
4	121	1.88	113	1.42
5	73	1.51	76	1.33

Table 4-8 examines whether the same pattern holds for C schools as for A schools. It was similar in that the SAT placed more students in Level 3 than did the FSA. However, the FSA placed more students in Level 2 than the SAT. So, in this case, a C school would be better off choosing

the SAT over the FSA Grade 10 ELA test. There were no significant differences across any Achievement Level for smaller C schools. However, starting at the school size of 262, there is a slight advantage to using SAT over the FSA to calculate the achievement indicator.

Table 4-8. Distribution of Scores across Achievement Levels for a "C" School with 411 Students per Grade

Achievement Level	FSA Mean	FSA SEM	SAT Mean	SAT SEM
1	121	1.30	122	1.09
2	139	1.99	133	1.60
3	64	1.24	69	1.83
4	59	1.29	59	1.33
5	26	0.81	26	0.91

Finally, examining F schools shows an even stronger advantage to using the SAT scores. Table 4-9 shows the data from a simulated large school, with a class size of 614. The FSA ELA grade 10 test placed more students in Level 1 than did the linked SAT scores. Conversely, the SAT scores placed more students at Levels 2, 3, and 4 than the FSA scores. Smaller schools, those with school sizes of 39, showed that the FSA ELA grade 10 test placed more students in Level 1 than did the linked SAT scores. The SAT scores placed more students at Levels 2 and 3 than the FSA scores. In fact, all school sizes showed that the FSA placed more students in Level 1 than did the SAT. Only the level that favored the SAT changed with the size of the school.

Table 4-9. Distribution of Scores Across Achievement Levels for an "F" School with 614 Students per Grade

Achievement Level	FSA Mean	FSA SEM	SAT Mean	SAT SEM
1	385	2.10	358	1.90
2	203	2.23	210	1.95
3	22	0.72	32	0.88
4	3	0.48	12	0.53
5	0	0.00	1	0.14

In conclusion, these simulations show that A schools will perform better on the FSA ELA Grade 10 test, but schools with Achievement Levels similar to C, D, and F schools would do better with the SAT. There were few significant differences for B schools.

a.2. Mathematics

To better understand the population of students who had an Algebra 1 EOC score, an ACT or SAT score, and a prior year score, the first analysis looked at the distribution of students who scored across the five achievement categories using both their FSA results and their ACT or SAT score linked back to the Florida scale.

Table 4-10. Distribution of Students Across Achievement Categories by Subject and Test

	Students who took both FSA and ACT		Students who took both FSA and SAT	
Math	**Algebra 1 EOC**	**ACT**	**Algebra 1 EOC**	**SAT**
Level 1	28.2%	39.7%	12.5%	15.4%
Level 2	12.7%	20.3%	6.6%	13.2%
Level 3	35.3%	26.5%	21.3%	36.0%
Level 4	12.2%	5.0%	18.4%	18.4%
Level 5	11.6%	8.6%	41.2%	16.9%
Levels 3 + 4 + 5	*59.1%*	*40.1%*	*80.9%*	*71.3%*

As shown in Table 4-10, more higher-achieving students took the SAT than the ACT. But, in both cases, a greater proportion of students would score below achievement level 3 if using the linked ACT or SAT score than if taking the FSA Algebra 1 EOC test. The skew is more pronounced for ACT at the lower levels. This trend could be due to the difference in content, with instruction focused on FSA Algebra 1 while the ACT and SAT assess much more than just Algebra 1 content.

Although there are fewer scores that can be used to analyze mathematics trends, the trends with the ACT are clear: the ACT places more students in Achievement Levels 1 and 2 and the FSA Algebra 1 EOC places more students in Levels 3 and 4. This trend was true regardless of school size.

A similar trend occurred for SAT. The SAT placed more students in Levels 1, 2, and 3 and the FSA Algebra 1 EOC placed more students in Level 5. These results should be interpreted with caution, however, as it represents a very small sample of students.

Summary of Achievement Simulations

With ELA, there was a sufficient sample to create schools of differing sizes and achievement spreads. There were no significant differences with using the FSA scores compared to ACT or SAT for small schools. As the size of the school grew, the achievement level of the students mattered more. The simulations show that A schools will perform better on the FSA ELA Grade 10 test, but schools with achievement levels similar to C, D, and F schools would do better with the SAT. There were few significant differences for B schools.

For mathematics, there were too few students to simulate results for the 35 school types. Looking at the group of students overall, and grouping them in different size schools from 25 to 300, the results consistently show that the Algebra 1 EOC results placed more students in higher achievement levels and the linked ACT or SAT scores placed more students in the lower levels.

Overall, a small school or a high performing one would look better in the A–F grading system if it used the FSAs, but a larger, lower-performing school would do better with ACT or SAT in ELA. These differences indicate that the answer to the question about fairness is no: schools that

use in state tests in their accountability system should not be compared to those that use the alternate tests.

B. Learning Gains

Learning gains were calculated using the Florida rules about increasing achievement levels from one grade to the next or increasing scale scores for the higher achievement levels.[5] It is important to remember that for ELA, the learning gain is calculated from Grade 9 to 10 for those taking the FSAs and from Grade 10 to 11 for those taking the alternate test. For each school, the number of students showing a gain was calculated and then converted to a percentage to allow for comparisons across different school sizes.

The first step was to calculate learning gains for the full sample and then to compare that to learning gains of the lowest achieving 25%. Because the sample sizes are small for the bottom 25%, this analysis uses the full sample. Table 4-11 shows fairly substantial differences between learning gains calculated using the FSA and learning gains calculated for the alternate test, particularly in the lowest achieving quartile.

Table 4-11. Percentage of Students Showing a Learning Gain, by Test, Subject, and Achievement

	N (denominator)	FSA	ACT/SAT
Full sample			
ACT ELA	59,935	46.3%	48.5%
SAT ELA	12,010	53.8%	53.7%
ACT MATH	1,056	47.2%	31.3%
SAT MATH	136	59.6%	33.8%
Bottom 25%			
ACT ELA	14,984	15.1%	35.9%
SAT ELA	3,002	17.6%	34.3%
ACT MATH	264	11.4%	29.2%
SAT MATH	34	14.7%	32.4%

ACT or SAT learning gains higher than FSA gains within the same school type are highlighted in green.
ACT or SAT learning gains lower than FSA gains within the same school type are highlighted in yellow.

For the full sample, slightly more students show a learning gain in ELA when measured by ACT. For math, more students show a learning gain with the FSA Algebra 1 EOC test than with either the ACT or SAT. However, for the bottom performing 25%, a significantly more students showed learning gains with the alternate test than on the FSA in both ELA and mathematics.

b.1. ELA

For ACT, school size is less of a determining factor, but the performance of the students impacts the magnitude of the differences between the gain scores calculated from the two Florida tests compared to the ACT and the Grade 10 test from the prior year. As shown in Table 4-12, for the

[5] See the Florida 2016-17 Guide to Calculating School and District Grades at http://www.fldoe.org/core/fileparse.php/18534/urlt/SchoolGradesCalcGuide17.pdf

A schools, there are no significant differences between the two learning gains calculations for schools of any size. For the C and F schools, however, all school sizes show a significant difference. B and D schools show more students with learning gains for ACT for all schools except for the smallest. In every case where a difference in the calculation occurs, ACT shows more students showing learning gains than the FSA.

Table 4-12 ELA Learning Gains Calculated for FSA and ACT for 35 School Types

	Percentage of Students with a Learning Gain									
Grade	**A**		**B**		**C**		**D**		**F**	
Size	**FL**	**ACT**	**FL**	**ACT**	**FL**	**ACT**	**FL**	**ACT**	**FL**	**ACT**
39	50.1	51.0	46.8	44.6	41.5	49.2	41.0	43.6	37.6	49.5
117	50.8	53.3	42.9	46.8	41.4	45.1	42.1	44.1	37.3	46.9
262	50.8	51.4	44.9	47.1	42.5	45.6	39.5	43.6	35.9	46.3
411	49.6	51.3	45.1	46.7	42.1	45.2	40.4	44.9	37.5	47.3
508	50.0	50.9	44.7	46.2	42.3	45.4	40.2	44.9	36.8	46.5
614	50.6	51.3	45.1	46.6	41.5	44.6	40.2	45.3	36.0	47.0
800	49.6	50.6	44.9	47.0	41.4	46.0	40.2	45.0	35.7	46.1

ACT learning gains higher than FSA gains within the same school type are highlighted in green.
ACT learning gains lower than FSA gains within the same school type are highlighted in yellow.

The pattern was a little different for SAT. The combination of large schools and lower performing schools were the ones that showed the largest gains for SAT over FSA, as demonstrated in Table 4-13. There were no significant differences in the number of students showing learning gains in A schools, with one exception – for the largest schools, with class sizes of 800, slightly more students showed a learning gain using the FSA than the SAT. All schools achieving at the D and F levels, regardless of size, had more students showing learning gains with the SAT than with the FSA.

Table 4-13. ELA Learning Gains Calculated for FSA and SAT for 35 School Types

	Percentage of Students with a Learning Gain									
Grade	**A**		**B**		**C**		**D**		**F**	
Size	**FL**	**SAT**	**FL**	**SAT**	**FL**	**SAT**	**FL**	**SAT**	**FL**	**SAT**
39	54.0	54.4	48.4	48.9	44.2	45.3	37.8	44.2	38.7	49.2
117	53.9	53.4	48.2	48.9	44.8	47.5	40.7	45.7	40.0	46.1
262	53.7	53.2	48.0	49.5	45.2	47.2	42.4	46.0	38.9	46.3
411	53.2	53.3	47.7	48.9	44.6	47.7	42.1	46.4	38.1	47.0
508	53.9	53.4	48.5	49.9	44.6	47.1	42.6	46.7	38.5	46.2
614	53.8	53.5	48.3	48.5	44.5	46.5	42.4	45.9	38.4	46.5
800	53.0	52.1	48.0	49.4	44.8	47.3	42.2	46.4	37.8	46.1

SAT learning gains higher than FSA gains within the same school type are highlighted in green.
SAT learning gains lower than FSA gains within the same school type are highlighted in yellow.

b.2. Mathematics

There were very few cases from which to calculate mathematics gains. For ACT, an effort was made to examine the effect by size, as the amount of variance in performance was too small to

try to simulate schools of different achievement levels. Regardless of school size, the findings were the same: more students showed learning gains using the FSA than the ACT. Because the FSA is the Algebra 1 EOC test, the primary learning gain calculation would be from the Algebra 1 test in the current year and the FSA Grade 8 math test in the prior year. This means that students in this calculation took Algebra 1 in grade 9. For a learning gain to be associated with the ACT, a mathematics course must have been taken in grade 10. For these simulations, the only mathematics data given was for Algebra 1. This means all of those learning gains represented students who took Algebra 1 in grade 10. Therefore, the students included in the ACT simulation were necessarily lower performing than those included in the FSA simulation. With that consideration, the fact that the learning gains indicator always favors the FSA in mathematics may reflect the achievement level of the students more than the low consistency in test scores.

Table 4-14. Mathematics Learning Gains Calculated for FSA and ACT, by School Size

Size of school	Algebra 1	ACT math
22	46.8%	32.1%
75	47.9%	35.1%
138	48.1%	32.7%
214	48.1%	33.1%
316	47.8%	33.4%

ACT learning gains lower than FSA gains within the same school type are highlighted in yellow.

