HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2017 No. 019

South Carolina Assessment Evaluation Report #1

Prepared for:	South Carolina Education Oversight Committee (EOC) 1205 Pendleton Street Room 502 Brown Building Columbia, SC 29201	**Prepared under:**	Contract Number: 4400014645
Authors:	Emily R. Dickinson Jing Chen Matthew Swain	**Date:**	April 24, 2017

South Carolina Assessment Evaluation Report #1

Table of Contents

List of Tables

List of Figures

South Carolina Assessment Evaluation Report #1

Executive Summary

The South Carolina Education Oversight Committee (EOC) contracted with the Human Resources Research Organization (HumRRO) to conduct a comprehensive evaluation of its state assessments. This is the first of three reports that will summarize that effort.

The EOC provides oversight of programs and expenditure of funds for the Education Accountability Act and the Education Improvement Act of 1984. As established in Section 59-6-10 of the South Carolina Code of Laws, the EOC's responsibilities include reviewing all assessments for approval as components of the state accountability system. As a part of this process, assessments are evaluated for alignment with the state standards, level of difficulty and validity, and for the ability to differentiate levels of achievement, and providing recommendations for change as needed. Based on these reviews, recommendations for change are made to the EOC, which in turns, shares the information with the State Board of Education, the South Carolina Department of Education (SCDE), the Governor, the Senate Education Committee, and the House Education and Public Works Committee. The SCDE will then report to the EOC on how it will address the recommendations. Then, the EOC will decide whether to approve the assessments for accountability purposes. HumRRO's comprehensive evaluation is intended to support the EOC in meeting these legislative mandates.

In order to meet federal accountability requirements, the SC Ready is administered annually to all public school students in grades 3-8 in the content areas of English Language Arts (ELA) and mathematics. The EOCEP is administered in ELA, mathematics, and science to all public school students by the third year of high school. HumRRO's evaluation includes the SC Ready for ELA and mathematics at all tested grade levels, as well as the EOCEP tests for Algebra 1, Biology 1, and English 1.

HumRRO's approach to the evaluation includes a series of separate but related tasks that focus on the key elements of assessment design and implementation. This report details methods and findings from the review of:

- item development processes for SC Ready and the EOCEP Algebra 1 test
- content alignment for the EOCEP Algebra 1 test
- test construction for SC Ready and the EOCEP Algebra 1 test.

Based on the results from these three tasks, we found that the SC Ready ELA/math and EOCEP Algebra 1 tests generally adhere to industry best practices, with some areas noted for improvement. HumRRO offers several findings and interim recommendations for each assessment reviewed. Each interim recommendation is accompanied by a priority rating using the following classification schema:

Priority Rating	Description of Priority Rating
Urgent	Definitely needs to be addressed; should be considered and addressed immediately.
High	Needs to be addressed; should be considered and addressed as soon as possible.
Medium	Should be considered and possibly addressed.
Low	Might be considered if time allows.

Subsequent reports will address additional and related aspects of test development and implementation, building toward a more complete understanding of the quality of the South Carolina assessments. HumRRO will provide final recommendations to the EOC in the third and final evaluation report.

Algebra 1

Item Development (Chapter 2)

Finding 1.1. The processes used to develop items for the EOCEP Algebra 1 tests adhere to industry best practices. Items undergo a multi-step process that includes review by expert judges regarding content and cognitive complexity alignment, as well as sensitivity and fairness.

Finding 1.2. Universal design principles are referenced, but different documents provide different details on how to fulfill these principles. Inconsistency and lack of detail was found in the presentation of check points (specific points of guidance for item developers) across documents, with missing check points to address the accessibility needs of all students. We did not see documents that clearly describe how empirical results and expert judgements are appropriately used to review items and scoring guides. It is difficult to judge whether empirical results and expert judgements are appropriately used when reviewing items and scoring criteria.

Finding 1.3. Documentation about the item management system (IDEAS) was not found. No documentation was provided on how items are stored, how item review feedback is saved, or how changes are tracked in the system. Currently, preliminary item information is only obtained from field testing.

Finding 1.4. Item development documentation does not clearly specify the intended uses of the test scores.

> **Interim Recommendation 1.1. Improve item development processes (High).** Item development processes could be improved in several ways. Aspects of the item development process to improve include expanded background information for item developers/reviewers on the goals of the assessment for which items are developed, and expanded item review checklists with clear guidance for evaluating item content, difficulty, clarity, and accuracy. Record keeping of the item development process should also be uniformly implemented and consistently documented. Cross-referencing should be added to item development documents to ensure easy access to all available information. Processes and documentation should clearly and consistently implement universal design principles. More detailed information about the background and characteristics of expert judges and quality assurance staff should be captured and documented.
>
> **Interim Recommendation 1.2. Continue to expand the available documentation describing processes and procedures for item development (High).** Standard 7.4 of *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Joint Standards*) highlights the importance of detailed documentation of all test development procedures. We found several areas where detailed descriptions were lacking in the available documentation, or where no formal documentation was available. There were also instances where inconsistent guidance was provided across documents. Although we were able to clarify our understanding through web searches and phone interviews with relevant staff, the assessment system could be improved through

continued expansion of the formal documentation that is available. We recommend that the Data Recognition Corporation (DRC) compile a technical manual that documents all aspects of item development.

Interim Recommendation 1.3. Consider adding item tryouts or cognitive labs to the item development process (Medium). Item tryouts, which use a smaller number of students than field testing, and which occur earlier in item development when changes can be made more easily, should be considered for subsequent item development. This would be particularly useful for developing novel item types.

Content Alignment (Chapter 3)

A content alignment study was conducted on two EOCEP Algebra 1 test forms (Spring 2017 and Winter 2016-17) to investigate how well the items align to the SCCCRS. Independent, external content experts served as panelists for this alignment workshop. The findings and recommendations follow.

Finding 2.1. Overall, the alignment results provide support for the content validity of the EOCEP Algebra 1 test. On average, panelists rated approximately 90% of the items as "fully aligned" to the SCCRS. We also investigated alignment using a modified Webb alignment methodology (1997, 1999, 2005). The Webb alignment criteria were investigated at the level of the content strand and at the level of the key concept. There was one Webb alignment criterion (categorical concurrence) that received a "partially aligned" rating at the content strand level and a "weakly aligned" rating at the key concept level on the Webb rating scale; *however*, the categorical concurrence criterion is intended to inform the minimum number of items required for each reporting category. Because SCDE does not report scores at the level of the content strand or at the level of the key concept, the lower alignment ratings on Webb's categorical concurrence criterion should be of no concern for the SCDE. The EOCEP Algebra I test meets the remaining Webb criteria for appropriate item difficulty (depth-of-knowledge) and coverage of the standards (range-of-knowledge correspondence and balance-of-knowledge representation). Finally, at the end of the workshop panelists were asked to provide a final holistic rating of the overall alignment between the EOCEP Algebra 1 test and the SCCCRS. Four of the five panelists (including the nationally recognized content expert) rated the overall alignment as "good."

Finding 2.2. As indicated in Finding 2.1, Webb's depth-of-knowledge consistency criterion was attained per the Webb rating scale. The depth-of-knowledge consistency criterion indicates whether there is consistency between the complexity of knowledge required by the standards and the complexity of knowledge required to correctly answer the items linked to those standards. Webb's suggested minimum for this criterion is that at least 50% of the items should have complexity ratings at or above the level of the corresponding standard. All the content strands meet this alignment criterion. At the level of the key concept, one of the four key concepts—Structure and Expressions—fell considerably short of meeting this criterion for both the Spring 2017 form and the Winter 2016-17 form. This finding suggests that the cognitive complexity required to correctly answer the items linked to the standards within this key concept is, on average, lower than the cognitive complexity required by the standards.

Finding 2.3. In addition to the Webb alignment criteria, we also compared the mean number of items linked to each content strand by the expert panelists to the target number of items in the test blueprint for each content strand. The mean number of items linked to each content strand was within the range specified in the test blueprint for all content strands, except for the Number and Quantity content strand, for which the mean number of linked items was 4.8 (*SD = 0.45)*

and 4.6 (*SD* = 0.55) for the spring and winter forms, respectively, which was slightly below the target of 5 – 9 items specified in the test blueprint.

Finding 2.4. The independent, external reviewers found an overwhelming majority of Algebra 1 items to be free of any issues related to clarity, accuracy, grade-level appropriateness, and biased content/presentation. There were only two items on which at least three of the five panelists expressed concerns about the items' clarity. All panelists' comments on items have been provided to DRC, separately from this report[1], for their consideration.

Finding 2.5. Three of the five panelists mentioned the limitations of multiple-choice tests such as the EOCEP Algebra 1 test for providing useful information about the South Carolina College- and Career-Ready Mathematical Process Standards, or to support research-based instruction. Four of the five panelists also mentioned that some items might be biased towards students with access to, and familiarity with, graphing calculators, though one panelist stated that this is common to most math tests.

Interim Recommendation 2.1. Monitor the cognitive complexity of the items intended to measure the Building Functions key concept (Medium). Consider enhancing the cognitive complexity required to answer the items intended to measure the Structure and Expressions key concept to ensure that there is consistency between the level of cognitive complexity required by the standards that comprise this key concept and the cognitive complexity required to correctly answer the items that measure this key concept. If any reporting were to be considered at the key concept level, this recommendation would become a higher priority.

Interim Recommendation 2.2. Continue to monitor the content representativeness of the item pool (Medium). All test items are linked to a content standard, and evidence from the alignment study indicates appropriate numbers of items for all content strands, with the possible exception of the Number and Quantity content strand. The SCDE may want to consider including an additional item or two to the measure the Number and Quantity content strand to ensure that the EOCEP Algebra 1 test is meeting the intent of the test blueprint. Also, should changes be made to reporting practices (e.g., reporting subscores), ongoing monitoring of the content standard(s) measured by items will help to ensure that there are sufficient numbers of items for such purposes.

Interim Recommendation 2.3. Consider including additional item types to the Algebra 1 test (Low). Item types other than traditional multiple choice would offer more opportunities for students to demonstrate, for example, relating problems to prior knowledge and identifying multiple paths to a solution. Such opportunities may better reflect the South Carolina College- and Career-Ready Mathematical Process Standards while also better supporting research-based instruction.

Test Construction (Chapter 4)

Finding 3.1. The processes and procedures for creating EOCEP Algebra 1 test forms generally reflect industry best practices as outlined in the *Joint Standards*.

Finding 3.2. Available documentation guiding test construction processes and procedures contains several gaps. For example, there is no mention of internal consistency reliability

[f]f

minimums or if this is considered when creating forms. The origin of item statistics used for test form construction (e.g., estimated during field testing or prior operational use) is not clearly stated, nor is it clear at what stage differential item functioning (DIF) is analyzed. Documentation also appears inconsistent regarding the use of classical test theory (CTT) and/or item response theory (IRT) statistics for forms assembly.

Finding 3.3. Item P-Values and point-biserial correlations associated with Algebra 1 forms administered in 2015-16 are within acceptable ranges. However, within the item bank, approximately 5% of items have P-Values below .2, and a small number of items have negative point-biserial correlations.

Interim Recommendation 3.1. Remove items with P-Values and/or point-biserial correlations outside of the acceptable ranges from the item bank (Urgent). Though item statistics are considered during form construction and previous operational test forms only contained items with CTT statistics within the acceptable ranges, removal of problematic items from the item bank would provide an extra quality assurance step. It would also provide a more accurate depiction of the strength of the available item pool and inform item development.

Interim Recommendation 3.2. Continue to expand the available documentation describing processes and procedures for test form construction (High). The content considerations of the test need to be more explicitly defined (e.g., paper/pencil vs computerized administration, procedures for replacing technology-enhanced items on a paper/pencil test). The conditions of administration need to be more clearly specified (time for testing, directions, administration guidelines), and the statistical targets for test development (test length, internal consistency reliability, target P-Values, target point-biserial correlations) need to be better specified. Specifically, we recommend a range of P-Values and a minimum point-biserial correlation be specified. We recommend that DRC compile a technical manual that documents all aspects of test construction, including evidence of all studies to investigate potential sources of construct irrelevant variance.