For the SAT, gain scores only existed for 136 students. Again, more students showed learning gains using their EOC Algebra 1 test than the SAT, approximately 59.6% compared to 33.8%, respectively.

Summary of Learning Gains Simulations

The learning gains simulations resulted in significant differences between the FSA and the college admissions test results in the majority of the scenarios. For ELA, A schools did the same or slightly better with the FSA ELA Grade 10 test. There were no differences among the smallest schools until the achievement level of the students matched a school grade of C (for ACT), D (for SAT), or F (for both). Then, schools showed more learning gains with the SAT or ACT. Overall, the majority of school types had a greater number of students showing learning gains in ELA with the college admissions test than with the FSA Grade 10 test.

For mathematics, the numbers were smaller, but overall, more students showed learning gains using the alternative test than the FSA Algebra 1 EOC test. Although the data could not be cut as many ways, this find appeared true across multiple school sizes. Because of the limitations of the sample, these simulations were very different for the two tests. For Algebra 1, the learning gain was calculated between Grade 8 math and those who took the Algebra 1 EOC test in Grade 9. For the alternative tests, the learning gain was calculated between those who took Algebra 1 in Grade 10 and the ACT or SAT in Grade 11. Therefore, the ages and mathematics exposure differed between the two comparisons.

C. Value-Added Model

The VAM requires three years of data: the current year and two prior years. From the two prior years, a trajectory is established and then the current year score is compared to the predicted score based on that trajectory to determine whether the student is maintaining that trajectory, has increased more than expected, or has fallen behind.

c.1. ELA

For the ELA scores, the FSA provides an easy data source for the VAM. Students take the FSA ELA at grades 3–10. The VAM score for high school could be determined by using the Grade 7 and 8 scores to project to Grade 9 and use the Grade 8 and 9 scores to project to Grade 10. However, if either the ACT or SAT replaces the Grade 10 ELA test, then schools will not be able to calculate a value-added growth score as there will be no ELA test preceding the ACT or SAT.

c.2. Mathematics

The picture is more complicated for mathematics. For students who take Algebra 1 in Grade 9 or earlier, a VAM score can be calculated from the FSA math tests. For students who take Algebra 1 in Grades 10, 11, or 12, however, most will not likely have a prior year test score. For schools that choose to use the ACT or SAT in grade 11, a student must have taken a mathematics test in both Grades 9 and 10 to receive a VAM score. Only one scenario provides those data, and that is the one where a student takes Algebra 1 in Grade 9 and Geometry in Grade 10. However, the student also must take the EOC tests in both subjects, and the current law appears to indicate that would not be necessary, since the ACT or SAT would take the place of the Algebra 1 EOC.

Implications for Florida's Accountability System

There are multiple implications of the policy to allow schools to replace the FSA ELA Grade 10 test and the FSA Algebra 1 EOC test with the ACT or SAT. The policy could result in fewer indicators being available for use in various accountability systems, as well as districts possibly selecting the test that favors their schools.

For the first implication, it is important to note that selecting the ACT or SAT ensures that all students would have an achievement score for that indicator. Selecting the FSA would ensure that all students would have an ELA score, but higher achieving students would not have a mathematics score associated with their high school. Only schools who selected the FSA would have VAM scores.

Districts with large numbers of A schools would be better served staying with the FSA, as their students will do better on ELA, on average, than they would if using the ACT or SAT. For districts with large numbers of F schools, the reverse is true: they would do better to select the ACT or SAT for their ELA test. Mathematics is a much more complicated picture, as the highest achieving students do not have their mathematics score associated with high school under the current EOC program. In those cases, schools with large numbers of high-performing students would be able to include their scores if their district selected the ACT or SAT. For the students who tend to take Algebra 1 in high school, the EOC test is more likely to place them in a higher achievement level than either the ACT or SAT. This could be because the ACT and SAT assess a much broader math construct than simply Algebra 1 content.

However, all of these findings indicate that comparisons between districts who select the FSA and those that select either the ACT or SAT are fraught with problems. In the following figure, a comparison of potential outcomes is presented on what may be compared in schools using FSA versus an alternative test (ACT or SAT):

Figure 4-1. Comparison of FSA with ACT/SAT Outcomes

FSA	ACT/SAT
Students who took Algebra 1 in eighth grade and Geometry in ninth grade will have their achievement and learning scores included	All students will have their mathematics achievement at Grade 11 included.
Students who took Algebra 1 prior to eighth grade will not be included in high school rating for Grade 9-12 schools	All students will have their mathematics achievement at Grade 11 included
Students who took Algebra 1 in ninth grade will have achievement and learning gains included in the high school rating. Learning gains scores will also include students who took Geometry in high school immediately after taking Algebra 1.	All students will have a mathematics achievement score but only those who take a FSA Geometry EOC test in Grade 10 will have a learning gains score.
Students who took Algebra 1 in tenth, eleventh, or twelfth grade will most likely only have achievement included	All students will have a mathematics achievement score but only those who take a FSA Geometry EOC test in Grade 10 will have a learning gains score.
Students will take ELA in Grade 10 and have it included as both an achievement and learning gains indicator	All students will have their ELA achievement included for Grade 11 and none will have a learning gains score.

The use of the FSA Geometry EOC scores could allow for more gain scores to be calculated for students taking ACT or SAT, but that concordance table has not been developed.

So for those districts adopting the ACT or SAT, for ELA there will be an achievement score for all students but no learning gain; for mathematics, schools will have an achievement score for all students, but learning gains scores for only the lower achieving students. None of these schools will have a VAM score.

Conclusion

Allowing districts to choose whether to take the FSA, ACT, or SAT will likely result in a very uneven accountability system in Florida. Moreover, because of the differences in results based on school size and achievement levels, districts may be swayed to select the test they think will give them better ratings rather than the one that is truly best for their students. These two factors combined raise large questions regarding whether any of the Florida accountability systems would result in scores that could be fairly compared across districts that have selected different high school tests.

Section 5 – Peer Review (Criterion 6)

Executive Summary

The Every Student Succeeds Act (ESSA) provides an opportunity for states to permit school districts to use a locally selected, nationally recognized test in place of the statewide assessments used previously. In 2017, the Florida legislature adopted legislation (HB 7069) to do just that. The legislation directed the Florida Department of Education to investigate the feasibility of permitting Florida school districts to choose to use either the ACT or the SAT in lieu of the statewide, standardized ELA and Algebra 1 end-of-course assessments for high school students.

There are several ESSA-required criteria that any assessment proposed for use to meet ESSA assessment requirements must address. These criteria apply equally to the Florida Standards Assessments and any assessments selected by districts to replace the statewide assessments. These criteria include

- Sufficient coverage of the state's ELA/reading and mathematics content standards,
- Technical qualities (reliability and validity) of the tests,
- Accessibility features and accommodations offered to students with disabilities and English learners to assure that they can take the assessments in a fair and accessible manner,
- Protections for the privacy of students,
- Achievement standards, and
- Reporting results in an understandable and useful manner.

An important part of determining part of determining whether states address these requirements is a process called "peer review" of a state's assessment programs. In 2015, the U.S. Department of Education (USED) sent each state the non-regulatory guidance that advised the states about the criteria their testing programs had to meet. The guidance includes 30 Critical Elements that the state had to adequately address. Each state is to assemble evidence that it has met each Critical Element. Evidence might come from the state education agency, its vendors, assessment consortium (to the extent that a state participated in one), or local districts.

Once the state's response is prepared and submitted to USED, a panel of assessment experts read and critique the evidence for completeness, sufficiency, and adequacy. The peer notes are reviewed by USED staff. A decision letter is then prepared for the signature of the U.S. Secretary of Education (or her designee) and sent to the Commissioner of Education in the state.

To test the acceptability of Florida's plan to offer its schools the option of using the ACT or SAT in lieu of the FSAs, ASG conducted a mock peer review, using evidence provided by ACT, the College Board, and the Florida Department of Education, as well as from ASG's studies of alignment, comparability, and accommodations. Experienced peer reviewers examined the evidence and prepared written notes similar to an actual peer review. A summary of the peer review results is shown in Table 5-1.

Table 5-1. Number of Peer Review Critical Element Determinations by College Entrance Test

Peer Review Determinations	ACT	College Board
Met Mock Peer Review Requirements	23	20
May Not Meet Mock Peer Review Requirements	1	6
Did Not Meet Mock Peer Review Requirements	6	3
TOTAL	30	29*

* One Critical Element is not applicable to the College Board SAT.

The ACT was judged to not meet six of the 30 mock peer review Critical Elements, and may not meet one other. The SAT was judged to not meet one of the 29 mock peer review Critical Elements, and may not meet six others.

For these reasons, the conclusion of ASG and its partners is that providing Florida districts with the option to use the ACT or SAT in lieu of the FSAs will likely not meet USED peer review requirements. Details of the process and findings that support this conclusion are provided in the following parts of this section.

Introduction

This section of the report covers the sixth and final Criterion – the likelihood that Florida's plan to offer the ACT or SAT tests to Florida school districts in lieu of Florida's FSA grade 10 English Language Arts (ELA) and Algebra 1 end-of-course (EOC) assessments will meet the federal requirements for peer review of standards and assessments used to assure compliance with the Every Student Succeeds Act (ESSA). It describes the processes used to address this issue and summarizes the findings of the actual mock peer review carried out for the ACT and SAT as if those tests were used as a part of the Florida Statewide Assessment System.

Peer Review Requirements

One key addition to federal standards and assessment requirements included in ESSA is the flexibility for a state to permit a school district to administer, in lieu of the statewide high school assessment, a "locally selected, nationally recognized" high school academic assessment that has been approved for use by the state (including submission for the U.S. Department of Education's assessment peer review process). ESSA requires that a state must determine that its assessments meet specific criteria.

As noted in a recent CCSSO publication (CCSSO, 2017a), there are several requirements for peer review:

> "ESSA specifies that certain technical criteria must be satisfied to receive approval for use by the state. These requirements should be considered minimum standards, meaning the state may establish additional requirements. ESSA requires that the assessment chosen by the state
>
> - Is aligned to and addresses the breadth and depth of the state's content standards

- Is equivalent to the statewide assessments in its content coverage, difficulty, and quality
- Provides valid and reliable data on student achievement for all students and subgroups as compared to the statewide assessments[6]
- Meets the criteria for technical quality that all statewide assessments must meet (e.g., peer reviewed)
- Provides unbiased, rational, and consistent differentiation among schools within the state's accountability system.

> "Additionally, the ESSA statute and relevant regulations stipulate that any approved assessment would be subject to peer review. While Congress and President Trump have voided ESSA regulations related to accountability, they left in place ESSA regulations related to Title I assessments.
>
> "The requirement for peer review signals that a locally selected test will be reviewed against the same set of technical and administrative criteria used to evaluate the state test. For states with different assessment systems across LEAs, there will be a need to reexamine processes for reporting data. For example, if the state uses computer-based testing and an LEA uses paper-based testing, neither program in itself has a threat to comparable interpretation of results (i.e., 'mode effect'), but the two programs together need to demonstrate there is not a mode effect.
>
> "The state should establish additional criteria to ensure that data from locally selected assessments will support valid assessment interpretations and required accountability uses" (p. 3-4).

The U.S. Department of Education (USED) prepared non-regulatory guidance in 2015 (USED, 2015) to inform states about the requirements for standards and assessment peer review, including directions for states to submit the required evidence and a peer review template to be used to do so. Included in the non-regulatory guidance is this excerpt designed to inform states about the requirements for their peer review submissions.