SC Ready

Item Development (Chapter 2)

Finding 4.1. The processes used to develop items for the SC Ready ELA/math tests adhere to industry best practices. Items undergo a multi-step process that includes review by expert judges regarding content and cognitive complexity alignment, as well as sensitivity and fairness.

Finding 4.2. Universal design principles are referenced, but different documents provide different details on how to fulfill these principles. Inconsistency and lack of detail was found in the presentation of check points (specific points of guidance for item developers) across documents, with missing check points to address the accessibility needs of all students. However, we did not see documents that clearly describe how empirical results and expert judgements are appropriately used to review items and scoring guides. It is difficult to judge whether empirical results and expert judgements are appropriately used when reviewing items and scoring criteria.

Finding 4.3. Documentation about the item management system (IDEAS) was not found. No documentation was provided on how items are stored, how item review feedback is saved, or

how changes are tracked in the system. Currently, preliminary item information is only obtained from field testing.

Finding 4.4. HumRRO's evaluation of a sample of items found that items generally adhered to item quality guidelines and various review feedback was incorporated to improve the quality of the items. However, we find readability and grade level appropriateness are specifically considered for the reading passages and related item stimuli as indicated in document 13, but not for math items.

Finding 4.5. Students' responses from field tests are used to refine the scoring rubrics for text dependent analysis writing prompts on the SC Ready ELA assessment. However, it is not clear how empirical data are used to review and improve scoring criteria (e.g., refine scoring guides, build training sets).

Interim Recommendation 4.1. Improve item development processes (High). Item development processes could be improved in several ways. Aspects of the item development process to improve include expanded background information for item developers/reviewers on the goals of the assessment for which items are developed, and expanded item review checklists with clear guidance for evaluating item content, difficulty, clarity, and accuracy. Record keeping of the item development process should also be uniformly implemented and consistently documented. Cross-referencing should be added to item development documents to ensure easy access to all available information. Processes and documentation should clearly and consistently implement universal design principles. More detailed information about the background and characteristics of expert judges and quality assurance staff should be captured and documented.

Interim Recommendation 4.2. Continue to expand the available documentation describing processes and procedures for item development (High). Standard 7.4 of the *Joint Standards* highlights the importance of detailed documentation of all test development procedures. We found several areas where detailed descriptions were lacking in the available documentation, or where no formal documentation was available. There were also instances where inconsistent guidance was provided across documents. Although we were able to clarify our understanding through web searches and phone interviews with relevant staff, the assessment system could be improved through continued expansion of the formal documentation that is available. We recommend that DRC compile a technical manual that documents all aspects of item development.

Interim Recommendation 4.3. Incorporate readability and grade-level appropriateness reviews for mathematics items and associated stimuli (High). The reading demand of the math items and associated stimuli may introduce construct irrelevant variance and affect students' performance. Adding these reviews during item development would further support the validity of test scores.

Interim Recommendation 4.4. Consider adding item tryouts or cognitive labs to the item development process (Medium). Item tryouts, which use a smaller number of students than field testing, and which occur earlier in item development when changes can be made more easily, should be considered for subsequent item development. This would be particularly useful for developing novel item types.

Test Construction (Chapter 4)

Finding 5.1. The processes and procedures for creating test forms generally reflect industry best practices as outlined in the *Joint Standards*.

Finding 5.2. Available documentation guiding test construction processes and procedures contains some gaps. For example, we found no guidelines surrounding issues of item parameter drift or non-convergence that might occur during the post-equating process. We also found no description of how comparisons between paper and computer-based item-level data are conducted, nor mention of forms-level comparisons between paper and computer forms.

Finding 5.3. Item statistics from the item bank demonstrate improvements in the available item pool over time. Items with statistics outside of the acceptable ranges were removed between 2016 and 2017.

> **Interim Recommendation 5.1. Continue to expand the available documentation describing processes and procedures for test form construction (High)**. Documentation should be expanded to ensure complete information is available for understanding how issues such as item parameter drift and non-convergence are evaluated and addressed. We recommend that DRC compile a technical manual that documents all aspects of test construction.
>
> **Interim Recommendation 5.2. Consider continuing the analysis of mode DIF and expand the available documentation describing these procedures (Medium).** Although there is a movement toward near universal online test administration, if there are paper forms administered then the analysis of any differences between paper and online forms should be conducted. Any such analyses should be described in detail in the technical documentation.

South Carolina Assessment Evaluation Report #1

Chapter 1: Introduction

The South Carolina Education Oversight Committee (EOC) contracted with the Human Resources Research Organization (HumRRO) to conduct a comprehensive evaluation of its state assessments. This is the first of three reports that will summarize that effort.

The EOC provides oversight of programs and expenditure of funds for the Education Accountability Act and the Education Improvement Act of 1984. As established in Section 59-6-10 of the South Carolina Code of Laws, the EOC's responsibilities include reviewing all assessments for approval as components of the state accountability system. As a part of this process, assessments are evaluated for alignment with the state standards, level of difficulty and validity, and for the ability to differentiate levels of achievement, and providing recommendations for change as needed. Based on these reviews, recommendations for change are made to the EOC, which in turns, shares the information with the State Board of Education, the South Carolina Department of Education (SCDE), the Governor, the Senate Education Committee, and the House Education and Public Works Committee. The SCDE will then report to the EOC on how it will address the recommendations. Then, the EOC will decide whether to approve the assessments for accountability purposes. HumRRO's comprehensive evaluation is intended to support the EOC in meeting these legislative mandates.

The state assessment program includes the South Carolina College-and Career-Ready (SC Ready) assessments for grades 3-8 and the End of Course Examination Program (EOCEP) for high school. Data Recognition Corporation (DRC) works in coordination with SCDE to develop, administer, and score the tests.

In order to meet federal accountability requirements, the SC Ready is administered annually to all public school students in grades 3-8 in the content areas of English Language Arts (ELA) and mathematics. The EOCEP is administered in ELA, mathematics, and science to all public school students by the third year of high school. HumRRO's evaluation includes the SC Ready for ELA and mathematics at all tested grade levels, as well as the EOCEP tests for Algebra 1, Biology 1, and English 1.

HumRRO's approach to the evaluation includes a series of separate but related tasks that focus on the key elements of assessment design and implementation. Specifically, HumRRO identified 7 tasks corresponding to the general requirements outlined in the Request for Proposals (RFP). These tasks include:

- Task 1: Review of Item Development Processes
- Task 2: Review of Items to Standards Alignment and Item Quality
- Task 3: Review of Test Construction Processes
- Task 4: Review of Test Administration Procedures
- Task 5: Review of Scaling, Equating, and Scoring Processes
- Task 6: Review of Psychometric Processing and Item Parameters
- Task 7: Review of Minimum Legal Requirements of SC Ready

Each of the above tasks will be completed for the SC Ready 3-8 ELA and mathematics tests, and the EOCEP tests for Algebra 1, Biology 1, and English 1, with one exception. Task 7 will be completed for the SC Ready tests only. To accomplish the above tasks, HumRRO coordinates with DRC to obtain the necessary documentation and data to conduct these tasks.

Per the requirements outlined in the RFP, the seven tasks will be completed in a staggered fashion and the results will be presented over a series of three reports. This report includes (a) Task 1 for SC Ready and the EOCEP Algebra 1 test, (b) Task 2 for the EOCEP Algebra 1 test, and (c) Task 3 for SC Ready and the EOCEP Algebra 1 test. HumRRO's second report will include (a) Task 1 for the EOCEP Biology and English 1 tests, (b) Task 2 for SC Ready and the EOCEP Biology and English 1 tests, (c) Task 3 for the EOCEP Biology and English 1 tests, (d) Task 4 for SC Ready and the EOCEP Algebra 1 and Biology tests, (e) Task 5 for SC Ready and the EOCEP Algebra 1, Biology and English 1 tests, (f) Task 6 SC Ready and the EOCEP Algebra 1, Biology and English 1 tests, and (g) Task 7 for SC Ready. HumRRO's third and final report will include Tasks 4-6 for the EOCEP English 1 test, along with final recommendations for the South Carolina Assessment System. Table 1.1 summarizes the tasks and tests that will be included in each report.

Table 1.1 Tasks and Tests Included in each HumRRO Report

Task	Reports			
	SC Ready	EOCEP Algebra 1	EOCEP English 1	EOCEP Biology 1
1 – Item Development	1	1	2	2
2 – Items to Standards	2	1	2	2
3 – Test Construction	1	1	2	2
4 – Test Administration	2	2	3	2
5 - Scaling, Equating, Scoring	2	2	2/3	2
6 – Psychometric, Item parameters	2	2	2/3	2
7 – Minimum Legal Requirements	2	--	--	--
Final	3	3	3	3

Chapters 2-4 of this report summarize the methods and results from Tasks 1-3, respectively. Chapter 5 discusses findings across the three tasks and provides interim recommendations.

Chapter 2: Review of Algebra 1 and SC Ready Item Development Processes

Introduction

HumRRO conducted an evaluation of the item development processes for the EOCEP Algebra 1 test as well as the SC Ready mathematics and English Language Arts (ELA) assessments. The purpose of this evaluation was to document the extent to which best practices are employed during item development to ensure the development of high-quality test items. It is worth noting that our evaluation was focused on the processes and procedures for initial item development and review and is therefore qualitative in nature. Subsequent HumRRO reports will include additional tasks that focus on item-level statistics and other quantitative data to further inform the quality of test items.

Methods

Our evaluation was conducted in three phases. First, we reviewed all available relevant documents and evaluated the processes described in these documents based on industry standards. Second, we conducted an interview with relevant staff from DRC and SCDE to ask clarifying questions and collect additional detailed information. Third, we collected and reviewed a set of sample items to see how individual items were developed, modified or dropped during the process. This helped us understand the implementation of procedures within the processes. The methods used in each phase are described in more detail below.

Phase I. Document Review

We worked in cooperation with the EOC, SCDE and DRC to obtain documentation of the South Carolina item development processes for each test. We also searched the SCDE website to identify relevant information. The documents we collected fall into several categories based on their foci. Table 2.1 lists all the documents we collected and reviewed. These documents provided useful information about various steps and procedures within the item development processes.

Table 2.1. Documents Reviewed for Task 1 – Item Development

Document Focus	Document File Name	Assessment(s) that the file applies to or comes from		
		EOCEP Algebra 1	SC Ready Math	SC Ready ELA
Flowchart of Item Development Process	Document 1: 001_Item Development Process_RE.pdf (document 1)	X	X	X
Item Review Checklist	Document 2: 002_Item Review Checklist_RE.pdf	X	X	X
Item Review Process	Document 3: 003_Item Review Process_E.pdf	X		
	Document 14: 014_ Item Review Process_R.pdf		X	X
Item Writer Training Materials	Document 7: 007_Training manuals for item developers_RE.pdf	X	X	X
	Documents 7F: 007F_Item Writer Training	X	X	X
Quality Assurance Procedures	Document 8: 008_Quality Assurance Procedures for Item Development_RE.pdf	X	X	X
Guidelines for Selecting/Developing Passages and Other Item Stimuli	Document 13: 013_Guidelines for Selecting developing passages and stimuli_RE.pdf	X		X
Test Administration Manual	018_ Spring 2017 SC READY Test Administration Manual.pdf		X	X
	019_Spring 2017 EOCEP Test Administration Manual.pdf	X		
Sample Item Full Development Documentation	1_Item Development Documentation ALGEBRA_E.pdf	X		
	2_Item Development Documentation ALGEBRA_E.pdf	X		
	3_Item Development Documentation ALGEBRA_E.pdf	X		
	1_Item Development documentation_G3_R.pdf			X
	2_Item Development documentation_G4_R.pdf		X	X
	3_Item Development documentation_G5_R.pdf		X	X
	4_Item Development documentation_G6_R.pdf		X	X
	5_Item Development documentation_G7_R.pdf		X	X
	6_Item Development documentation_G8_R.pdf		X	X
	1_Item development documentation_G3_1_R.pdf		X	
	2_Item Development documentation_G3_2_R.pdf		X	

This evaluation of the item development processes and resulting test items was informed by industry best practices as outlined in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Joint Standards*). Two HumRRO researchers identified the standards that were relevant to the item development processes from the *Joint Standards*. The relevant *Joint Standards* identified by each researcher were compared and discussed to reach final consensus on the selected *Joint Standards*. Four *Joint Standards* were agreed to be directly relevant to the scope of item development processes included in this task. These *Joint Standards* can be found in the results section of Chapter 2.

In addition to identifying relevant *Joint Standards*, we developed a rating scale to evaluate the item development processes per each standard. The rating scale ranges from a score of 1 to 5, increasing with level of compliance with the standard. The rating scale is presented in Table 2.2. For each of the identified *Joint Standards*, we assign an overall rating based on information collected from all three phases.

Table 2.2. Rating Scale for Relevant Joint Standards

Score Level	Description of Score Level
1	No evidence of the Standard found in the documents/interview.
2	Little evidence of the Standard found in the documents/interview; less than half of the Standard covered in the documents/interview and/or evidence of key aspects of the Standard could not be found.
3	Some evidence of the Standard found in the documents/interview; approximately half of the Standard covered in the documents/interview, including some key aspects of the Standard.
4	Evidence in the documents/interview mostly covers the Standard; more than half of the Standard covered in the documents/interview, including key aspects of the Standard.
5	Evidence in the documents/interview fully covers all aspects of the Standard.

Phase II. Interview with Key Staff

We conducted a phone interview with key staff from DRC and SCDE to ask clarifying questions and collect in-depth information. The interview allowed us to gain better understanding of the item development and review processes and identify any potential issues related to the established processes. The interview questions are listed in Appendix A. Two HumRRO researchers conducted the interview. One was responsible for asking questions and the other was responsible for taking detailed notes. The interview lasted approximately one hour. Approximately ten staff from SCDE and DRC participated in the interview, including staff from test development, psychometrics and program administration.

Phase III. Item Review

Finally, we conducted a targeted review of a sample of items from each test. The purpose of the review was to track a sample of items from initial draft through the item develop process to see how they were either modified or dropped for operational use. To do this, we collected item cards[2] for sample items. The item cards include each iteration of an item through the development process with reviewer comments. These provided concrete examples that illustrated the item review and revising procedures. We requested and reviewed all available

[2] Item cards capture each iteration of an item during development and revision including reviewer comments. In addition, the item cards identify the standard and sub-standard targeted and a conceptual level of item difficulty (easy, moderate, or hard).

documentation for a representative sample of items. These items were selected for each relevant content area and grade span, representing a range of content standards, item types, and item difficulties. Sixteen items were selected that included three items from the EOCEP Algebra 1 test, six items from the SRC Ready Reading assessment and seven items from the SC Ready math assessment, with approximately one item at each grade level.

As mentioned previously, we did not collect empirical item level statistics specific to each assessment. The information we collected through the three phases described above indicates that the item development processes are generally the same for the EOCEP Algebra 1 test and the SC Ready assessments. Furthermore, the results from our evaluation do not differ substantively across these assessments. Consequently, our results are presented across the assessments.

Results

Results are organized around the relevant *Joint Standards* and include details from our process documentation review, interviews with key item development staff, and targeted item review to support judgments about the extent to which industry standards are met. Table 2.3 provides an overall rating for each relevant *Joint Standard* after reviewing all available information from each assessment using the scale in Table 2.2.

Table 2.3. Evaluation Results Based on the Joint Standards

Standard Number	Standard Content	Rating
Standard 3.2	Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.	4
Standard 3.3	Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.	4
Standard 4.0	Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.	3
Standard 4.8	The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	4

The following section is organized by the *Joint Standards* used in our evaluation. For each standard, we describe the rationales of our rating and explain to what extent the standard is met. We also provide suggestions for improvement to better align with the standard.

Standard 3.2 – Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

Evidence from the documents and the interview suggests that key aspects of Standard 3.2 are covered. Responses from the interview indicate that each item goes through five rounds of review in the item development processes to ensure the item measures the intended construct. During the first and second rounds of review, item editors and content leaders examine alignment for each item to see if the item represents intended standard(s). They review how well the cognitive rigor and complexity level of each item represent test specifications. The third round of review addresses editorial issues. The fourth round of review is an external review focusing on bias and sensitivity issues. The last round of review is usually performed by SCDE staff to ensure alignment with South Carolina standards and style. Thus, each item goes through multiple rounds of reviews to ensure the quality of the item.

In the item development documentation that we received for the sample items, information such as the standard(s) that the item addressed, the depth of knowledge (DOK), and the estimated item difficulty were included. These provide evidence that each item is designed to measure the intended construct(s) at the intended difficulty level. However, it is possible that even when items are carefully designed, there might still be gaps between the items and the standards to be covered. Alignment study results presented in Chapter 3 provide additional information about how well items measure the intended standards/sub-standards and represent the same DOK level as the standards.

Evidence from relevant documents collected in the Phase I document review and item cards collected in Phase III provide evidence that items are carefully reviewed and edited to minimize the potential for tests to be affected by construct irrelevant characteristics. For example, the item review checklist includes check points on the linguistic, communicative, cognitive, and other important characteristics of the items. In the item review process files, the authors presented a multi-aspect review process that includes reviews on content alignment, rigor-level alignment, technical design, universal design, and bias/fairness/sensitivity issues. These review processes are helpful to minimize construct irrelevant variance for the items.

However, in both the item review checklist file and the item review process files (documents 2, 3, and 14), some review check points are described vaguely without concrete criteria. It is difficult for item reviewers, especially less experienced ones, to make good judgements about the extent to which guidelines have been followed without more detailed descriptions of what is expected. Examples of check points with ambiguity are provided below:

- Are the content expectations appropriate?
- Is the difficulty appropriate?
- Are supporting graphics necessary, appropriate, and clear?
- Are all distractors plausible and logical?

We also find that the item review guidelines and checklists vary in their comprehensiveness across documents. For instance, document 2 only provides a very brief item review checklist. However, one of the item writer training files (document 7F) provides a very detailed content review checklist. It may be worth adding references to detailed guidelines and checkpoints in all documents so that item writers or reviewers can use all available information to review items and check the quality of items from all possible aspects.

Standard 3.3 – Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test.

Evidence from the documents and the interview indicates that key aspects of Standard 3.3 are covered. As described in the item review checklist file and the item review process files, all items are reviewed for bias, fairness and sensitivity. In document 7F, detailed guidelines for reviewing bias, fairness, and sensitivity issues are described. The authors provide definitions of bias and sensitivity, introduce different types of bias, and describe topics to avoid, topics of concern, and special circumstances. Sample items with bias, fairness, and/or sensitivity concerns are provided to support the training of bias and sensitivity review. All these helped to improve the quality of items and ensure the reliability and validity of the assessment for examinee subgroups.

Furthermore, based on responses from the interview, SCDE recruited an external review committee to review issues related to bias, fairness and sensitivity within the item development processes. Currently, the review committee has 33 members from different organizations (e.g., the African American studies program of university of South Carolina, the psychology department of university of South Carolina) with diverse experience and academic backgrounds. The committee receives trainings at the beginning of the review sessions and reviews the items with a focus on bias, fairness and sensitivity issues. All these practices suggest the item development processes generally met the second standard. Relevant subgroups are considered during the item development and review processes.

Accessibility issues are addressed to some extent during the item development processes. Accommodations for students with disabilities are provided in both SC Ready and EOCEP assessments. Customized formats (e.g., braille, large-print, loose-leaf) are available for students with documented disabilities. In the quality assurance file, the authors describe the implementation of universal design principals as a way to improve examinees' participation of the assessment. It is described that all item developers, editors, graphic artists, and publications experts are trained in applying universal design principles.

However, we think there is room for improving the accessibility of items since some of the item review checklists (e.g., document 2) do not include checkpoints on accessibility related issues. We also find some inconsistencies between the documents regarding DRC's practices to follow the universal design principles. In the quality assurance file, the authors list five current item writing and editing practices to comply with the universal design principles (see document 8). However, in one of the item writer training files (Making Assessments Accessible and Inclusive), the authors include a much more comprehensive list of actions that should be taken to follow universal design principles and guidelines. Because of the inconsistencies between the documents, we are not sure of the current practices that DRC takes to ensure the accessibility of individual items.

Standard 4.0 – Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

Information from the documents and the interview provides some evidence of the Standard 4.0 being met. Approximately half of the standard is covered. In all the documents that we received

to date, the test developers do not clearly describe the intended uses of the test scores for the intended examinees. How items are developed to support the intended uses of the test scores for individuals in the intended examinee population is not clear (e.g., whether the test is designed for summative purposes at the school or district-level or for individual students).[3] Though the documented item development processes generally provide some evidence of test fairness, reliability and validity, the connection between the item development processes and the goal of the assessment can be strengthened to better align with the standard.

The test developers documented some steps taken during the item development processes to provide evidence of fairness, reliability, and validity. For example, the overall item development process is documented in the flowchart presented in document 1. The item review guidelines and checklists are documented in documents 2, 3, and 14. Item writer selection procedures, training activities, and item writer training materials are documented in documents 7 and 7F. Quality assurance procedures are documented in document 8. All the documented processes and procedures provide some evidence that the SC Ready and EOCEP tests are designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. However, more details and additional information need to be added to the current documents, and additional documents need to be developed, to complete the documentation of the item development processes. Without more detailed information, it is unclear how each step is conducted to control the quality of items. For example, documents 2 and 3 only include item review guidelines and checklists. Other item review procedures such as how item review feedback is tracked and used to improve items are not clearly documented.

Standard 4.8 – The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

Evidence from the documents and the interview suggests that key aspects of Standard 4.8 are met, though processes and documentation can be improved to address this standard better. Empirical results and expert judgements are used to review items. However, it's unclear whether empirical analyses and expert judgements are most appropriately used. The test developers documented the item writer selection criteria, training activities, and training materials, but we are not sure to what extent SCDE and DRC document experts' or item writers' qualifications, relevant experiences, and demographic characteristics.

Results from empirical analyses are used to review and select items. For both SC Ready and EOCEP assessments, a set of psychometric guidelines is used for selecting items. These guidelines include the recommended ranges for P-Values, item-total correlations, and differential item functioning (DIF) values. However, as discussed in Chapter 4, it is not clearly documented whether empirical analyses are conducted using the field test data, operational data, or a combination of both to calculate the item statistics for item selection and review purposes.

Empirical results are also used to review scoring criteria. The only constructed response (CR) items are the text dependent analysis writing prompts in the SC Ready ELA assessment. For these items, students read a piece of text or passage and draw upon that text for their extended

[3] Information about the intended uses of test scores is available on the SCDE website. However, this information should be made more explicit in item development documentation.

written responses (e.g., support their responses with evidence from the text). Responses from the interview suggest that students' responses from field tests are used to refine the scoring rubrics for these items. However, we did not receive documentation that describes how empirical data are used to review and improve scoring criteria (e.g., refine scoring guides, build training sets). It is difficult for us to judge whether empirical results are appropriately used when reviewing item scoring criteria.

Expert judges are used when conducting the bias, fairness and sensitivity reviews. But it's also not clear to what extent DRC and SCDE document the qualification, relevant experiences, and demographic characteristics of item writers or expert judges. Document 7 only lists the qualifications used to select item writers. Similarly, in the quality assurance (QA) file, the experience and qualifications of staff that perform QA procedures at different levels are only briefly described. The existing documentation should be expanded to include more detailed information on the background and characteristics of expert judges and other QA staff.

Discussion

We conducted an evaluation of DRC's item development processes for SC Ready and EOCEP Algebra 1 assessments. Our evaluation is based on available documentation and conversation with key item development staff from SCDE and DRC. It is likely that we did not capture all important information within the item development processes. We plan to conduct at least one site visit to observe one or more processes related to item development prior to our second report that will further explore item development processes relative to the EOCEP Biology and English 1 tests.

We generally found the processes used to develop items for the SC Ready ELA/math and EOCEP Algebra 1 tests adhere to industry best practices. We found that that some or all of the key aspects of the relevant *Joint Standards* are met. Items undergo a multi-step process that includes reviews by expert judges regarding content and cognitive complexity alignment, as well as sensitivity and fairness.