> "A key purpose of Title I of the ESEA is to promote educational excellence and equity so that by the time they graduate high school all students master the knowledge and skills that they need in order to be successful in college and the workforce. States accomplish this, in part, by adopting challenging academic content standards that define what the State expects all students to know and be able to do. States must develop and administer assessments aligned to those standards, and adopt academic achievement standards aligned to the academic content standards to define levels of student achievement on the assessments…." (p. 1).

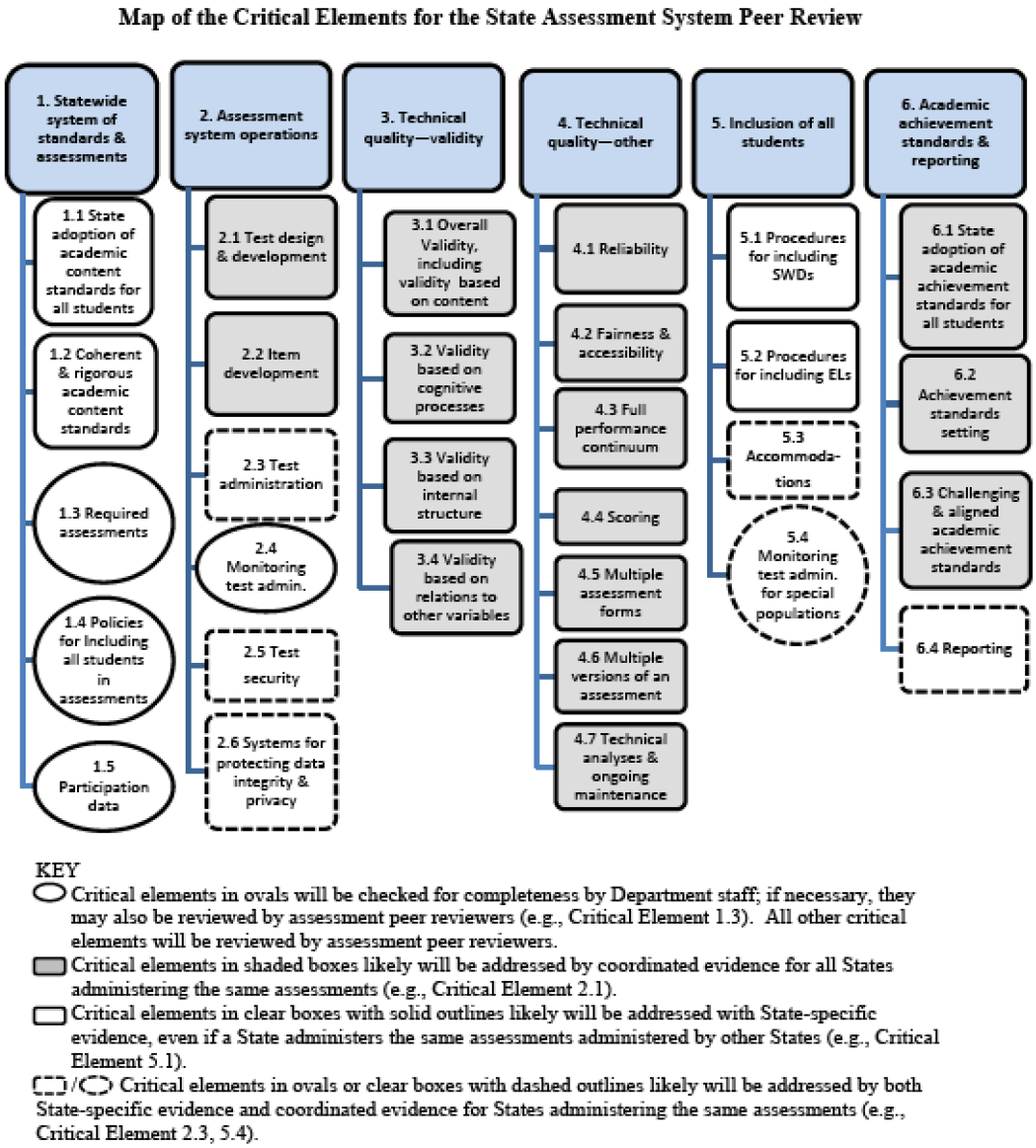
The non-regulatory guidance goes on define Critical Elements in the following manner.

[6] The December 2016 regulations indicate that comparability between the locally selected test and the state test is expected at each academic achievement level.

> **"Critical Elements.** The critical elements in Part II of this document represent the ESEA statutory and regulatory requirements that State assessment systems must meet. The six sections of critical elements that cover these requirements are: (1) Statewide System of Standards and Assessments, (2) Assessment System Operations, (3) Technical Quality – Validity, (4) Technical Quality – Other, (5) Inclusion of All Students, and (6) Academic Achievement Standards and Reporting. The map of critical elements included in Part II provides an overview of the six sections and the critical elements within each section" (USED, 2015, p. 4).

The map of the 30 ESSA peer review Critical Elements is shown in Figure 5-1 (USED, 2015, p. 19).

Figure 5-1. Map of the Elements for the ESSA State Assessment System Peer Review



The USED non-regulatory guidance also more fully describes a Critical Element:

> "*Critical Element.* The critical element is a statement of the relevant requirement, and a State must submit evidence to document that its assessment system meets the requirement. The set of evidence submitted for each critical element, collectively, should address the entirety of the critical element" (p. 12).

The statutory peer review requirements listed in the ESSA legislation can be found in **Appendix 5-A** (CCSSO, 2017b).

The potential to use consortium-based information for peer review is also spelled out in directions from the non-regulatory guidance on peer review (USED, 2015) provided to states. These directions are excerpted below.

> **"Coordination of Submissions for States that Administer the Same Assessments**
> In the case of multiple States administering the same assessment(s), the Department will hold one assessment peer review for those assessments in order to reduce the burden on States and to promote consistency in the assessment peer review. This includes groups of States that formed consortia for the purpose of developing assessments and States that administer the same commercially developed assessments (e.g., multiple States that are all administering the same commercially developed test as their high school assessment).
>
> "For evidence that is common across an assessment administered in multiple States, the submission of evidence should be coordinated, with one State submitting the evidence on behalf of all States administering the assessment. Each State also must submit State-specific evidence that is not common among States that administer the same assessment(s). As described below, in their State-specific submissions, individual States should cross-reference coordinated submissions. A State for which a coordinated submission of evidence is part of its evidence for assessment peer review is encouraged to submit its State-specific evidence at the same time as the coordinated submission" (USED, 2015, p. 11).

The directions that USED provided to states in its non-regulatory guidance regarding peer review serve to notify states about which organization is required to prepare what information for peer review, when a state chooses to use a consortium-developed or administered assessment or the same test as other states. Florida's decision to offer schools the option of using a college entrance test in lieu of the Florida Standards Assessments does not make the college entrance test officially "a consortium-based assessment." However, the fact that a number of states currently use either the ACT or the SAT for state-based ESSA high school accountability means that the evidence needed for Florida's peer review submission might be available from the evidence other states used in the past to meet the ESSA standards and assessment peer review requirements, or that the Florida Department of Education might be able to use the same evidence along with other states in a concurrent peer review submission. Note that this is commonly done by states that use a common assessment.

The USED non-regulatory guidance specifies two types of evidence for the Critical Elements that states might provide: 1) state-specific evidence and 2) evidence provided from assessments

"common among States that administer the same assessments" (USED, 2015). An additional type of evidence – a hybrid of these two information types – might also be necessary. USED provided a chart that classified each of the 30 Critical Elements in the peer review into one of these three categories. USED gave an example of hybrid evidence – the use of common training materials such as assessment administration manuals and training presentations used in a number of states administering the same assessment, but also individual state-created assessment administration training procedures and materials created and used as well.

Figure 5-2 shows the types of evidence needed for each Critical Element to satisfy the requirements of ESSA (USED, 2015, p. 12).

Figure 5-2. Type of Evidence for Each Critical Element

Evidence	**Critical Elements**
State-specific evidence	1.1, 1.2, 1.3, 1.4, 1.5, 2.4, 5.1, 5.2 and 6.1
Coordinated evidence for States administering the same assessments	2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 6.2 and 6.3
Hybrid evidence	2.3, 2.5, 2.6, 5.3, 5.4 and 6.4

These then provide the guidance for the peer review portion of the work on this project. The goal is to assemble the most complete evidence possible from all sources – from the Florida Department of Education, ACT, and the College Board, plus the studies of alignment, accommodations, and comparability from this project, and prepare this information for the equivalent of an external and independent peer review, with the goal of advising the Florida Department of Education about the likelihood that it would receive approval from USED for its plan to provide local districts in Florida with the flexibility to choose which high school test to use in the future.

A Word of Caution – Predicting a favorable outcome from peer review before the underlying assessment program is implemented by a state and its districts is a fraught with challenges for several reasons.

First, ESSA requires that all students take an assessment that measures the on-grade-level standards and do so with assessments aligned to the depth and breadth of its content standards in mathematics and reading/ELA. The standard for what constitutes "comparability" between tests when different measures are used by different students has not been established. This is essentially an issue of "how close is close enough?"

Second, since offering schools the option of using a nationally-available test in lieu of the state assessment is a new option offered to states in ESSA, the full nature of the evidence that will be sufficient for a state to fully meet the requirements of peer review when each district can choose which test to use has not been completely determined. Until one or more states goes through the official peer review process, sees what the peers deem to be adequate or incomplete information, and have received their decision letters from USED, the nature and depth of the information that states are required to gather and submit in their peer review submission materials is not fully known.

Third, a final decision from the USED will not occur until a plan such as Florida's is implemented, since several of the Critical Elements rely on evidence that can only be collected during an actual administration. An example of this is Critical Element 1.5 – Participation Data.

Finally, while USED has gone to great lengths to "standardize" the peer review process by the training provided to peer reviewers in advance of their work, peer review does remain to some extent an idiosyncratic process. It is possible for one peer panel to react a bit differently to the state-provided evidence than another peer panel would, due in part both to differences in how the peer review information is provided and differences in who serves to review the evidence.

Mindful of these caveats, however, ASG has submitted a prediction about the likelihood of a favorable outcome from peer review (plus an indication of additional information that might be provided by the state, the local districts in Florida, and/or the test vendors) by the conclusion of the project. This prediction can be found in Section 6 – Summary and Conclusions.

Peer Review Process Used in This Project

Planned Peer Review Process – The ASG team's innovative and comprehensive approach to the mock peer review under Criterion 6 included the following activities:

1. Examine current information and evidence from states that have used the ACT and the SAT as their high school ESSA accountability assessment on how they have addressed the requirements in the USED peer review, in particular those Critical Elements related to alignment, test development, accommodations, validity and other aspects of technical quality.

2. Create a hybrid peer review template for use by the Florida Department of Education, ACT, and the College Board to submit their evidence to demonstrate how adequately they have addressed each peer review Critical Element.

3. Request the Florida Department of Education, ACT, and the College Board provide the pertinent statements and supporting peer review evidence for each peer review Critical Element.

4. Add the statements and supporting peer review evidence for each peer review Critical Element exactly as submitted by the Florida Department of Education, ACT, and the College Board. Any unclear statements or incomplete references to evidence would be clarified with the respective organization or test vendor.