The documentation of item development processes could be improved to provide more validity evidence for the assessments. The test developer documented some steps taken during the item development processes, but not others. For example, we did not find documentation about the item management system (IDEAS). There is no documentation on how (a) items are stored, (b) item review feedback is saved, and (c) changes are tracked in the system. We did not see documents that clearly describe how empirical results and expert judgements are used to review items and scoring guides. The lack of documentation around development does not necessarily imply that it did not follow a rigorous process of development and review. Rather, it is not feasible to review the processes. It's also possible that there are existing documents but we were not able to get these documents for some reason. We recommend DRC compile a technical manual that documents all aspects of item development and test construction processes.

For the documents that we received and reviewed, oftentimes, more details and additional information need to be added. Without detailed information, it's unclear how each step is conducted to control the quality of items. We find some inconsistencies or variations among documents which should be modified to clarify procedures. For example, the item review guidelines and checklists vary with respect to their comprehensiveness across documents. The description about the actions that should be taken to comply with universal design principles and guidelines also varies across the QA file and the item writer training file.

Besides documentation, we find that item development processes can be further improved with the addition of an item tryout step. Item tryouts and cognitive labs can be useful in early design to ensure an item is comprehensible to students, assessing what is intended, and solution strategies are consistent with intended purposes. Because item tryouts only collect data from a small sample of students, it is a quicker and less expensive way to identify problems with the items, as opposed to the field tests that typically collect 2500-3000 responses per item (according to the responses from interview).

HumRRO's evaluation of the sample items found that items generally adhered to item quality guidelines and various review feedback was incorporated to improve the quality of the items. However, we find readability and grade level appropriateness are specifically considered for the reading passages and related item stimuli as indicated in document 13, but not for math items. The reading demand of the math items may introduce construct irrelevant variance and affect students' performance on the items. It is important to consider readability and grade level appropriateness not only for ELA assessments, but also for math assessments.

Chapter 3: Review of Algebra 1 End-of-Course Examination Content Alignment

Introduction

Alignment studies address a vital question related to the validity of test scores: "Does the test content adequately reflect the content knowledge and skills that students are expected to learn as outlined in the state standards?" School curriculum must be designed to meet the goals specified by the state standards and consequently assessments should measure the same content. South Carolina's Code of Laws mandates the review of end-of-course assessments for alignment with the state standards.

HumRRO conducted an alignment workshop during which a panel of educators and content experts reviewed the South Carolina College- and Career-Ready Standards (SCCCRS) and a sample of items from the End-of-Course Examination Program (EOCEP) Algebra 1 test to evaluate the extent to which students' test scores reflect content knowledge and skills at the breadth and depth outlined in the content domain. This chapter describes the alignment methods and results, along with discussion of the overall alignment of the EOCEP Algebra 1 test to the SCCCRS.

Methods

Several methods of alignment are in current use (e.g., Forte, 2017; Porter, 2002; Webb, 1997, 1999, 2005). These methods involve panelists subjectively evaluating several aspects of the content standards and test items. The data from panelists' evaluations are analyzed statistically to determine the extent of alignment. HumRRO modified the method developed by Norman Webb to evaluate the alignment of the EOCEP Algebra 1 test to the SCCCRS for Algebra 1. Webb's alignment methodology is the most widely used in the United States.

Webb Alignment Method

The Webb alignment method (1997; 1999; 2005) was originally designed to align content standards with large-scale assessments. Dr. Norman Webb has researched and refined this method over time, and his approach is supported by the Council of Chief State School Officers (CCSSO).[4]

The Webb method includes four major criteria to evaluate alignment. These criteria link with statistical procedures used to assess how well items on the assessment, regardless of item type and point value, and the state's standards document actually match. The four alignment criteria are: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance-of-knowledge representation.

Categorical concurrence is a basic measure of alignment between content standards and test items. This term refers to the proportion of overlap between the content stated in the standards document and that assessed by items on the test.

Depth of knowledge (DOK) measures the type of cognitive processing required by items and content standards. For example, is a student expected to simply identify or recall basic facts or use reason to manipulate information, or to strategize how to best solve a complex problem? Using Science as an example, a student may be asked to identify the planets of our solar

[4] See http://www.ccsso.org/Documents/2006/Creating_Aligned_Standards_2006.pdf for background information on alignment.

system among several answer choices. This task should be less complex than trying to compare and contrast the composition of the planets in preparation of landing unmanned probes.

The purpose of using DOK as a measure of alignment is to determine whether a test item and its corresponding standard are written at the same level of cognitive complexity. Panelists make two separate judgments about cognitive complexity, one rating for the standard and one rating for the item. These two judgments are compared to determine whether the item is written at the same level as the standard to which it is linked. Webb (1997) refers to this comparison as *Depth-of-Knowledge consistency*.

Range-of-knowledge correspondence examines the range-of-knowledge correspondence between the assessment and content standards. The range-of-knowledge correspondence measure looks in greater detail at the breadth of knowledge represented by test items. Categorical concurrence simply notes whether a sufficient number of items on the test covers each general content topic (i.e., standard). However, states usually lay out more specific content objectives (i.e., grade level expectations, evidence outcomes), under each strand. The range-of-knowledge correspondence indicates the number of specific content objectives assessed by items.

Balance-of-knowledge representation focuses on content coverage in yet more detail. In this case, the number of items matched to the content objective does matter. The balance of representation determines whether the assessment measures the content objectives equitably within each content topic using only the content objectives identified by panelists and not all content objectives eligible to be assessed. Based on Webb's (1997) method, items should be distributed evenly across the objectives per content topic for good balance. The balance-of-knowledge representation is determined by calculating an index, or score, for each content topic. Each should meet or surpass a minimum index level to demonstrate adequate balance.

All of Webb's (1997) measures begin with calculations for each panelist's data and build up to a summary of results across panelists. To calculate categorical concurrence, range-of-knowledge, and balance-of-knowledge statistics, individual ratings of the measured content standard of each item are analyzed by panelist and then averaged across panelists. To calculate depth of knowledge statistics, consensus ratings of the DOK of each content standard are compared to individual ratings of item-level DOK. The number of items rated below, at, or above the DOK level of the linked content standard are then averaged across panelists. The procedures for collecting these consensus and individual ratings are described in a subsequent section.

Scope of Alignment Evaluation

The alignment evaluation performed for this study involved a comparison of the EOCEP Algebra 1 test items to the SCCCRS. Highly qualified educators provided alignment ratings for the evaluation. To maintain the independent and external nature of the study, SCDE did not take part in this process. This process was conducted and directed solely by HumRRO.

The content alignment review involved two major tasks that collected the data necessary for evaluating Webb's four alignment criteria: (a) providing depth of knowledge (DOK) ratings for each content standard within the SCCCRS for Algebra 1, and (b) evaluating the EOCEP Algebra 1 test items by providing an item DOK rating, verifying the content standard that each item is intended to measure, and rating the quality of alignment between the item and the linked content standard. In addition to the traditional alignment ratings, panelists also provided two

other ratings: (a) a consensus rating of how well the test blueprint reflects the intent of the SCCCRS, and (b) an independent flagging of items for issues related to clarity, accuracy, grade-level appropriateness, and bias in content or presentation.

Panelists

HumRRO recruited a nationally recognized content expert and four Kentucky educators to serve as panelists. The purpose of using panelists external to South Carolina was to ensure an independent review by highly qualified educators who are experienced with implementing rigorous content standards. The Kentucky Academic Standards for Algebra 1 are very similar to the SCCCRS for Algebra 1 in both organization and language (see Appendix B for a crosswalk between the two sets of standards).

Educators were selected based on their prior experience teaching Algebra at the high school level and familiarity with the Kentucky Academic Standards. We sought recommendations for a nationally recognized content expert, and we also recruited from a pool of experienced alignment panelists who participated in prior alignment studies for other state or national testing programs. Table 3.1 presents selected professional and demographic characteristics of the panel.

Table 3.1. Professional and Demographic Characteristics of Panelists

			Gender	
Number of Panelists	Average Years of Experience (SD)	Percent with Master's Degree	Female	Male
5	11 (2.55)	100%	40%	60%

Training

All panelists received training at the outset of the study. HumRRO introduced key alignment concepts and a general overview of study processes and procedures. All panelists signed non-disclosure agreements before proceeding to the more detailed training on the study materials and data collection tools.

Materials

During the alignment workshop, panelists evaluated the alignment of the EOCEP Algebra 1 test items with the SCCCRS for Algebra 1 by reviewing paper copies of test items (including screen shots from two technology-enhanced items) and completing electronic rating forms adapted from Webb (2005). All rating forms were completed electronically in Excel®. The item presentation and rating forms are discussed in further detail below. Excerpts from the rating forms are also presented in Appendix C.

Test Items. Panelists evaluated operational Algebra 1 items from the Winter 2016-17 test form and from the Spring 2017 test form. The assessments are administered in paper and online format. However, most items are identical in either format. For the two technology-enhanced items presented on the Spring 2017 test form, screen shots of the technology enhancements (e.g., dropdown menus) were provided in paper format. Table 3.2 lists the number of items from each form. One item was presented on both forms, so was only reviewed once. Because the test items are secure, this report does not include any examples of items or references to specific item content.

Table 3.2. Number of Algebra 1 Items Reviewed

Form	Total Items
Spring 2017	49
Winter 2016-17	50
Total	99

Rating Forms and Instructions. Panelists were given instruction sheets describing the rating tasks, the codes to be used, and the excel documents used during the review. Panelists completed two rating forms, the first was completed as a group (by consensus) to provide depth of knowledge (DOK) ratings for the content standards and the second form, an item rating form, captured individual ratings for the items. Samples of the materials are found in Appendix C.

Procedures

HumRRO conducted the alignment study over a two-day period at its office in Louisville, Kentucky. The workshop began with training and orientation to materials and rating procedures. Two HumRRO staff were available throughout the workshop to assist with logistics and questions. Prior to beginning their review, panelists read and signed affidavits of nondisclosure for the secure materials they would be reviewing during the workshop.

Before each of the rating tasks, a HumRRO staff member trained panelists on the procedures to complete the task, answered questions on the rating criteria, and facilitated a short calibration activity to ensure panelists were comfortable applying ratings. HumRRO staff provided general suggestions and comments when appropriate; however, they emphasized to panelists that staff would not give explicit direction on how to rate standards or items because panelists were valued as content experts. Each panelist was assigned a workstation with rating forms already uploaded on their assigned laptop computer. HumRRO staff provided instructions as needed for working with the electronic rating forms.

Panelists began with DOK evaluations of the content standards. Panelists started this process by independently assigning a DOK level to one standard and then discussing their individual ratings with the group until a consensus rating was reached. When all panelists felt comfortable with the task they followed a similar process in which they provided independent ratings for each standard prior to identifying a group consensus rating. To make the consensus rating of the test blueprint, panelists reviewed the full set of SCCCRS for high school mathematics after completing their consensus DOK ratings of the standards. Panelists were instructed to consider if they felt that any critical standards were omitted from the test blueprint. They then engaged in a group discussion until arriving at a single Yes or No rating of whether the test blueprint adequately reflected the intent of the standards as a whole. Regardless of their rating, panelists were also instructed to provide consensus comments. A volunteer scribe within each group recorded all consensus ratings (e.g., DOK of the content standards, overall rating of the test blueprint).

Next, panelists received specific instructions for rating the items. As a calibration activity, HumRRO staff asked panelists to rate the first two items individually and then discuss the ratings as a group. Once panelists were comfortable using the ratings, they continued the item rating activity on their own. A recalibration activity was conducted at the beginning of the second day of the workshop to ensure that panelists maintained a common approach to the rating tasks.

Panelists rated the individual items on the test forms on several dimensions: (a) depth of knowledge required by the item, (b) content match to the SCCCRS for Algebra 1, and (c) the degree of alignment (i.e., how well the item links to the identified standard). Within the content match dimension, panelists verified the standard to which the item had been linked, but they could identify additional standards if the item seemed to assess another standard as well (or nearly as well) as the existing linked standard. Finally, panelists were asked to flag items if they saw an issue related to clarity of presentation, accuracy of content, appropriateness of content or presentation for high school students, and any potential bias against student subgroups in terms of content of presentation. Panelists were reminded that items had undergone extensive review and field testing, and to only flag items in which they identified a serious issue. Any item flags were required to be accompanied by an explanation.

All panelists finished their rating tasks within the 2 days allotted for the workshop. Once panelists finished the review, they completed a session debriefing form in which they provided comments about the overall alignment of the items and content standards, as well as feedback about the quality of workshop training, processes, and materials.

Results

The following section summarizes the results from the analysis of panelists' ratings. We first report on the rate of agreement among panelists, followed by Webb's four alignment criteria and the additional ratings of the quality of test items and the test blueprint.

Interrater Reliability

Table 3.3 presents the interrater reliability coefficients for panelists' independent ratings of item DOK and of the quality of the item-to-standards link. We used the intraclass correlation coefficient (ICC; Landers, 2015; Shrout and Fleiss, 1979) as a measure of consistency in the ratings among the panelists. An ICC of .70 is generally considered sufficient for research purposes, although ICCs of .80 and above are preferred when ratings are used for making important decisions (e.g., promotion) (Graham et al., 2012). For both independently made ratings, panelists demonstrated acceptable levels of consistency.

Table 3.3. Interrater Consistency Coefficients

Rating	ICC
Item DOK	.812
Quality of Item-Standard Link	.810

Webb Alignment Results

All of Webb's (1997) measures begin with calculations for each panelist and build up to a summary of results across panelists. First, we calculated the mean ratings across items for each panelist. Next, we determined the mean rating across panelists.

The 37 Algebra 1 content standards are organized around a series of key concepts (n = 10), which are further organized into four broader content strands in the Algebra 1 test blueprint (Algebra, Functions, Number and Quantity, and Statistics and Probability). All alignment results are reported at both the content strand and the key content levels. However, South Carolina does not report student subscores for either key content strands or key concepts.

Categorical Concurrence

Categorical concurrence describes the extent to which the EOCEP Algebra 1 items, regardless of item type and point value, cover the content of the SCCCRS for Algebra 1. Webb (1997, 1999, 2005) recommends a minimum of six test questions to adequately assess each *reported* entity. This criterion serves as a guideline for reasonable content coverage based on earlier research on the reliability of tests compared to the number of items (Subkoviak, 1988).

The organization of the standards provides important context for interpreting alignment results. Each key concept consists of one or more standards. For key concepts with fewer associated standards, it would be difficult to meet Webb's criterion for adequate categorical concurrence.

Table 3.4 summarizes the results for categorical concurrence. We present the results organized by both broader content strand (n = 4) and key concept (n = 10) to illustrate that categorical concurrence will vary depending on the breadth of the level at which it is measured. The content strands and key concepts that meet Webb's criterion are in bold, italicized text. Appendix D also contains the standard deviations for each. Table 3.4 indicates that the Algebra and Functions content strands meet the categorical concurrence criterion for both the Spring 2017 Form and the Winter 2016-17Form; however, the Number and Quantity and Statistics and Probability content strands do not. Table 3.4 also illustrates that three of the 10 key concepts meet the categorical concurrence criterion for the Spring 2017 Form and Winter 2016-17 Form. It is important to reiterate that Webb's categorical concurrence criterion is intended to inform the minimum number of items required for reporting. Because South Carolina does not report scores at either the level of the content strand nor the level of the key concept, the fact that there are content strands and key concepts that do not meet the categorical concurrence criterion should be of no concern for the SCDE. Should the SCDE desire to report scores at a finer-grain level, the results in Table 3.4 suggest that, based on Webb's categorical concurrence criterion, they could do this for the following content strands: Algebra and Functions, and for the following key concepts: Reasoning with Equations and Inequalities, Interpreting Functions, and Linear, Quadratic, and Exponential. It is more informative to compare the number of items linked to content strands to the target number of items specified in the test blueprint for each content strand. As shown in the far-right column of Table 3.4, the mean number of items linked to each content strand was within the range number of items specified in the test blueprint for all content strands, except for the Number and Quantity content strand. For the Number and Quantity content strand the mean number of linked items was slightly below the 5 – 9 target (i.e., a mean of 4.8 and 4.6, respectively, for the Spring Form and Winter Form). This suggests that the SCDE might want to consider including an additional item or two to measure the Number and Quantity domain to help ensure that the assessment is meeting the intent of the test blueprint.

Table 3.4. Summary of Categorical Concurrence Results

Content Strands	Key Concepts	Number of Standards	Mean Number of Items Linked		Target Number of Items in Test Blueprint
			Spring Form	Winter Form	
Algebra		15	***21.8***	***22.8***	21 - 25
	Arithmetic with Polynomials and Rational Expressions	1	2.0	2.0	
	Creating Equations	3	4.4	4.0	
	Reasoning with Equations and Inequalities	8	***12.0***	***12.4***	
	Structure and Expressions	3	3.4	4.4	
Functions		13	***20.0***	***20.0***	18 - 22
	Building Functions	1	1.0	1.0	
	Interpreting Functions	8	***12.0***	***12.0***	
	Linear, Quadratic, and Exponential	4	***7.0***	***7.0***	
Number and Quantity		6	4.8	4.6	5 - 9
	Quantities	3	1.0	1.0	
	Real Number System	3	4.0	3.6	
Statistics and Probability		3	2.0	2.0	2
	Interpreting Data	3	2.0	2.0	
Content Strands Meeting Criteria			**2 of 4**	**2 of 4**	
Key Concepts Meeting Criteria			**3 of 10**	**3 of 10**	

In addition to verifying the content assessed by each item, we asked panelists to indicate *how well* the item assessed the content. Although this is not one of Webb's alignment criteria, it provides additional information about the strength of the alignment between items and content standards. This is especially useful when panelists verify, rather than create, the match between items and standards. Panelists rated the extent of item alignment to the content on a 3-point scale ranging from '0- not aligned' to '2- fully aligned'. An item was considered to be 'fully aligned' if all the content measured by the item was contained in the associated standard. An item was considered to be 'partially aligned' if some of the content measured by the item was not contained in the associated standard. An item was considered to be "not aligned" if none of the content of the item was contained in the associated standard.

Table 3.5 presents the mean number of items (across panelists) at each degree of alignment. On average, panelists rated approximately 90% of the items as 'fully aligned' across both forms.

Table 3.5. Panelist Ratings on Overall Item Alignment

	Spring Form			Winter Form		
Degree of Alignment	Mean Number of Items (*N*=50)	*SD*	Percent of Items	Mean Number of Items (*N*=50)	*SD*	Percent of Items
Not aligned	1.80	0.45	3.60	1.00	1.00	2.00
Partially aligned	1.60	1.14	3.20	4.20	3.11	8.40
Fully aligned	46.60	1.34	93.20	44.80	2.95	89.60

Depth-of-Knowledge Consistency

Analyses of depth-of-knowledge (DOK) measure the type of cognitive processing required of students. The DOK requirements implied by the standards should be matched by assessment items. To confirm this match, panelists were asked to rate the standards and the Algebra 1 items separately. Webb's (1997) *depth-of-knowledge consistency* criterion indicates whether there is consistency between the complexity of knowledge required by the standards and the complexity of knowledge required to correctly answer the items linked to those standards.

To make their ratings, panelists used a rating scale (adapted from Webb, 2005) with four levels of cognitive complexity.

- Level 1 Recognition – simple recall of information (i.e., facts, terms); sequencing; more automatic.
- Level 2 Skills/Concepts – beyond habitual response; applying concepts; problem-solving.
- Level 3 Strategic Thinking – requires basic reasoning, planning, or use of evidence; generating hypotheses.
- Level 4 Extended Thinking – complex reasoning; evaluation of multiple sources or independent pieces of evidence; often over an extended period of time.

Table 3.6 summarizes the depth-of-knowledge consistency results. Webb's (1997) suggested minimum for this alignment criterion is that at least 50% of the items should have complexity ratings at or above the level of the corresponding standard. The mean percentages of content strands and key concepts that reach the 50% criterion are bolded and italicized. Appendix E also contains the standard deviations for each. All the content strands meet Webb's criterion for depth-of-knowledge consistency. At the key concept level, slightly less than 50% of the items linked to the Building Functions key concept were rated at a cognitive complexity level at or above the cognitive complexity level of the corresponding standard for the Spring 2017 Form. And, for both the Spring 2017 Form and the Winter 2016-17 Form, the Structure and Expressions key concept had considerably less than 50% of its linked items with cognitive complexity levels at or above the cognitive complexity level of the linked standard. The depth-of-knowledge consistency criterion was obtained for all other key concepts. This finding suggests that the SCDE may want to review to the cognitive complexity of the items intended to measure the Structure and Expressions key concept to consider whether the cognitive complexity of the items may be enhanced so that there is greater consistency between the cognitive complexity of the items and the cognitive complexity of the standards to which the items are linked.

Table 3.6. Summary of Depth-of-Knowledge Consistency Results

Content Strands	Key Concepts	Number of Standards	Percent of Items with DOK at or Above the Level of the Linked Standard	
			Spring Form	Winter Form
Algebra		15	***66.03***	***65.81***
	Arithmetic with Polynomials and Rational Expressions	1	***100.00***	***100.00***
	Creating Equations	3	***65.00***	***50.00***
	Reasoning with Equations and Inequalities	8	***71.67***	***77.31***
	Structure and Expressions	3	26.67	33.00
Functions		13	***65.00***	***82.00***
	Building Functions	1	40.00	***60.00***
	Interpreting Functions	8	***73.33***	***86.67***
	Linear, Quadratic, and Exponential	4	***54.29***	***77.14***
Number and Quantity		6	***100.00***	***100.00***
	Quantities	3	***100.00***	***100.00***
	Real Number System	3	***100.00***	***100.00***
Statistics and Probability		3	***100.00***	***60.00***
	Interpreting Data	3	***100.00***	***60.00***
Content Strands Meeting Criteria			4 of 4	4 of 4
Key Concepts Meeting Criteria			8 of 10	9 of 10

Range of Knowledge Correspondence

The *range-of-knowledge correspondence* measure examines in greater detail the breadth of knowledge covered by the assessment. In addition to evaluating which content standards are assessed, we must look at how many of the content standards within each content strand and within each key concept are represented by items. Webb's (1997) minimum level of acceptability for range-of-knowledge correspondence is that at least 50% of standards per key reported entity link with items. Table 3.7 summarizes the range-of-knowledge results. The content strands and key concepts that meet Webb's criterion are in bold, italicized text. The range-of-knowledge criterion was met for all content strands, and for all but one of the key concepts for both the Spring 2017 Form and the Winter 2016-17 Form. The range-of-knowledge criterion was not met for the Quantities key concept. Because the SCDE does not report at the level of the key concept, this finding should be of no concern for the SCDE. Should the SCDE wish to report at the key concept level in the future, the SCDE should consider adding an additional item to the Algebra 1 assessment to address another standard within the Quantities key concept. Tables presenting means and standard deviations for each key concept and content strand are presented in Appendix F.

Table 3.7. Summary of Range-of-Knowledge Results

Content Strands	Key Concepts	Number of Standards	Percent of Standards Matched to at Least One Item	
			Spring Form	Winter Form
Algebra		15	***93.33***	***100.00***
	Arithmetic with Polynomials and Rational Expressions	1	***100.00***	***100.00***
	Creating Equations	3	***100.00***	***100.00***
	Reasoning with Equations and Inequalities	8	***100.00***	***100.00***
	Structure and Expressions	3	***66.67***	***100.00***
Functions		13	***100.00***	***92.31***
	Building Functions	1	***100.00***	***100.00***
	Interpreting Functions	8	***100.00***	***87.50***
	Linear, Quadratic, and Exponential	4	***100.00***	***75.00***
Number and Quantity		6	***63.33***	***66.67***
	Quantities	3	33.33	33.33
	Real Number System	3	***100.00***	***100.00***
Statistics and Probability		3	***66.67***	***66.67***
	Interpreting Data	3	***66.67***	***66.67***
Content Strands Meeting Criteria			4 of 4	4 of 4
Key Concepts Meeting Criteria			9 of 10	9 of 10

Balance-of-Knowledge Representation

The fourth measure of alignment included in the Webb (1997) method is *balance-of-knowledge representation*. This measure describes the distribution of items linked to each standard within each key concept and content strand. The number of items should be distributed rather evenly between the key concepts and content strands to achieve good balance.