5. Prepare the relevant parts of the actual peer review document as if for submission to USED. Once the ASG partners' studies (i.e., alignment, accommodations, comparability, and accountability) are completed, add the evidence from each partner to that of the ACT and College Board for each Critical Element where relevant. This evidence will be clearly distinguished from that of the ACT or the College Board by separate headings for ACT/College Board evidence and that from ASG and its partners.

6. Carefully review and comment on the pieces of evidence submitted by ACT and the College Board, plus the ASG partners, to determine the completeness and accuracy of the information to support the use of these assessments at the high school level in place of the FSAs.

7. Provide a professional judgment on the likelihood of the ACT and/or SAT, when used as an optional high school test in place of the state's test, being approved by USED following actual peer review. ASG and its partners used a "peer review-like" process to accomplish this. The ASG team gathered peer evidence from the Florida Department of Education, ACT, and the College Board, including their draft of the pertinent sections of the actual *State Assessment Peer Review Submission Cover Sheet and Index Template,* the document that each state uses to submit its evidence of the technical qualities of its proposed assessments. The evidence from the Florida Department of Education was not subjected to peer review in this project because it had already been submitted to USED and the actual federal peer review of it had already taken place.

The peer review evidence compilation was split into two tracks: 1) the evidence for Critical Elements responded to by ACT or the College Board, and 2) supporting evidence related to these Critical Elements from the work that ASG and its partners carried out in Criteria 1-5. This included providing comments on the strengths and weaknesses of the evidence that was provided by each organization and recommendations about areas where improvements or additional evidence may be needed.

Note: If the option to offer the ACT or the SAT to high schools for use in lieu of the FSAs is implemented and the actual peer review evidence to seek federal approval of the option is assembled, not all evidence for peer review will be provided by ACT, the College Board, or for that matter, by the state. There are a few Critical Elements for which evidence in support the use of the ACT or the SAT will need to come from the school districts that use each test after the initial administration of these assessments in Florida's districts, which will be challenging for the state to collect.

An example of this is evidence of the participation of students with disabilities and English learners. Unless information on the participation of these students in the assessments can be collected by the Department as part of routine data collection, assembling this "local-use" data will add to the complexity of assembling the peer review evidence by the Florida Department of Education for submission to the USED.

Peer Review Expertise – The peer review process used by ASG was carried out by an individual (Ed Roeber) with extensive peer review experience, dating back to the first USED peer review in 1995. He has served as a reviewer of numerous states' peer review applications and supporting documentation under the current and two previous authorizations of the ESEA, i.e., IASA, NCLB, and ESSA). In addition, as director of assessment and accountability for the Michigan Department of Education (MDE), Dr. Roeber directed the submission of several peer review applications with supporting documentation from MDE. The most pertinent of these was Michigan's application to use the ACT as the high school state accountability assessment in

2007. Although Michigan was not the first user of a college entrance test for NCLB accountability purposes, it was the first state to be approved through the peer review process.

Dr. Roeber was joined in the peer review process by Dr. John Olson. They both served as peer reviewers in a USED-led peer review in 2017, and are scheduled for upcoming peer reviews in 2018. Other ASG team members, including Sheryl Lazarus and Martha Thurlow, also have served as peer reviewers in the past decade. ASG team members Norman Webb and Sara Christopherson have provided alignment evidence used by states in their peer review applications, while others such as Ed Roeber, John Olson, and Marianne Perie have assisted states in the preparation of their peer review applications.

Implementation of the Peer Review Process

Shortly after the award of the contract to ASG, preparatory work for peer review began. The following steps were those carried out. As noted above, ASG proposed to carry out several activities and most of these took place as planned. Each of the planned processes, and the results of what actually occurred, are described below.

1. Examine current information and evidence from states that have used the ACT and the SAT as their high school ESSA accountability assessment on how they have addressed the requirements in the USED peer review, in particular those Critical Elements related to alignment, test development, accommodations, validity, and other aspects of technical quality.

   Activities – When this search had concluded, a determination was made that information for the College Board SAT was not yet available. Calls to current users of the SAT for ESSA high school accountability (e.g., the Michigan assessment director) confirmed this. The College Board indicated that this information should be forthcoming at about the same time as this work concludes.

   On the other hand, there was considerable information on the use of the ACT as a state high school accountability exam. Unfortunately, quite a bit of this information for the use of the ACT preceded ESSA and its current peer review elements. For example, Michigan submitted its evidence to use the ACT as part of its NCLB high school accountability in 2007 and was fully approved that year, the first state to be so designated.

   Evidence from one state (Wisconsin) that used the ACT that submitted evidence for peer review under ESSA was located and examined (the USED decision letter can be found in **Appendix 5-B**).

   In its decision letter dated January 2017, the USED found that the state only "partially meets requirements" for the use of the ACT as its high school accountability assessment. Additional evidence was requested of Wisconsin for the peer review Critical Elements shown in Figure 5-3.

Figure 5-3. Critical Element for Which Wisconsin was Asked to Provide Additional Information on the ACT

Critical Element	Name of Critical Element
2.1	Test Design and Development
3.1	Overall Validity, including Validity Based on Content
3.3	Validity Based on Internal Structure
3.4	Validity Based on Relationships with Other Variables
4.1	Reliability
4.4	Scoring
4.7	Technical Analyses and Ongoing Maintenance
5.1	Procedures for Including Students with Disabilities
5.2	Procedures for Including ELs
5.3	Accommodations
5.4	Monitoring Test Administration for Special Populations
6.2	Achievement Standards-Setting
6.3	Challenging and Aligned Academic Achievement Standards
6.4	Reporting

Note that the Wisconsin Department of Public Instruction was asked to provide additional information for almost half of the peer review Critical Elements. As can been seen by examining the USED letter, some of the requests for additional information included items such as whether the state's standards are fully assessed by the ACT tests and whether the structure of some of state's standards (such as in ELA) match the manner in which the ACT results are reported. These and other issues will bear on the sorts of evidence that Florida will need to provide in its actual peer review submission.

2. Create a hybrid peer review template for use by the Florida Department of Education, ACT, and the College Board to submit their evidence to demonstrate how adequately they have addressed each peer review Critical Element.

   Activities – The latest version of the USED peer review document (*State Assessment Peer Review Submission Cover Sheet and Index Template)* was obtained. Normally, this is a three-column document providing locations for the Critical Elements to be listed and described in column 1, a brief overview of the evidence along with the list of evidence to address the Critical Element indicated in column 2, and pertinent notes that explains the evidence further, shown in column 3.

   For this project, ASG created a hybrid peer review form with four columns in it. The first column, copied from the USED document, contained the Critical Element number, title and description. This information was copied as is (that is, not changed on the hybrid peer review form). The second through fourth columns provided the locations for the peer review evidence to be pasted that was submitted by the Florida Department of Education, ACT and the College Board, respectively, to be placed.

The Florida Department of Education information was limited to information for just the high school FSA ELA and Algebra 1 EOC assessments . ASG requested that the Florida Department of Education provide the actual peer review submission information from 2016 that had resulted in a "substantially meets" designation from USED in early 2017 (see **Appendix 5-C**). The USED decision letter contained a list of the Critical Elements for which FDOE was asked to provide addition information. ASG requested that FDOE indicate what information it planned to submit (in December 2017) in response to the requested information.

The Florida Department of Education provided its submission with the information for just the FSA high school ELA and mathematics assessments shown. The information that USED requested was highlighted in yellow underneath the Critical Element description in column 1, and the information that Florida Department of Education planned on submitting was listed in column 2 in the Summary Statement row beneath the row of peer review evidence.

Table 5-2 shows the types of evidence expected from the Florida Department of Education, ACT, and the College Board, as well as from ASG's partners – WCEPS, NCEO, and CAARD.

Table 5-2. List of Critical Elements and ASG Study's Contributions to the Peer Review Evidence

Critical Element Number	Critical Element	FL DOE	ACT	College Board	Additional Evidence (ASG & Partners)			Peer Rview (ASG)
					WCEPS	NCEO	CAARD	
	Statewide System of Standards and Assessments							
1.1	State Adoption of Standards for All Students	NN	NN	NN				X
1.2	Coherent/Rigorous Standards	NN	S+	S+	X			X
1.3	Required Assessments	NN	NN	NN				X
1.4	Policies for Including All Students in Assessments	NN	NN	NN		X		X
1.5	Participation Data	NN	S/D+	S/D+				X
	Assessment System Operation							
2.1	Test Design & Development	X+	X	X	X			X
2.2	Item Development	X+	X	X	X			X
2.3	Test Administration	NN	X	X				X
2.4	Monitoring Test Administration	X+	X	X				X
2.5	Test Security	NN	X	X				X
2.6	Systems for Protecting Data integrity and Privacy	NN	X	X				X
	Validity							
3.1	Overall Validity, including Validity Based on Content	X+	X	X	X			X
3.2	Validity Based on Cognitive Processes	X+	X	X	X			X
3.3	Validity Based on Internal Structure	NN	X	X				X
3.4	Validity Based on Relationships with Other Variables	X+	X	X				X
	Reliability							
4.1	Reliability	X+	X	X			X	X
4.2	Fairness and Accessibility	NN	X	X		X		X
4.3	Full Performance Continuum	NN	X	X			X	X
4.4	Scoring	NN	X	X				X
4.5	Multiple Assessment Forms	NN	X	X				X
4.6	Multiple Versions of an Assessment	X+	X	X				X
4.7	Technical Analysis and Ongoing Maintenance	NN	X	X				X
	Inclusion of All Students							
5.1	Procedures for Including Students with Disabilities	NN	X	X		X		X
5.2	Procedures for Including ELs	NN	X	X		X		X
5.3	Accommodations	X+	X	X		X		X
5.4	Monitoring Test Administration for Special Populations	X+	S+/D+	S+?D+				X
	Academic Achievement Standards and Reporting							
6.1	State Adoption of Academic Achievement Standards for All Students	NN	NN	NN				X
6.2	Achievement Standards-Setting	NN	S+	S+			X	X
6.3	Challenging and Aligned Academic Achievement Standards	NN	S+	S+			X	X
6.4	Reporting	X+	D+	D+				X
	KEY TO CODES							
NN	Not Necessary to add evidence to what Florida has already submitted							
S+	State evidence needs to be augmented by ACT/College Board							
D+	Dstrict evidence may need to be collected to supplement state-collected data or if state data cannot be collected							
X+	Florida is providing additional information to USED							
X	Additional Evidence to the Supplied by WCEPS							
X	Additional Evidence to the Supplied by NCEO							
X	Additional Evidence to the Supplied by CAARD							

3. <u>Request to the Florida Department of Education, ACT, and the College Board to provide the pertinent statements and supporting peer review evidence for each peer review Critical Element</u>

<u>Activities</u> – On November 1, 2017, the FDOE, ACT, and the College Board were sent an e-mail (see **Appendix 5-D**) from ASG requesting that each organization provide the information needed for peer review by the deadline of November 25.

A secure Dropbox folder was set up for the College Board to upload its peer review template, a list of evidence provided, and each of the 33 pieces of supporting evidence. Each piece of evidence was given a unique ID number that corresponded to the Critical Element that it supported. Another secure Dropbox folder was provided to ACT for it to upload its peer review template, a list of evidence, and each piece of evidence. It also gave each piece of information a unique ID number.

4. Add the statements and supporting peer review evidence for each peer review Critical Element exactly as submitted by the Florida Department of Education, ACT, and the College Board. Any unclear statements or incomplete references to evidence would be clarified with the organization or testing vendor.

   Activities – The Florida Department of Education was first to respond. There were some minor editorial issues spotted by ASG and the Florida Department of Education was asked to accept these minor formatting changes, which it did promptly.

   The template provided by the College Board was reviewed. It was viewed as complete, and no follow-up was necessary. The submission from the College Board was included as-is in the master combined peer review document.

   The ACT-supplied information did not provide all of the requested information, so ASG requested additional detailed information from ACT. This included contextual information for the evidence citations and specific page number(s) for the evidence cited. ACT responded promptly to this request and uploaded a revised template to the secure Dropbox site. The revised template provided by ACT required no additional changes and the revised submission from ACT was added to the master combined peer review document.

5. Prepare the relevant parts of the actual peer review document as if for submission to USED. Once the ASG partners' studies (i.e., alignment, accommodations, comparability, and accountability) are completed, add the evidence from each partner to that of the ACT and College Board for each Critical Element.

   Activities – Next, each ASG partner was contacted about providing summary information from their studies, along with evidence citations (document name and page numbers) that could be added to the master peer review template. As mentioned above, this information was added to the master peer review template below the information provided by ACT and the College Board, in a specially-marked section. Since these studies could inform the peer review process, a summary of this information was added to the peer review template.

   Evidence from the studies carried out by ASG and its partners on alignment, accommodations, comparability, and accountability that yielded results pertinent to the peer review Critical Elements was added beneath the ACT- and College Board-submitted evidence in the peer review evidence rows. This evidence came from studies on the alignment of the ACT and SAT to Florida's content standards (Criteria 1 & 2), the comparability of the ACT and the SAT with comparable FSA assessments (Criterion 3), and

the accommodations afforded students with disabilities and English learners by ACT and the College Board (Criterion 4).

An example of the hybrid peer review template, with the evidence from ACT and the College Board, along with the ASG partner information, and with the addition of ASG peer review commentary, is shown in Figure 4.

Figure 5-4. Excerpt from the ASG-Prepared Hybrid Peer Review Template

	Evidence from Each of the Current or Proposed Assessments		
Critical Element	**Florida Standards Assessment**	**ACT**	**SAT**
1.4 – Policies for Including All Students in Assessments The State requires the inclusion of all public elementary and secondary school students in its assessment system and clearly and consistently communicates this requirement to districts and schools. • For students with disabilities (SWD), policies state that all students with disabilities in the State, including students with disabilities publicly placed in private schools as a means of providing special education and related services, must be included in the assessment system; • For English learners (EL): o Policies state that all English learners must be included in the assessment system, unless the State exempts a student who has attended schools in the U.S. for less than 12 months from one administration of its reading/ language arts assessment; If the State administers native language assessments, the State requires English learners to be assessed in reading/language arts in English if they have been enrolled in U.S. schools for three or more consecutive years, except if a district determines, on a case-by-case basis, that native language assessments would yield more accurate and reliable information, the district may assess a student with native language assessments for a period not to exceed two additional consecutive years.	**ELA and Mathematics:** **Evidence 01** – USDOE approval letter for Florida's most recent ESEA flexibility request. **Evidence 02** – Florida's ESEA flexibility request for 2014-2015, including rigorous academic standards in math and ELA/reading (with 02a, Appendix), pgs. 46-51. **Evidence 05** – Florida Statute 1008.22 - Student assessment program for public schools; subsection (3) **Evidence 06** – Rule 6A-1.0943 Statewide Assessment for Students with Disabilities. **Evidence 06EL** – Rule 6A-6.0909 Exemptions Provided to English Language Learners **Evidence 06TAP** – Statewide Assessment for SWDs Technical Assistance Paper **Evidence 07** - Rule 6A-1.09422 Statewide Standardized Assessment Program Requirements; (1)(c) **Evidence 07EL** – District English Language Learner Plan – Template **Evidence 18ELrule** – Rule 6A-6.09091, Accommodations for English Language Learners **Evidence 19** – Statewide Assessment Accommodations	**State-Specific Evidence** [The evidence provided by the state should be sufficient.] **Additional Evidence of Inclusion of Students with Disabilities (ASG/NCEO)** [The Accommodations Studies did not address participation requirements other than indicating that if the students needed accommodations, they needed to apply for them.]	**State-Specific Evidence** [The evidence provided by the state should be sufficient.] **Additional Evidence of Inclusion of Students with Disabilities (ASG/NCEO)** [The Accommodations Studies did not address participation requirements other than indicating that if the students needed accommodations, they needed to apply for them.]
Section 1.4 Summary Statement			
___ No additional evidence is required, or ___ Additional evidence is needed/provide brief rationale [List additional evidence needed with brief rationale]		**Peer Review (ASG)** The evidence provided by the State indicates that it has adequately addressed this Critical Element.	**Peer Review (ASG)** The evidence provided by the State indicates that it has adequately addressed this Critical Element.

6. Carefully review and comment on the pieces of evidence submitted by ACT and the College Board, plus the ASG partners, to determine the completeness and accuracy of the information to support the use of these assessments at the high school level in place of the FSAs.

Activities – Once all of the evidence from ACT, the College Board, and ASG partners had been received and added to the hybrid combined peer review template, the actual peer review process began. First, the evidence for the 16 Critical Elements that did not require information beyond that submitted by the ACT or the College Board was reviewed. This review proceeded through the template from the first to the last Critical Element. This first part of the review was completed before the results of the ASG partner studies were received so as to facilitate the timely review of the evidence.

For each Critical Element, first the text supplied by ACT was reviewed, along with each supporting document. The evidence was first reviewed to make sure it was pertinent to the Critical Element, which was almost always the case. Then, the evidence was reviewed for sufficiency – does the set of evidence (looking across all pieces submitted by ACT for the particular Critical Element) sufficiently address the ESSA requirements for that Critical Element? Finally, the set of evidence was reviewed for adequacy – does the evidence adequately address the requirements of the Critical Element. The results of this review were recorded in the Summary box under the ACT-provided evidence.

This process was repeated for the same Critical Element for the College Board-supplied evidence for the SAT. The same questions were addressed – pertinence, sufficiency, and adequacy. All of the College Board-supplied evidence was reviewed, and the summary of the review was also recorded in the Summary box, this time under the College Board column.

This process was repeated for each of the 15 remaining Critical Elements for which ASG partner information was unnecessary.

Once ASG partner information was received, it was added to the combined peer review form. At this point, the ACT and partner information, as well as the College Board and partner information could be reviewed in combination. The same process of examining the combined evidence for pertinence, sufficiency, and adequacy was used. And, once again, each review was summarized in the Summary box under either the ACT or the College Board columns.

Before the combined peer review template was finished, the draft peer review summaries from ASG staff for Critical Elements on which ASG partner information had been added was provided to the pertinent ASG partner for their review and approval. This was done to assure that the conclusions drawn by ASG would be supported by the ASG partners.

In total, there are 30 Critical Elements and a review of the ACT and of the SAT for each of these Elements. This included providing comments on the strengths and weaknesses of the evidence that was provided and recommendations on the areas where improvements or additional evidence may be needed.

7. Provide a professional judgment on the likelihood of the ACT and/or SAT, when used as an optional high school test in place of the state's test, will be approved by USED following actual peer review.

Activities – Once the overall combined peer review template had been filled out with the summaries for each test and for each Critical Element, the ASG peer review team provided its overall judgment across all 30 Critical Elements to render its judgment as to the likelihood that offering Florida high schools an option of which exam to use (the FSAs, the ACT, or the SAT) would receive federal approval. We note that not all evidence for peer review will be provided by ACT or the College Board. If a program such as Florida's is actually implemented, there will peer review evidence in support of the use of the ACT or the SAT that will need to come from school districts that use each exam, after the initial administration of these assessments in Florida's districts. Collecting this "local-use" data will add to the complexity of Florida Department of Education's ultimate peer review submission to the USED.

Findings

The complete combined peer review template, containing the evidence for peer review provided by the Florida Department of Education, ACT, College Board, and the three ASG partners (WCEPS, NCEO, and CAARD), along with the peer review commentary from ASG, is found in Appendix 5E. The peer review commentary is shown in the following tables.

The peer review can also be summarized in a more qualitative manner as to how the evidence was judged through the mock peer review process conducted by ASG. These qualitative results are shown in Table 5-3.

Table 5-3. ASG Summary Judgments of Peer Review Evidence

Critical Element	ACT	College Board
1.1 – State Adoption of Academic Content Standards for All Students		
1.2 – Coherent and Rigorous Academic Content Standards		
1.3 – Required Assessments		
1.4 – Policies for Including All Students in Assessments		
1.5 – Participation Data		
2.1 – Test Design and Development		
2.2 – Item Development		
2.3 – Test Administration		
2.4 – Monitoring Test Administration		
2.5 – Test Security		
2.6 – Systems for Protecting Data Integrity and Privacy		
3.1 – Overall Validity, including Validity Based on Content		
3.2 – Validity Based on Cognitive Processes		
3.3 – Validity Based on Internal Structure		
3.4 – Validity Based on Relationships with Other Variables		
4.1 – Reliability		
4.2 – Fairness and Accessibility		
4.3 – Full Performance Continuum		
4.4 – Scoring		
4.5 – Multiple Assessment Forms		

4.6 – Multiple Versions of an Assessment		Not Applicable
4.7 – Technical Analysis and Ongoing Maintenance		
5.1 – Procedures for Including Students with Disabilities		
5.2 – Procedures for Including ELs		
5.3 – Accommodations		
5.4 – Monitoring Test Administration for Special Populations		
6.1 – State Adoption of Academic Achievement Standards for All Students		
6.2 – Achievement Standards-Setting		
6.3 – Challenging and Aligned Academic Achievement Standards		
6.4 – Reporting		

Legend
Met Mock Peer Review Requirements
May Not Meet Mock Peer Review Requirements
Did Not Meet Mock Peer Review Requirements

The qualitative judgments shown in Table 5-3 can be summarized in an overall look at peer review shown in Table 5-4.

Table 5-4 Number of Peer Review Critical Element Determinations by College Entrance Test

Peer Review Determinations	ACT	College Board
Met Mock Peer Review Requirements	23	20
May Not Meet Mock Peer Review Requirements	1	6
Did Not Meet Mock Peer Review Requirements	6	3
TOTAL	30	29*

* One Critical Element – related to online assessment - is not applicable to the current College Board SAT.

The complete summaries from ASG in the peer review process concerning the evidence provided for each Critical Element is displayed in Table 5-5. The complete combined peer review template can be found in **Appendix 5-E**.

Table 5-5 ASG Summaries for Peer Review Evidence

Critical Element	ACT	College Board
1.1 – State Adoption of Academic Content Standards for All Students	The evidence provided by the State indicates that it has met the requirements for this Critical Element.	The evidence provided by the State indicates that it has met the requirements for this Critical Element.
1.2 – Coherent and Rigorous Academic Content Standards	The evidence provided by the State indicates that it has met the requirements for this Critical Element.	The evidence provided by the State indicates that it has met the requirements for this Critical Element.
1.3 – Required Assessments	The evidence provided by the State indicates that it has met the requirements for this Critical Element.	The evidence provided by the State indicates that it has met the requirements for this Critical Element.