The content balance is determined by calculating an index, or score, for each key concept.[5] According to Webb (1997), the minimum acceptable index for a single reported entity is 70 (on a scale of 0 to 100 with 100 representing perfect balance). An index of 70 or higher suggests that items broadly assess the standards within a key concept area and content strand instead of clustering around one or two standards.

It is important to note that only those standards that were indicated by panelists as being partially or fully aligned to an item are included in calculations of the balance index. A given key concept may include more standards than were verified by panelists as being linked to items. Recognizing this feature of the balance index is important in cases when the range measure and balance measure produce seemingly contrasting results.

[5] The exact formula for calculating the balance index is explained in detail in Webb's (2005) alignment training manual: http://www.wcer.wisc.edu/WAT/index.aspx.

Table 3.8 summarizes the results on balance-of-knowledge representation. The EOCEP Algebra 1 test surpassed the minimum level of acceptability (index of 70) for demonstrating good content balance among those standards linked to items within each content strand and key concept for both forms. Tables containing means associated with the calculation of the balance index are presented in Appendix G.

Table 3.8. Summary of Balance-of-Knowledge Representation Results

Content Strands	Key Concepts	Number of Standards	Mean Balance Index	
			Spring Form	Winter Form
Algebra		15	**79.20**	**79.49**
	Arithmetic with Polynomials and Rational Expressions	1	**100.00**	**100.00**
	Creating Equations	3	**84.67**	**83.33**
	Reasoning with Equations and Inequalities	8	**75.00**	**77.88**
	Structure and Expressions	3	**90.00**	**79.33**
Functions		13	**81.15**	**80.00**
	Building Functions	1	**100.00**	**100.00**
	Interpreting Functions	8	**79.17**	**79.17**
	Linear, Quadratic, and Exponential	4	**89.29**	**90.48**
Number and Quantity		6	**84.67**	**91.00**
	Quantities	3	**100.00**	**100.00**
	Real Number System	3	**83.33**	**90.00**
Statistics and Probability		3	**100.00**	**100.00**
	Interpreting Data	3	**100.00**	**100.00**
Content Strands Meeting Criteria			4 of 4	4 of 4
Key Concepts Meeting Criteria			10 of 10	10 of 10

Additional Ratings

Evaluation of Test Blueprints

Panelists indicated via consensus that the EOCEP Algebra 1 test blueprint adequately reflects the intent of the SCCCRS for high school mathematics. In their accompanying consensus comments, they made reference to two additional content standards which they felt were reflected in the test items, but not explicitly enumerated in the test blueprint. They were:

1. FBF.2 Write arithmetic and geometric sequences both recursively and with an explicit formula, use them to model situations, and translate between the two forms.
2. AREI.7 Solve a simple system consisting of a linear equation and a quadratic equation in two variables algebraically and graphically. Understand that such systems may have zero, one, two, or infinitely many solutions.

Panelists stated that FBF.2 addresses sequences, a concept reflected in the test items, but not explicitly enumerated in the test blueprint. Panelists stated that AREI.7 also seems to be reflected in the test items, but is not included on the blueprint and is not identified as an SCCCR Graduation Standard.

Evaluation of Item Quality

Panelists' independent evaluations of item quality were analyzed to identify any items that were flagged on the same element of item quality by the majority of panelists (i.e., 3 of 5). Two of the 99 items reviewed (2%) were flagged by at least three panelists on the same element. Specifically, the majority of panelists raised concerns about the clarity of presentation of these two items (e.g., not clear what is being asked, labeling or other information is missing). Because the test items are secure, this report does not reference specific item content. At the EOC's request[6], a password-protected document of the panelists' item quality ratings, including panelists' comments/explanations of their item quality ratings, was shared with DRC.

Discussion

The overall alignment results provide positive support for the content validity of the EOCEP Algebra 1 test. Summary alignment judgments are based on Webb's summary criteria (2005). These summary judgments focus on the percentage of content strands and key concepts well represented by the assessment. Webb outlined a scale with a range of potential alignment outcomes applied to each of the four criteria:

- Fully aligned – assessment aligns to all content strands or key concepts (91%–100%),
- Highly aligned – assessment aligns to the majority of content strands or key concepts (70%–90%),
- Partially aligned – assessment aligns well to some content strands or key concepts (50%–69%),
- Weakly aligned – assessment aligns to less than half the content strands or key concepts (below 50%).

Webb's (1997) alignment method does not allow for a *single* judgment of overall alignment across the four alignment criteria. However, one can get a sense of overall alignment between the assessments and standards by looking at the alignment criteria altogether. Tables 3.9 (Spring 2017 Form) and 3.10 (Winter 2016-17 Form) present the summary alignment outcomes for the EOCEP Algebra 1 test based on the above scale. The table includes a summary judgment for each Webb alignment criterion based on the percentage of content strands or key concepts that met the target. As shown in Tables 3.9 and 3.10, per Webb's scale, for categorical concurrence there is partial and weak alignment for content strands and key concepts, respectively (for both the Spring and Winter forms). However, it is important to note that categorical concurrence is intended to inform the minimum number of items required for a reporting category. Because the SCDE does <u>not</u> report scores at the level of the content strand or at the level of the key concept, the "partial" and "weak" alignment on Webb's categorical concurrence criterion should be of no concern for the SCDE. It is more informative to compare the number of items linked to content strands to the target number of items specified in the test blueprint for each content strand. This comparison shows that the mean number of items linked to each content strand was within the range number of items specified in the test blueprint for all content strands, except for the Number and Quantity content strand. For the Number and Quantity content strand, the mean number of linked items was slightly below the target. This suggests that the SCDE might want to consider including an additional item to measure the Number and Quantity domain to help ensure that the assessment is meeting the intent of the

[6] Teleconference meeting with EOC staff, Melanie Barton and Kevin Andrews, on March 22, 2017 to discuss initial draft report.

test blueprint. For all the remaining Webb criteria, Tables 3.9 and 3.10 show that there is high to full alignment even at these finer-grain descriptor levels. Based on these Webb alignment criteria, the alignment study results show that the EOCEP Algebra 1 test items reflect the range of intended content domain.

Table 3.9. Summary Alignment Outcomes on Each Webb Criterion (Spring 2017 Form)

Percentage of Key Concepts and Content Strands Meeting Webb Criteria							
Categorical Concurrence		Depth-of-Knowledge Consistency		Range-of-Knowledge Correspondence		Balance-of-Knowledge Representation	
Content Strand	Key Concept	Content Strand	Key Concept	Content Strand	Key Concept	Content Strand	Key Concept
Partially aligned (50%)	Weakly aligned (33%)	Fully aligned (100%)	Highly aligned (80%)	Fully aligned (100%)	Highly aligned (90%)	Fully aligned (100%)	Fully aligned (100%)

Table 3.10. Summary Alignment Outcomes on Each Webb Criterion (Winter 2016-17 Form)

Percentage of Key Concepts and Content Strands Meeting Webb Criteria							
Categorical Concurrence		Depth-of-Knowledge Consistency		Range-of-Knowledge Correspondence		Balance-of-Knowledge Representation	
Content Strand	Key Concept	Content Strand	Key Concept	Content Strand	Key Concept	Content Strand	Key Concept
Partially aligned (50%)	Weakly aligned (33%)	Fully aligned (100%)	Highly aligned (90%)	Fully aligned (100%)	Highly aligned (90%)	Fully aligned (100%)	Fully aligned (100%)

Additional ratings by panelists provide further support for the EOCEP Algebra 1 test as a strong measure of the intended content domain. Panelists agreed that the test blueprint captures the intent of the South Carolina College- and Career-Ready Standards for High School Mathematics. Furthermore, items were found to be of sufficient quality, with only a very small number of items rated by the majority of panelists as having issues related to clarity of presentation.

Finally, panelists were also asked to provide general opinions via the session debriefing form administered at the end of the workshop. Comments provided offer additional support for the positive results reflected in the analysis of item ratings. Four of the five panelists (including the nationally recognized content expert) rated the overall alignment between the EOCEP Algebra 1 items and the SCCCRS as '4 - Good' on a scale ranging from '1 - Not aligned in any way' to '5 - Perfect Alignment.' Three of the five panelists also commented that they felt the cognitive/performance expectations reflected in the test items reflected the appropriate range for students tested at the high school level.

However, panelists also raised a couple concerns in their comments. Three of the five panelists mentioned the limitations of multiple-choice tests, such as the EOCEP Algebra 1 test, for providing useful information about the South Carolina College-and Career-Ready Mathematical Process Standards, or to support research-based instruction. Four of the five panelists also mentioned that some items might be biased towards students with access to, and familiarity with, graphing calculators, though one panelist stated that this is common to most math tests.

Chapter 4: Review of Algebra 1 and SC Ready Test Construction

Introduction

Forms construction refers to the assembly of test items into forms that meet certain specifications for content, statistical properties, and construct representation (e.g., a test blueprint). The *Joint Standards* describe best practices surrounding the forms construction process. HumRRO identified eight *Joint Standards* that were directly related to aspects of forms construction for this evaluation task. A rationale for the *Joint Standards* incorporated into this review is described in Appendix H.

The current report evaluates test construction processes for (a) EOCEP Algebra 1, (b) SC Ready Math, and (c) SC Ready ELA. DRC provided test form metadata for four Algebra 1 forms (Winter 2016-17 and Spring 2017, Online and Print), and item bank metadata for Algebra 1 and SC Ready Math and ELA. We organized this chapter by assessment program followed by the area of evaluation for Test Construction: (a) fidelity of documented practices to relevant *Joint Standards* and (b) analysis of test forms and/or item bank metadata.

Methods

Documents and Datasets

Documents relevant to forms construction were provided to HumRRO by DRC to satisfy the review of forms construction processes for the three aforementioned exams. Reviewed documents and datasets are presented in Table 4.1.

Table 4.1. Documents and Datasets Reviewed for Task 3 – Forms Construction

Report Section	Document Filename
End-of-Course Exams (Algebra 1)	
Fidelity to Forms Construction Standards	004_SCCCR Algebra 1 Test Blueprint_E.pdf
	005_EOCEP 2016-2017 Form Construction Guidelines_E.pdf
	006_Guidelines for making changes within a test form_RE.pdf
	011_Test Form Construction Process_E.pdf
	012_Quality Assurance Procedures for Test Construction_RE.pdf
	015_ Guidelines for Ordering Items_E.pdf
Test Form Metadata	EOCEP Algebra Fall_Winter 16_17 Online Metadata.xlsx
	EOCEP Algebra Fall_Winter 16_17 Print Metadata.xlsx
	EOCEP Algebra Spring 17 Print Metadata.xlsx
	EOCEP Algebra Spring 17 TE Items Online Metadata.xlsx
Item Bank Metadata	017_Item Metadata Available Pool_E.xlsx
SC Ready (Math and ELA)	
Fidelity to Forms Construction Standards	004_SC READY English Language Arts Blueprint_R.pdf
	004_SC READY Mathematics Blueprint_R.pdf
	006_Guidelines for making changes within a test form_RE.pdf
	009_SC READY Item Development Plans_R.pdf
	010_Test Form Construction Process_R.pdf
	012_Quality Assurance Procedures for Test Construction_RE.pdf
	016_Guidelines for Item Analysis and Form Construction_R.pdf
Item Bank Metadata	Math Item Metadata_2016_2017.xlsx
	ELA Item Metadata_2016_2017.xlsx

Following a review of the available documentation, we conducted a phone interview with staff from SCDE and DRC. The purpose of this call was to ask follow-up questions about the documentation and fill in any gaps in our understanding.

A global numeric rating was assigned to each Standard after reviewing all documents for each exam and conducting the phone interview. The scale used is presented in Table 4.2. The goal was to quantify the fidelity of the practices as described in the forms construction documents to the *Joint Standards* for each exam. In addition to the numeric rating, specific aspects of the *Joint Standard* that were missing from the documentation are listed in the comment following the rating. The following section is organized by *Joint Standard* number and includes the text of the standard, the assigned rating, and an explanation of what was not found in the documentation provided by the testing contractor.