1.4 – Policies for Including All Students in Assessments	The evidence provided by the State indicates that it has met the requirements for this Critical Element.	The evidence provided by the State indicates that it has met the requirements for this Critical Element.
1.5 – Participation Data	The evidence provided by the State indicates that it has met the requirements for this Critical Element.	The evidence provided by the State indicates that it has met the requirements for this Critical Element.
2.1 – Test Design and Development	The evidence provided by ASG/WCEPS indicates that the ACT does not meet the requirements for content coverage of this Critical Element. Without augmentation of the ACT, the ACT would likely not be approved for use in lieu of the Florida Standards Assessment in ELA or in Mathematics. Note: the evidence described in the ACT Technical Report is how ACT described alignment. It is not an independent alignment study for the ACT.	The evidence provided by ASG/WCEPS indicates that there was acceptable alignment of one SAT form with the LAFS, but another form that would require slight adjustments to meet the minimum cutoffs for full alignment. Similarly, for mathematics, there was acceptable alignment of one SAT form with the Algebra 1 standards, but another form that would require slight adjustments to meet the minimum cutoffs for full alignment. As currently configured, the SAT might not meet peer review requirements, and thus, might not be approved for use in lieu of the Florida Standards Assessment in ELA or in Mathematics.
2.2 – Item Development	The evidence provided by ACT and in the ASG/WCEPS alignment study show that adequate attention was provided to the process of item development. As shown in Tables 1b-9.2a and 1b-9.2b in the ASG/WCEPS ELA report, the ACT did not meet the minimum DOK Consistency requirements in two of the four reporting categories for one form, and for the other, did not meet the minimum requirement for one reporting category and did so weakly for another. In Mathematics (Tables 1a-9.3 and 1a-9.4), the ACT showed acceptable DOK Consistency	The evidence provided by the College Board and in the ASG/WCEPS alignment study show that adequate attention was provided to the process of item development. As shown in Tables 1b-9.3a and 1b-9.3b in the ASG/WCEPS ELA report, the SAT did not meet the minimum DOK Consistency requirements for one out of four reporting categories in each form studied, and weakly met the minimum requirements in another reporting category in one form. In Mathematics (Tables 1a-9.5 and 1a-9.6), the ACT showed acceptable DOK Consistency levels in all three reporting categories.

	levels in all three reporting categories. The Range of Knowledge Correspondence was judged to be weak or low in all three reporting categories. Thus, the ACT did not meet the requirements for this Critical Element.	The Range of Knowledge Correspondence was judged to be weak or low in two out of three reporting categories. Thus, the SAT may not meet the requirements for this Critical Element.
2.3 – Test Administration	The materials provided by ACT indicate that it has adequately established and communicated to educators the same clear, thorough and consistent standardized procedures for the school day administration as those used in national testing administrations. ACT has established procedures to ensure that all individuals responsible for administering the ACT receive training on the established administration procedures for its assessments; All educators involved with testing receive training and are certified. Note: Should Florida elect to offer schools the ACT, participating districts will need to provide evidence that educators involved with the administration of the ACT have been adequately trained before the administration of the ACT. ACT has met the requirements for this Critical Element.	The materials provided by the College Board indicate that it has adequately established and communicated to educators the same clear, thorough and consistent standardized procedures for the school day administration as those used in national testing administrations. The College Board has established procedures to ensure that all individuals responsible for administering the SAT receive training on the established administration procedures for its assessments; All educators involved with testing receive training and are certified. (See Evidence #2.1.1, 2.3.1, 2.3.2, 2.3.3, and 2.3.4.) Note: Should Florida elect to offer schools the SAT, participating districts will need to provide evidence that educators involved with the administration of the SAT have been adequately trained before the administration of the SAT. The College Board has met the requirements for this Critical Element.
2.4 – Monitoring Test Administration	ACT has adequately summarized the procedures it has in place to monitor the administration of the ACT when used as a state assessment. These include observations of assessment administration, records collected from each site, and port-test	The College Board has extensively and adequately documented the procedures it has in place to monitor the administration of the SAT when used as a state assessment. These include observations of assessment administration, records collected

	checks on students' responses (see Exhibit #8). Considerable monitoring of the assessment administrations occurs. These procedures help to ensure that school day administrations are comparable to national test date administrations, with test administration procedures implemented with fidelity across districts and schools. Note: Should Florida elect to offer schools the ACT, participating districts will need to monitor the administration of the ACT in addition to the monitoring provided by the ACT. ACT has met the requirements for this Critical Element.	from each site, and port-test checks on students' responses. (See Evidence 2.1.1, 2.3.1, 2.3.2, 2.3.3, and 2.3.4.). Considerable monitoring of the assessment administrations occurs. These procedures help to ensure that school day administrations are comparable to national test date administrations, with test administration procedures implemented with fidelity across districts and schools. Note: Should Florida elect to offer schools the SAT, participating districts will need to monitor the administration of the SAT in addition to the monitoring provided by the College Board. The College Board has met the requirements for this Critical Element.
2.5 – Test Security	The evidence submitted by ACT shows that it has adequate policies and procedures in place to prevent test irregularities and ensure the integrity of test results. ACT has procedures for maintaining the security of test materials, proper administration procedures (see Exhibit #8), incident-reporting procedures (see Exhibits #8 and #10), investigations of incidents, consequences for confirmed violations of test security (see Exhibits #8 and #10), and requirements for annual training for all individuals involved in test administration. It also has procedures to detect test irregularities and investigates alleged or actual test irregularities (see Exhibit #8).	The evidence submitted by the College Board shows that it has adequate policies and procedures in place to prevent test irregularities and ensure the integrity of test results. The College Board has procedures for maintaining the security of test materials (Evidence #2.1.1 and 2.3.1), proper administration procedures (see Evidence #2.3.1, 2.3.2, 2.3.3, and 2.3.4), incident-reporting procedures (see Evidence #2.5.1), investigations of incidents, consequences for confirmed violations of test security (see Evidence # 2.1.1, 2.3.1, 2.3.2, 2.3.3, 2.3.4, and 2.5.1), and requirements for annual training for all individuals involved in test administration (see Evidence #2.3.4)

	ACT has met the requirements of this Critical Element.	It also has procedures to detect test irregularities and investigates alleged or actual test irregularities. The College Board has met the requirements of this Critical Element.
2.6 – Systems for Protecting Data Integrity and Privacy	The evidence submitted by ACT shows that the ACT has adequate procedures in place to protect the integrity of its test materials (see Exhibit #8). ACT's procedures to protect student privacy and the confidentiality of student-identifiable information are shown in Evidence test data (see Exhibits #10 and 11). ACT has met the requirements for this Critical Element.	The evidence submitted by the College Board indicates that it has robust procedures in place to protect the integrity of its testing materials is outlined in Evidence #2.3.1 and 2.1.1. College Board's procedures to protect student privacy and the confidentiality of student-identifiable information are shown in Evidence #2.6.1, 2.6.2, 2.6.3, 2.6.4, and 2.6.5. The College Board has met the requirements for this Critical Element.
3.1 – Overall Validity, including Validity Based on Content	The evidence provided through the WCEPS indicates that a significant number of items (see above) would need to be modified or replaced for the ACT to measure the knowledge and skills specified in Florida's mathematics and ELA content standards. The alignment between Florida's content standards (including the content knowledge and processes, range, balance of content, and cognitive complexity) and the ACT appears to be inadequate in both subject areas. This would require augmenting the ACT, adding testing time, complexity (e.g., an additional test day), and costs to the use of the ACT in lieu of the FSA. The alignment study also demonstrated that there are some form-to-form differences on the ACT. This might require annual adjustments to the augmented section for the ACT, which would	The evidence provided through the WCEPS alignment study indicates that some items would need to be modified or replaced for the SAT to measure the knowledge and skills specified in Florida's mathematics and ELA content standards. The alignment between Florida's content standards (including the content knowledge and processes, range, balance of content, and cognitive complexity) and the SAT appears to be nearly acceptable in both subject areas. This might require augmenting the SAT, adding testing time, complexity (e.g., an additional test day), and costs to the use of the SAT in lieu of the FSA. The alignment study also demonstrated that there are some form-to-form differences on the SAT. This might require annual adjustments to the augmented section for the SAT, which would add additional costs and

	add additional costs and complexities of offering the test to high schools. The ACT has not met the requirements for this Critical Element.	complexities of offering the test to high schools. The College Board may not have met the requirements for this Critical Element.
3.2 – Validity Based on Cognitive Processes	The evidence provided through the WCEPS alignment study lends further support to the need to augment the ACT with additional items, in this case for the ELA test (see Critical Element 3.1). Ten or more revised or replaced items are needed for the ELA assessment to strengthen the DOK Consistency and Range of Knowledge of the ELA test. It may also be advantageous to augment the Mathematics test as well with seven or eight items revised or replaced in order to strengthen its DOK Consistency and Range of Knowledge. Without the augmentation spelled out for this Critical Element (and Critical Element 3.1), the ACT does not meet the requirements for this Critical Element (and Critical Element 3.1, too).	The evidence provided through the WCEPS alignment study lends further support to the need to augment the SAT ELA with additional items (see Critical Element 3.1). Adding or replacing five to seven items in the ELA assessment would strengthen the DOK Consistency of the ELA test. This augmentation might also address the weakness in one Range of Knowledge criterion. In mathematics, four to seven replaced or revised items are needed to address the Range of Knowledge reporting category issues. Without the augmentation spelled out in this Critical Element and Critical Element 3.1, the College Board may not meet the requirements for this Critical Element (and Critical Element 3.1, too).
3.3 – Validity Based on Internal Structure	ACT has provided adequate information on the internal structure of the ACT tests (see Evidence #14). The sub-score structure of the ACT sub-tests supports the structures of the ACT test reports (see Evidence #14 and #13). The DIF analyses indicated the vast majority of items do not function differently for different subgroups of students. This provides evidence of another aspect of validity. The ACT has met the requirements for this Critical Element.	The College Board has provided adequate statistical evidence of the internal consistencies of each test (see Evidence #2.1.1, Table A-6.2) and test inter-correlations (see Evidence #2.1.1 Tables A-6.9.1-A-6.9.3). Evidence that the SAT tests are related to the structures of the standards on which they are based is cited in Evidence #2.1.2. The College Board has met the requirements for this Critical Element.

3.4 – Validity Based on Relationships with Other Variables	The ACT has adequately documented the key validity evidence that the ACT scores are related as expected with other variables. These include high school and college course work (see Exhibit #4 (Table 5.20) and Exhibits #16 and #17, as well Evidence of college performance (see Exhibits #20-#22) and evidence of college freshman success (see Exhibits #21 and 26). The relationships between the ACT and other external measures are described in Exhibits #23-26. Note: If the ACT is augmented to better align the ACT with Florida's standards (before it is used by Florida districts in place of the Florida Standards Assessments), these analyses may need to be conducted anew. The ACT has met the requirements for this Critical Element.	The College Board has adequately documented the key validity evidence that the SAT scores are related as expected with other variables – previous versions of the SAT, college freshman GPA, and college and career readiness benchmarks (see Evidence #2.1.1). Old and new SAT tests were concorded, and the ACT and the SAT tests were concorded as well. A predictive validity study was also carried out (which included high school GPA and SAT scores - see Table 7.4, as well as college freshman grades - see Table 7.5). Finally, college readiness thresholds were also set (see Table 7.6). Note: If the SAT is augmented to better align the SAT with Florida's standards (before it is used by Florida districts in place of the Florida Standards Assessments), these analyses may need to be conducted anew. The College Board has met the requirements for this Critical Element.
4.1 – Reliability	ACT has met the requirements for this Critical Elements.	The College Board has met the requirements for this Critical Elements.
4.2 – Fairness and Accessibility	The evidence provided by ACT indicates that it has taken reasonable steps to ensure that its assessments are accessible to all students. The ASG/NCEO accommodations study (indicates that the ACT provides a full range of accommodations for students with disabilities, ELs, and ELs with disabilities. This will permit these students to adequately access the ACT and receive a score that can be used for accountability purposes.	The evidence provided by the College Board indicates that it has taken reasonable steps to ensure that its assessments are accessible to all students. The ASG/NCEO accommodations study indicates that the College Board provides a full range of accommodations for students with disabilities on the SAT. This will permit these students to adequately access the SAT and receive and receive a score that can be used for accountability purposes.

	The ACT has met the requirements for this Critical Element.	The range of accommodations for English learners for ELs is not quite as broad as for students with disabilities though it would not be anticipated that it should be since the SAT list of accommodations are for ELs only (not those with a disability), meaning that the College Board list is not comparable to that of ACT. The College Board has met the requirements for this Critical Element.
4.3 – Full Performance Continuum	ACT provides somewhat less coverage of the full performance continuum, but does provide a precise measure for ELA. The precision for mathematics is more limited. The ACT has met the requirements of this Critical Element.	SAT adequately meets the range of the performance continuum and is similar in precision to the FSA for both the ELA and mathematics. The College Board has met the requirements for this Critical Element.
4.4 – Scoring	The ACT has adequately documented the standardized scoring procedures and protocols it uses that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State's academic achievement standards. For example, scaling of the ACT tests is described in Exhibits #4, pages 40-45. Scoring of the ACT Writing test is described in Exhibits #13, pages 12-13. The ACT has met the requirements of this Critical Element.	The SAT has adequately documented the standardized scoring procedures and protocols it uses that are designed to produce reliable results, facilitate valid score interpretations, and report assessment results in terms of the State's academic achievement standards (see Evidence #2.1.1 (pages 59 and 72-88 and 4.1.2 (pages 8-74). The College Board has met the requirements for this Critical Element.
4.5 – Multiple Assessment Forms	The equating procedures used by ACT to assure the comparability of multiple forms of the ACT are described adequately in Exhibits #4 and #27. The construction of multiple forms is not described adequately in Exhibit #5.	The equating procedures used by College Board to assure the comparability of multiple forms of the SAT described in Evidence #2.1.1. However, peers feel that additional evidence is needed that

	The ACT has met the requirements of this Critical Element.	the College Board develops multiple forms for state use. The College Board may not have met the requirements of this Critical Element.
4.6 – Multiple Versions of an Assessment	The comparability of the ACT as administered on paper and via computer is described in Exhibit #4, in sections 4.1-4.3. Section 4.3.2 specifically presents information from a mode study that ACT carried out. In the first study it carried out, ACT found that online scores were slightly higher for the students who took the ACT online and mode adjustments were made (and online timing was adjusted). Student showed no differences in performance in the second mode study. The ACT has met the requirements for this Critical Element	This Critical Element is not applicable to the College Board.
4.7 – Technical Analysis and Ongoing Maintenance	ACT provided considerable and adequate evidence on how it maintains and improves the qualities of its assessments. It uses a TAC (technical advisory committee --see Exhibit #29), periodically updates the content standards measured through its National Curriculum Study (Exhibit # 17), and implements ongoing technical analyses and improvements (Exhibit #4). The ACT met the requirements for this Critical Element.	The College Board did not respond adequately to this Critical Element, since the evidence cited above by the College Board for this Critical Element does not indicate work to maintain and improve the qualities of its assessments. The evidence cited for Critical Element 4.7 relates to the creation and validity of the SAT tests, not how the tests are maintained and improved, which is the subject of this Critical Element. Some relevant information is provided in the Technical Report (Evidence #2.1.1) The College Board did not meet the requirements for this Critical Element.
5.1 – Procedures for Including Students with Disabilities	The evidence provided by ACT indicates that ACT has	The evidence provided by the College Board indicates that it has

	adequately addressed this Critical Element. The ASG/NCEO accommodations study indicates that the ACT provides a full range of accommodations for students with disabilities. This will permit these students to adequately access the ACT and receive a score that can be used for accountability purposes. The ACT has met the requirements for this Critical Element.	adequately addressed this Critical Element. The ASG/NCEO accommodations study indicates that the SAT provides a full range of accommodations for students with disabilities. This will permit these students to adequately access the SAT and score that can be used for accountability purposes. The College Board has met the requirements for this Critical Element.
5.2 – Procedures for Including ELs	The ACT provides adequate information on the availability and use of accommodations for English learners as well as available tools and accessibility features. The list of available accommodations from the ACT is so large because the list includes accommodations for ELs and ELs with disabilities. The ACT has met the requirements for this Critical Element.	The College Board provides adequate information on the availability and use of accommodations for English learners. The College Board provides a small number of accommodations for English learners because its list is for ELs only. This may mean that College Board has not adequately provided guidance regarding that group of students who are <u>both</u> ELs and students with disabilities (i.e., ELs with disabilities). The College Board may not have met the requirements for this Critical Element.
5.3 – Accommodations	The evidence provided by ACT and the ASG/NCEO study indicates that a range of accommodations is available to students with disabilities and students covered by Section 504. Data from validity studies would bolster this evidence. The ACT and ASG/NCEO evidence also indicates that the ACT provides an appropriate range of accommodations to English learners and English learners with disabilities. Data	The evidence provided by the College Board and the ASG/NCEO study indicates that a range of accommodations is available to students with disabilities and students covered by Section 504. Data from validity studies would bolster this evidence. The ACT and ASG/NCEO evidence also indicates that the SAT provides an appropriate range of accommodations to English learners who do not have

	from validity studies would also bolster this evidence The ACT has met the requirements for this Critical Element.	disabilities. Data from validity studies would also bolster this evidence. The College Board and ASG/NCEO evidence is less clear as to whether the College Board provides accommodations that are appropriate and effective for English learners with disabilities since the needs of this population were not specifically addressed in tables provided by College Board. This could mean that fewer English learners with disabilities will have access to accommodations that will result in a score that can be used for accountability purposes, and that will allow for meaningful interpretations. The College Board may not have met the requirements for this Critical Element.
5.4 – Monitoring Test Administration for Special Populations	ACT has described what accommodations are available, how these can be requested, what information is needed to approve accommodations, and the supports available to students. These could form the bases for monitoring. However, this evidence does not indicate how ACT monitors whether students receive suitable accommodations, that its accommodations are in line with state accommodations policies, and that the accommodations are those used during instruction, consistent with the students' IEPs or 504 plans. No evidence of a state or district training plan and supporting materials state ACT school-day administrations are provided.	The College Board has described in the evidence submitted for Critical Elements 5.1 and 5.2 how it reviews and approves requests for accommodations as well as the resources available to ELs (word-to-word glossaries and test directions). These could form the bases for monitoring. This evidence does not indicate how the College Board monitors whether students receive suitable accommodations, that they are in line with state accommodations policies, and that they are those used during instruction, consistent with the students' IEPs or 504 plans. No evidence of a state or district training plan and supporting materials for state SAT school-day administrations are provided.

	The ACT has not met the requirements for this Critical Element.	The College Board has not met the requirements for this Critical Element.
6.1 – State Adoption of Academic Achievement Standards for All Students	The State has met the requirements for this Critical Element under its submission for the current FSA components.	The State has met the requirements for this Critical Element under its submission for the current FSA components.
6.2 – Achievement Standards-Setting	The ACT has met the requirements for this Critical Element.	The College Board has met the requirements for this Critical Element.
6.3 – Challenging and Aligned Academic Achievement Standards	The ACT scores are not interchangeable with the FSA scores: "…it appears the ACT and SAT do not produce results comparable to the FSA and should not be considered alternatives to the ELA grade 10 or Algebra 1 EOC assessments." The ACT has not met the requirements for this Critical Element.	The SAT scores are not interchangeable with the FSA scores: "…it appears the ACT and SAT do not produce results comparable to the FSA and should not be considered alternatives to the ELA grade 10 or Algebra 1 EOC assessments." The College Board has not met the requirements for this Critical Element.
6.4 – Reporting	ACT provided examples of the reports that are used to convey the ACT score information to various audiences (see Exhibit #3, pages 3.1 – 3.7). However, ACT did not provide information on how its reports of results "…facilitate timely, appropriate, credible, and defensible interpretations and uses of results for students tested by parents, educators, State officials, policymakers and other stakeholders, and the public." In addition, ACT did not provide information on the use of results, "including itemized score analyses, to districts and schools so that parents, teachers, principals, and administrators can interpret the results and address the specific academic needs of students… and also provide interpretive guides to support appropriate uses of the assessment results."	The College Board provided adequate examples of the reports that are used to convey the SAT score information to various audiences (see Evidence #6.4.2, 6.4.3, 6.4.4, and 6.4.6). The College Board has provided information on how it supports users in using the SAT results to inform instruction (see Evidence #6.4.4 and 6.4.5). Evidence of the timely delivery of SAT score reports and providing the reports in alternative formats (e.g., in languages other than English) was not provided. The College Board has met the requirements of this Critical Element.

	Also, evidence of the timely delivery of the ACT score reports and providing the reports in alternative formats (e.g., in languages other than English), were not provided in either the document cited above or in the online *The ACT User Handbook for Educators.* The ACT may not have met the requirements of this Critical Element.	

Discussion

As can be seen, it is the judgment of the ASG peer reviewers that the plan to offer the use of the ACT and/or the College Board in lieu of the Florida Standards Assessments has important issues in it that are relevant for peer review. These include issues in the following areas:

1. Alignment – The ASG/WCEPS alignment study showed areas where additional assessment items would need to be added in order to enhance the alignment of the FSAs with their counterparts in the ACT and the SAT. Without such augmentation, it is the judgment of ASG that peer approval of the ACT will likely not occur and such approval of the SAT may not occur.
2. Accommodations – A substantial number of accommodations are offered for students with disabilities on the ACT and the SAT as well as the FSAs. English learners do not have accommodations that result in college-reportable scores available for the SAT in the same proportion as students with disabilities, an unfairness that may affect peer approval.
3. Comparability – While it is possible to determine equivalent scores on all three tests (FSAs, ACT, and SAT), when this is done, the levels of classification consistency (i.e., the frequency with which high school students are placed at the same performance level on each of the three tests) is unacceptably low (less than 50% in many cases). For peer review, this indicates an unacceptable level of alignment of achievement standards between the three tests and, thus, does not meet the mock peer review requirements.
4. Accountability – The lack of classification consistency means that performance of students at each performance level will not be comparable across schools in the state using different tests. This jeopardizes the school performance designations, a key aspect of a state's accountability system.

Summary and Conclusions

The fundamental goal of this project was to determine if offering the use of the ACT and the SAT to Florida's schools in lieu of the Florida Standards Assessments would likely receive approval through the USED ESSA peer review process. Several steps were followed to assemble evidence from the Florida Department of Education, ACT, and the College Board. This evidence was added to a modified peer review template designed to show each of the 30 peer review Critical Elements, as well as the evidence in support of the Florida Standards Assessments, ACT, and SAT. Evidence from other ASG partners (WCEPS, alignment; CAARD, comparability and accountability; and, NCEO, accommodations) was also added to the modified peer template.

Once all evidence was added to the peer review template, ASG staff with substantial experience in the USED peer review process reviewed the evidence, commented on its completeness, sufficiency, and adequacy for each Critical Element, and then rendered a judgment on each Critical Element as to whether each test met peer review requirements, may not meet peer review requirements, or did not meet peer review requirements.