Table 4.2. Rating Scale for Joint Standards

Score Level	Description of Score Level
1	No evidence of the Standard found in the test forms construction materials[a].
2	Little evidence of the Standard found in the test forms construction materials[a]; less than half of the Standard covered in the documents and/or evidence of key aspects of the Standard could not be found.
3	Some evidence of the Standard found in the test forms construction materials[a]; approximately half of the Standard covered in the materials[a], including some key aspects of the Standard.
4	Evidence in the test forms construction materials[a] mostly covers the Standard; more than half of the Standard covered in the materials,[a] including key aspects of the Standard.
5	Evidence in the test forms construction materials[a] fully covers all aspects of the Standard.

[a]Materials include all documents provided, any emails or phone calls with SCDE/DRC staff, as well as what could be found online.

Results

Algebra 1

We first examined the documentation provided concerning Algebra 1 forms assembly for adherence to the *Joint Standards*. Then, we analyzed the *test form metadata* of the four operational test forms to determine if the forms met the content and statistical specifications. We also analyzed *item bank metadata* that was provided (see Table 4.1 above for DRC's document file names). Results for SC Ready are presented after Algebra 1 results, followed by a combined discussion of SC Ready and Algebra 1.

Fidelity to Forms Construction Standards

Table 4.3 presents the rating assigned to each *Joint Standard* under review relative to the documentation available on the EOCEP Algebra 1 test. In subsequent paragraphs, we explain each rating, pointing out areas where the documentation exceeded or perhaps did not meet the *Joint Standard*.

Table 4.3. Algebra 1 Evaluation Results Based on the Joint Standards

Joint Standard Number	Standard Content	Rating
Standard 4.1	Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	4
Standard 4.2	In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	3
Standard 4.4	If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.	3
Standard 4.5	If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5
Standard 4.7	The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5
Standard 4.9	When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	5
Standard 4.10	When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	3
Standard 4.13	When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	3

Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

The purpose and definition of the EOCEP Algebra I exam is not defined in the *005_EOCEP 2016-2017 Form Construction Guidelines_E.pdf*; however, the state website describes its use. The test blueprint also clearly specifies the content strands, key concepts and content standards that must comprise each form. The intended examinee population is inferred to be high school but no grade range is explicitly mentioned. The item review guidelines, however, do recommend reviewing for appropriate grade-level wording. The purpose is clear as an End-of-Course exam; however, the "high-stakes" use of test results is mentioned but not defined. For example, it is unclear whether the exams are used for summative purposes at the school or district-level or for individual students? Without an explicit mention of the intended test use, it is challenging to recommend what validity studies should be conducted to support those uses.

Standard 4.2 – In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.

The documentation provides a clear description of test content, length, item format, and some psychometric properties and item order. However, there is no mention of internal consistency reliability minimums or if this is considered when creating forms. There is no mention of time allowed for testing in the documentation, but the online system directions do mention that there is no time limit (this is not included in the paper form directions). Although included on the actual exam, the test specifications should include the wording of the directions. Also missing are administration guidelines, materials, or scoring and reporting procedures. It is mentioned that a Rasch model is the measurement model but there is no mention of what scores students receive (i.e., scale score) and their range.

Standard 4.4 – If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.

Paper-based tests (PBT) and non-adaptive computer-based test (CBT) forms are assembled using the same test specifications. Our understanding is that all PBT items are ported to the CBT with a few substitutions. That is, some item types can only be administered on computer (technology enhanced [TE] items) and these are swapped for items in the same content standard on the PBT version of an exam. There is no mention of different psychometric targets for CBT forms, if they are assumed to be the same as PBT. Although we are aware that mode DIF analyses are done, these procedures should be made clear in the documentation to satisfy the Standard. As mentioned in a phone interview on March 1, 2017, item-level mode differential item functioning (DIF) is explored using ETS's Delta method. Items with category "C" DIF are

sent to item developers for review, although SCDE staff indicated that items rarely reach that level of DIF for mode comparisons.

Standard 4.5 – If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

Variations in administration conditions that this Standard covers are accommodations for students with disabilities. There are some tools available on the online testing platform that appear to be for students with disabilities (i.e., visual impairments); however, these features are not described in the documents provided. There are some descriptions of accommodations online for EOCEP exams (http://ed.sc.gov/tests/assessment-information/testing-swd/accommodations-and-customized-forms/).

Standard 4.7 – The procedures used to develop, review, and try out items and to select items from the item pool should be documented.

The procedure for selecting field test (FT) items is well-documented in terms of number of items, their placement, and statistics.

Standard 4.9 – When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.

The FT design uses an embedded approach where FT items are spread throughout an operational form in a standard testing environment (forms are then scrambled with the intention of administering FT items to a random sample of students). This approach ensures that items are field tested using a sample of students that come from the same population that complete the operational, scored items. This also allows for accurate item parameter estimation given that students are unaware of which items are scored and which are being field tested.

Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

Based on the test forms construction documents provided, it is not clear if item statistics used for form assembly are estimated during FT, prior operational use, or whatever the recent parameter estimate is regardless of operational or FT use. The primary documentation mentions classical test theory (CTT) item parameter targets. That is, a mean P-Value and median point-biserial range is provided to guide form-level evaluation. However, the "quality assurance" document mentions that "[t]he use of standardized-test construction software enables the construction of forms with similar test characteristic functions and standard errors of

measurement curves." So, it is not clear if forms are evaluated using CTT or item response theory (IRT) procedures like the quality assurance document appears to suggest, or perhaps both are used. It is mentioned that items with differential item functioning (DIF) is considered using the ETS Delta method. Items with DIF flags of "C" are not considered but those with "B" may be considered. It is not clear when DIF is evaluated, FT or operational, or after every administration.

Standard 4.13 – When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

The documentation provided does describe a paper-based and computer-based form, comprised of the same items, but differing in presentation. There was no evidence of a study investigating if test scores of these two modes are comparable, or if item parameters are similar. If data from paper-based and computer-based test forms are combined to estimate (calibrate) item parameters, and these parameters are used to assemble forms, there could be a situation where the "true" item parameters (that is, with mode effects removed) do not meet the psychometric guidelines. Mode differences are just one source of possible construct-irrelevant variance. The documentation does not provide evidence of any studies to investigate other possible sources of irrelevant variance.

Test Form Metadata

Test form metadata was provided for four Algebra 1 forms. The mean P-Value target for Algebra 1 forms is 0.65 and the median point-biserial target is between 0.35 and 0.45. The four forms meet the point-biserial target but each are more difficult than the target mean P-Value (see Table 4.4). These forms also seem to be more difficult than 2015 forms (mean P-Value = 0.59). The form assembly documents do state that content specifications are prioritized over psychometric guidelines, so it is possible that high-demand strands happen to have more difficult items, thus lowering the form's overall difficulty.

Table 4.4 Algebra 1 Classical Test Theory Descriptive Statistics

	P-Values				Point-Biserial Correlations			
Form	Min	Max	Mean	Median	Min	Max	Mean	Median
Winter Online	0.23	0.87	**0.52**	0.49	0.18	0.66	0.36	**0.36**
Winter Print	0.23	0.87	**0.53**	0.51	0.18	0.66	0.36	**0.36**
Spring Online[a]	0.23	0.90	**0.56**	0.55	0.10	0.62	0.37	**0.37**
Spring Print	0.26	0.90	**0.58**	0.57	0.10	0.62	0.37	**0.36**

Note. Mean P-Value target = 0.65, median point-biserial target between 0.35 and 0.45. We assume that online and paper items are calibrated separately.
[a]Spring Online form contains the same items as Spring Print form except for two technology enhanced (TE) items that substitute two items from the same content standard.

Comparing the four forms provided to the test specifications revealed that each form met the content specifications at the Key Concept level. There were some minor violations for the "range # items" column in the test blueprint, but these were not apparent in the broader Key Concept levels. These "range # items" are simply possible ways to reach the Key Concept levels. According to our call with DRC and SCDE staff on March 1, 2017, the forms assembly software algorithm only considers the Key Concept level when assembling forms.

EOCEP Algebra 1 Item Bank Metadata

An item bank containing content codes and item statistics was provided by DRC to HumRRO. A total of 725 items were contained in the item bank. Of these 725 items, 384 (52.97%) had content codes that contained "ALG15." Of these 384 items, 295 (76.82%) had no missing item stats. These conditions qualified items for selection on forms according to an email from DRC staff on March 1, 2017. Table 4.5 contains classical item statistics for the eligible item bank (k = 295).

Table 4.5 Algebra 1 Item Bank Descriptive Statistics

	k	Min	Max	Median	Mean	SD
P-Values	295	0.00	0.90	0.47	0.47	0.21
Point-Biserial Correlations	295	-0.10	0.69	0.33	0.32	0.15

Notably, the mean P-Value is .47; however, the target for form assembly is .65, which is considerably easier than what the bank provides overall. Figure 4.1 depicts the distribution of P-Values for the eligible bank. The target mean P-Value of .65 also shows how many items are less than the target mean difficulty. A lack of easy items in the bank explains why the Algebra 1 forms are more difficult than the test specifications. However, it is important to note that approximately 8% of items have a P-Value less than 0.2.

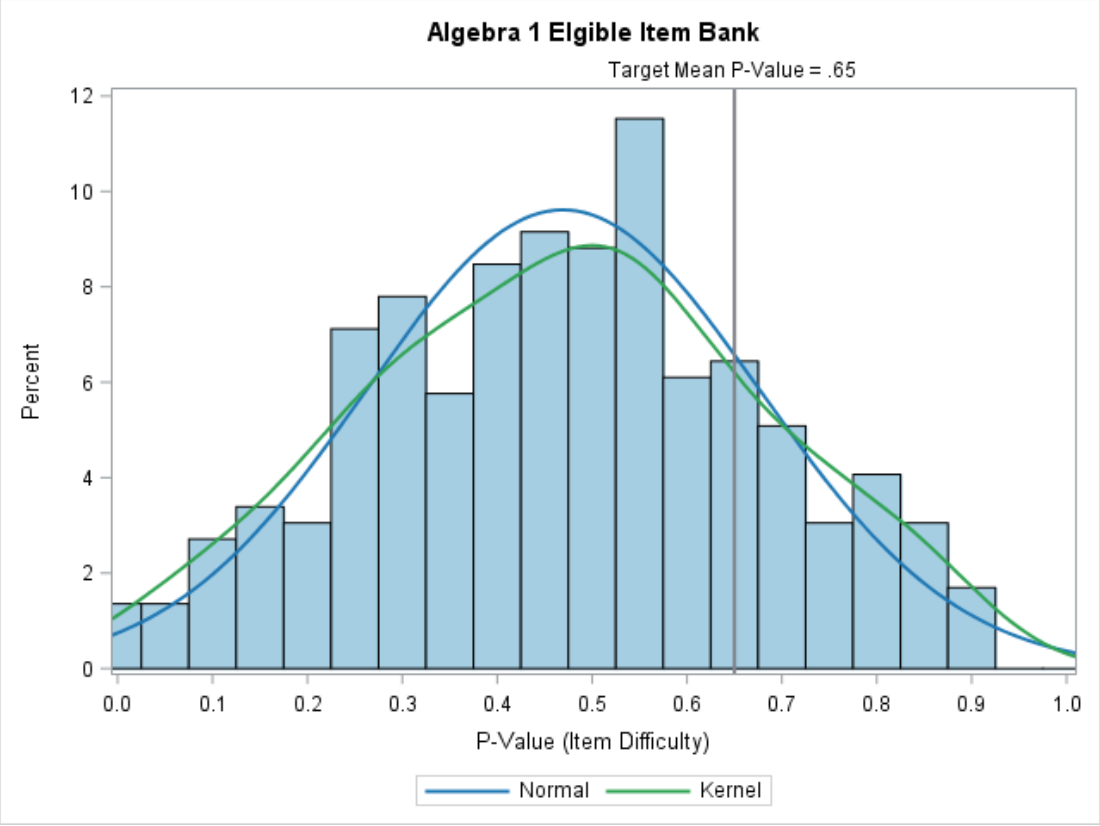

Figure 4.1. Distribution of P-Values from eligible item bank.