When the mock peer review was completed by ASG, the peer review template was reviewed by the ASG partners for the Critical Elements relevant to the work of each partner. When these reviews were completed, the peer review template was updated, as were the tables in this report. The following main points were made:

- The alignment study demonstrated that a significant percent of ELA and mathematics items would need to be revised or added to the ACT and to a lesser extent, the SAT, adding additional testing time, administration complexity, and costs to the state assessments. This means that the ACT will likely not be peer-approved without augmentation. The SAT may not been peer-approved, depending on the test form used.

- The comparability study indicated unacceptably low classification consistency between the three tests, meaning that many of the students who achieve any of the five Florida performance levels on the FSAs would not be placed at the same performance level on the ACT or the SAT (i.e., some would be placed in higher performance levels and others placed at lower levels).

- The accommodations studies showed that all three tests offer many accommodations to students with disabilities, while English learners (especially on the SAT) have fewer accommodations available to them than students with disabilities that result in non-college-reportable scores. This may jeopardize the peer approval of the SAT. Additional accommodations that result in college reportable scores, especially for English learners, would increase the fairness of the SAT. However, it is not certain that these changes would help a plan such as Florida's meet peer review requirements.

While Congress felt that the ACT or the SAT could be used as *the* high school exam for ESSA accountability purposes, and perhaps could be used interchangeably in a state with each other or perhaps the state assessments, such assumptions were included in the ESSA legislation

without having been tested in advance. Congress intends for states to do this testing and still meet all of the requirements for high quality assessments, since it included a proviso that any nationally recognized assessment that was selected still needs to meet peer review requirements (see **Appendix 5.A**, 200.3(b)(1)-(3)) in the legislation).

Florida did exactly this by conducting this study. In the case of Florida, using an end-of-course test in Algebra 1 limits the alignment and comparability of the ACT and SAT with the FSAs. The tests could be augmented, but this adds complexity, costs, and added testing sessions and time to each test. Whether this would improve the classification consistency to acceptable levels, however, is uncertain.

In conclusion, it is the belief of ASG and its partners is that the option to provide Florida districts with the option of using the ACT or SAT as they currently exist, in lieu of the FSAs, will likely not meet ESSA peer review requirements.

References

Bureau of Contracts, Grants, and Procurement Management Services (2017). *Request for proposals: Feasibility of the use of the ACT and SAT in lieu of statewide assessments (RFP2018-48).* Orlando FL: Florida Department of Education.

Council of Chief State School Officers. (2017). *Implementing the Locally-Selected, Nationally-Recognized High School Assessment Provision of the Every Student Succeeds Act: Key Questions and Considerations.* Washington, DC: Author.

Council of Chief State School Officers. (2017). *An Implementation Framework for the Locally-Selected, Nationally-Recognized High School Assessment Provision of the Every Student Succeeds Act* Washington, DC: Author.

U.S. Department of Education. (2015) Peer Review of State Assessment Systems Non-Regulatory Guidance for States for Meeting Requirements of the Elementary and Secondary Education Act of 1965, as amended. Washington, DC: Author.

Section 6 - Summary and Conclusions

The Every Student Succeeds Act (ESSA) allows states to approve a school district to administer, in lieu of the statewide high school assessment, a "locally selected, nationally recognized" high school academic assessment that has been approved for use by the state (including submission for the U.S. Department of Education (USED) assessment peer review process).

ESSA specifies that certain technical criteria must be satisfied to receive approval for use by the state. These requirements should be considered minimum standards, meaning the state may establish additional requirements. ESSA requires that the assessment chosen by the state

- Is aligned to and addresses the breadth and depth of the state's content standards
- Is equivalent to the statewide assessments in its content coverage, difficulty, and quality
- Provides valid and reliable data on student achievement for all students and subgroups as compared to the statewide assessments[7]
- Meets the criteria for technical quality that all statewide assessments must meet (e.g., peer reviewed)
- Provides unbiased, rational, and consistent differentiation among schools within the state's accountability system." (CCSSO, 2017a).

While Congress allowed states to approve a nationally recognized assessment for local use, it left it to the states to ensure that their revised assessment systems continued to meet all of the technical requirements (stated above) necessary to pass peer review. The standards required for states submitting assessment systems that include the use of a nationally recognized assessment to meet peer review have yet to be determined. Until one or more states goes through the official peer review process, sees what the peers deem to be adequate or incomplete information, and have received their decision letters from USED, the actual results are not known.

In 2017, the Florida legislature passed HB 7069 to contract for an independent study to determine whether the SAT and ACT may be administered in lieu of the grade 10 statewide, standardized ELA assessment and the Algebra 1 EOC assessment for Florida students consistent with federal requirements for precisely these reasons. Prior to revamping its state assessment system, the legislature wanted an independent assessment of whether such a system, allowing use of the FSAs, ACT, or SAT as chosen by each district, would provide comparable scores that are valid, reliable, and technically sound, and ultimately pass USED peer review requirements.

This overall study examined, in detail, six key criteria that a Florida assessment system allowing district choice of either the FSA in grade 10 ELA and Algebra 1 EOC or the ACT and SAT would need to satisfy in order to meet USED peer review requirements:

[7] The December 2016 regulations indicate that comparability between the locally selected test and the state test is expected at each academic achievement level.

- Alignment to Florida content standards for grade 10 ELA and Algebra 1 (Criteria 1 and 2)
- Comparability of assessment results (Criterion 3)
- Accommodations for students with disabilities and ELs (Criterion 4)
- Accountability (Criterion 5)
- Peer Review (Criterion 6)

It should be noted that this study only examined the likelihood of an assessment system that allows district the choice of the grade 10 FSA in ELA and Algebra 1 EOC, ACT and/or SAT would pass peer review. Determination of the likelihood that a state assessment system consisting of only the ACT or SAT for use in high school would pass peer review was beyond the scope of this paper.

The results of the study with respect to each of the six criteria are outlined in Table 6-2, below. As a preface to the results, Florida's use of an Algebra 1 EOC test, while perfectly acceptable in a state assessment system, creates some difficulties when determining the alignment and comparability of the assessment system to one using, at district choice, high school assessments such as the ACT and SAT and/or an Algebra 1 EOC test that is available in multiple grades. This is because Algebra 1 is a specific course with standards that are assessed at a deep level using an end-of-course assessment. This is in contrast with the ACT and SAT which are high school mathematics assessments that test far more standards than Algebra 1 and are, by definition, broader in scope.

Table 6-1 Summary of Findings from the Five Studies

Criteria	Summary Result Highlights ACT	Summary Result Highlights SAT
1–Alignment (Criteria 1 & 2)	**Math** • Insufficient number of items corresponding to a sufficient number of standards for all 3 reporting categories to address the breadth of MAFS, as measured by Range of Knowledge alignment criterion • Both test forms needed slight adjustment to meet minimum cutoffs for full alignment with Algebra 1 MAFS • 7 or 8 items required to be added for each test form analyzed to reach minimum cutoffs for full alignment with Algebra 1 MAFS • Only 1/3 of items correspond to Algebra 1 MAFS **ELA** • Insufficient number of items corresponding to a sufficient number of standards for one reporting category	**Math** • Insufficient number of items corresponding to a sufficient number of standards for 2 of 3 reporting categories to address the breadth of MAFS as measured by Range of Knowledge alignment criterion • 4 or 7 items required to be added for each test form analyzed to reach acceptable alignment with Algebra 1 MAFS • Only 2/3 of items correspond to Algebra 1 MAFS **ELA** • Insufficient number of items corresponding to a sufficient number of standards for one reporting category (RC3: Integration of Knowledge and Ideas) to address breadth of LAFS, as

	(RC3: Integration of Knowledge and Ideas) to address breadth of LAFS, as measured by Range of Knowledge alignment criterion • Both test forms weakly met or did not meet the DOK expected by the standards within Reporting Category 2 (Craft and Structure) and Reporting Category 4 (Language and Editing) • Both test forms needed major adjustment to meet minimum cutoffs for full alignment with Grade 10 LAFS • 12 or 17 items required to be added for each test form analyzed to reach minimum cutoffs for full alignment with Grade 10 LAFS	measured by Range of Knowledge alignment criterion • Both test forms did not meet the DOK expected by the standards within Reporting Category 4 (Language and Editing) • One test form was acceptably aligned and the other test form needed slight adjustment to meet minimum cutoffs for full alignment with Grade 10 LAFS • 5 or 7 items required to be added for each test form analyzed to reach minimum cutoffs for full alignment with Grade 10 LAFS

3–Comparability (Criterion 2)	• The data provided for this study are neither representative of the full population of students nor were the tests taken close enough in time to assume little to no learning occurred
	  ○ Only about 50% of tenth or eleventh grade students take either the ACT or the SAT prior to graduating high school (under-represents full sample of FSA test takers – which includes students only taking the FSAs which are, presumably, lower-scoring students)
	  ○ 83% of students took the Algebra 1 EOC test one to two years before taking the ACT or SAT. Large distances between time of testing in the two tests increases measurement error as learning likely occurred between those test administration times.
	• Several different statistical analyses were run to try to account for the data issues and determine if the test results are comparable.
	  ○ An important analysis was how often students would be placed at the same level of performance on the three tests. The results of this classification consistency analysis indicate that many students would be placed at different performance levels on the three tests, some by as much as four performance levels.
	• Thus, districts using the FSA option may have very different results than districts using either the ACT or SAT options. This casts serious doubt on the interchangeability of the three tests, and the soundness of making accountability decisions based on them.	

4–Accommodations (Criterion 3)	• With respect to provision of accommodations, the ACT could provide comparable benefit to the FSA for purposes of school accountability and graduation • Provides a greater number of accommodations to SWDs than the FSAs • Issue of a lack of transparency in the accommodations process about which the accommodation would result in a college reportable score. This would	• With respect to provision of accommodations, the SAT could provide comparable benefit to the FSA for purposes of school accountability and graduation ○ This was less evident for ELs • Provides a greater number of accommodations to SWDs than the FSAs • Issue of a lack of transparency in the accommodations process about which the accommodation would result in a college reportable score. This would

	likely result in non-comparable scores for some student groups	likely result in non-comparable scores for some student groups
5–Accountability (Criterion 4)	• Simulated schools were created to examine the effects of calculating school-level indicators using the different tests. • Overall, differences are shown across all three indicators. The results show that the numbers going into the accountability determination would differ for many schools by the test selected. • The differences shown for ELA vary by type of school. Larger schools with a greater number of lower performing students are advantaged by using the alternate tests (ACT or SAT). • There will often be very different students being compared in the growth models depending on which test is administered. • It does not appear to be fair to compare schools that use the state tests in their accountability system to those that use the alternate (ACT or SAT) tests.	
6–Peer Review (Criterion 5)	• 23 Critical Elements (CE) met peer review requirements • 1 Critical Element may not meet peer review requirements • 6 Critical Elements likely did not meet peer review requirements	• 20 Critical Elements met peer review requirements • 6 Critical Element may not meet peer review requirements • 3 Critical Elements likely did not meet peer review requirements • 1 Critical Element is N/A

Overall Conclusion

It is the opinion of ASG and its partners that due to the alignment, comparability, and accountability system issues associated with the ACT and SAT tests, allowing districts to pick which of the three tests to administer to its students is not appropriate, will not provide valid scores that are comparable, and, most likely, will not meet federal ESSA peer review requirements.