Point-biserial correlations in the bank are a bit closer to the form targets. The median and mean point-biserial is at .33, which is fairly close to the target median range of .35 to .45. Figure 4.2 depicts the number of items that are close to this range. Also notable are the handful of items

with point-biserial correlations around or below 0. These items contribute little or no information to the goal of determining student performance and should be re-evaluated or made ineligible for placement on future forms. Items with negative point-biserial correlations may actually reduce the overall validity of the assessment.

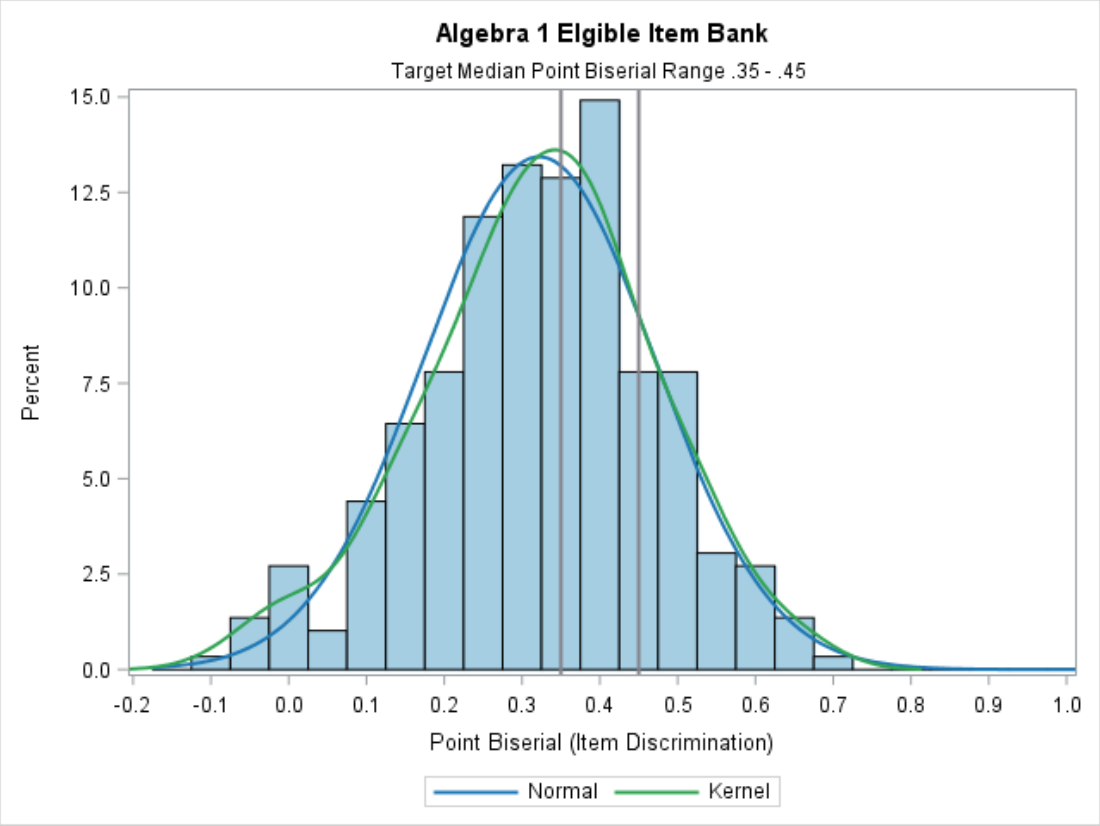

Figure 4.2. Distribution of point-biserial correlations from eligible item bank.

SC Ready

Similar to the EOCEP Algebra 1 exam section, SC Ready forms construction documents were first reviewed to gauge their fidelity with the same *Joint Standards* identified above. Due to the same documents and procedures for both Math and ELA exams, a single set of ratings are provided for both subjects. Item bank metadata was then analyzed for SC Ready Math and ELA separately, as well as separately by year.

It is important to note that for SC Ready, SCDE leases items from DRC's college and career readiness (CCR) item bank, which is also used by other DRC clients. Consequently, for security purposes, metadata was only provided for those items used by South Carolina for the Spring 2016 and Spring 2017 SC Ready assessments. It is also important to note that DRC's final data quality checks on the SC Ready metadata had not yet been completed by the date on which the metadata was provided to HumRRO. DRC provided the metadata to HumRRO prior to completion of its final QA check in order to meet the due date for this first interim report.[7]

[7] Email communication from Shar Moseng at DRC on February 24, 2017.

Fidelity to Forms Construction Standards

Table 4.6 presents the rating assigned to each *Joint Standard* under review relative to the documentation available on the SC Ready ELA and mathematics tests. In subsequent paragraphs, we explain each rating, pointing out areas where the documentation exceeded or perhaps did not meet the Standard.

Table 4.6. SC Ready Evaluation Results Based on the Joint Standards

Joint Standard Number	Standard Content	Rating
Standard 4.1	Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	5
Standard 4.2	In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	5
Standard 4.4	If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.	3
Standard 4.5	If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	5
Standard 4.7	The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	5
Standard 4.9	When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	5
Standard 4.10	When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	4
Standard 4.13	When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	3

Standard 4.1 – Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

The purpose of the SC Ready exams was not stated in the documents provided, although it does appear in the SC Ready Test Administration Manual obtained here (*http://ed.sc.gov/tests/tests-files/sc-ready-files/2016-sc-ready-test-administration-manual-tam/*). One of the primary uses of test scores are to meet the annual accountability requirements defined by SC law. The intended examinee population is inferred to align with the grade level of the test. Although inferred, the definition of the ELA and Math constructs can be defined by a test blueprint, which were provided for ELA and Math.

Standard 4.2 – In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.

The *016_Guidelines for Item Analysis and Form Construction_R.pdf* document describes in detail the assembly of test items into forms including: item order, item statistics, cueing, answer key repetitions, and content specifications, among other characteristics. Any details that were not immediately clear in the provided documentation (e.g., test format, time), were found in the Test Administration Manual online.

Standard 4.4 – If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.

The *016_Guidelines for Item Analysis and Form Construction_R.pdf* document states that computer-based test forms are first constructed and then paper-based forms have the same items with a few substitutions. There is no discussion on the difference of psychometrics specifications although it can be assumed that these companion items presented on the paper forms have item characteristics similar to their computer-based form counterparts. That is, assuming there are no item-level mode effects or differences in performances based on mode of delivery.

Standard 4.5 – If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.

The SC Ready exams are administered online to the majority of students. Accommodated online and paper and pencil exams are allowed for students who (a) have an IEP or 504 plan

that specifies paper-based testing only or (b) have a waiver for the computer-based requirement.

Standard 4.7 – The procedures used to develop, review, and try out items and to select items from the item pool should be documented.

The *016_Guidelines for Item Analysis and Form Construction_R.pdf* document explains that about 25% of a Math or ELA form are refreshed each year with field test items.

Standard 4.9 – When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.

FT items are embedded into operational forms so the sample used to calibrate these new items is very similar to the sample that will be used operationally in the future.

Standard 4.10 – When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.

The psychometric guidelines for SC Ready and the EOCEP exams are identical in terms of their CTT targets. The guidelines for picking "good" items are also identical and satisfy that portion of the *Joint Standards*. It appears that CTT parameters are the only psychometric evaluation of a test form for the SC Ready. According to the *016_Guidelines for Item Analysis and Form Construction_R.pdf* document, a Rasch model is used to estimate item difficulties as well as determine a test form's level of difficulty. However, this process only appears to be used for equating purposes, not forms construction. During the phone interview with SCDE and DRC staff on March 1, 2017, DRC staff confirmed this assumption. There were no guidelines surrounding issues of item parameter drift or non-convergence, if that occurs, in the post-equating process.

Standard 4.13 – When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.

According to a phone interview with SCDE and DRC staff on March 1, 2017, SCDE staff indicated that all items are subjected to comparison between paper and computer-based data for mode differences. On this call, SCDE staff indicated that no items have been categorized as an ETS "C" level since 2008. If any items were to reach that level, they would then be sent for content review and not immediately made ineligible for future forms. Even though we are aware of item-level mode difference comparisons, the documentation we received does not describe that these comparisons are conducted.

There was no mention of forms-level comparisons between paper and computer forms. This may be because small sample sizes for paper forms preclude these types of comparisons. South Carolina is moving toward a near 100% online assessment. It may also be the case that the forms are equated to eliminate forms-level differences. We would expect forms comparisons to be documented if they exist, or we would expect the rationale for not conducting such comparisons to be documented.

SC Ready Metadata

SC Ready metadata was provided by DRC to HumRRO in two separate datasets, one for Math and one for ELA (see Table 4.1 above for DRC's document file names). The data were further divided by grade (3 – 8) and year (2016 and 2017), thereby representing the items available for forms assembly by subject and grade between 2016 and 2017. As noted above, because SCDE leases items from DRC's CCR item bank, for test security purposes, metadata was only provided for those items used for the spring 2016 and spring 2017 SC Ready assessments.

MATH

SC Ready metadata for mathematics was analyzed first. Several grades had duplicate item IDs within the same grade; these duplicates had the same item statistics and content codes. Therefore, they were removed prior to any analyses. Tables 4.7 and 4.8 below represent the unique items provided in the SC Ready metadata. Furthermore, a few items were missing data for one or both CTT parameters; consequently, descriptive statistics were calculated separately by statistic to incorporate as much information as possible.

Table 4.7 contains descriptive statistics of the P-Values and point-biserial correlations (i.e., CTT parameters), by grade for the spring 2016 metadata. Table 4.8 contains the same information, but from the spring 2017 metadata. Recall the psychometric targets for forms were P-Values between .30 and .85 and point-biserial correlations greater than or equal to .20. These columns are in bold face in Tables 4.7 and 4.8. Notably, in 2016, the mean P-Values were at the lower end of the target range for forms and the median point-biserial correlations were low but above the .20 cut-off, except for Grade 3. In 2017, the item bank shifted higher in mean P-Value (difficult items were removed, easier items were added, or both) and the median point-biserial correlations increased for all grades.

Table 4.7. 2016 SC Ready (Math) Item Bank Descriptive Statistics

	P-Values						Point-Biserial Correlations					
Form	k	Min	Max	Mean	SD	Median	k	Min	Max	Mean	SD	Median
Grade 3	33	.07	.84	**.36**	.18	.32	32	-.31	.62	.17	.25	**.14**
Grade 4	40	.09	.86	**.44**	.16	.47	40	-.11	.71	.35	.20	**.33**
Grade 5	48	.05	.81	**.44**	.19	.45	48	-.05	.75	.35	.19	**.38**
Grade 6	45	.14	.86	**.45**	.20	.43	45	-.11	.64	.39	.16	**.38**
Grade 7	44	.07	.95	**.38**	.22	.36	44	-.16	.64	.34	.18	**.34**
Grade 8	46	.06	.82	**.35**	.17	.33	46	-.42	.85	.25	.29	**.30**

Note. Mean P-Value target between 0.30 and 0.85, median point-biserial target >= 0.20.

Table 4.8. 2017 SC Ready (Math) Item Bank Descriptive Statistics

	P-Values						Point-Biserial Correlations					
Form	k	Min	Max	Mean	SD	Median	k	Min	Max	Mean	SD	Median
Grade 3	49	.23	.84	**.59**	.15	.61	49	.20	.59	.42	.09	**.41**
Grade 4	52	.30	.85	**.56**	.13	.56	52	.12	.63	.41	.10	**.40**
Grade 5	53	.22	.88	**.53**	.17	.52	53	.13	.68	.42	.11	**.43**
Grade 6	42	.37	.87	**.58**	.13	.60	42	.20	.62	.41	.10	**.42**
Grade 7	37	.21	.73	**.49**	.14	.51	37	.17	.59	.37	.12	**.37**
Grade 8	59	.17	.82	**.49**	.15	.48	59	.12	.69	.40	.11	**.41**

Note. Mean P-Value target between 0.30 and 0.85, median point-biserial target >= 0.20.

ELA

As with Math, several grades had duplicate item IDs and item statistics within the grade for ELA. Duplicate items were removed prior to any analyses. The numbers in Tables 4.9 and 4.10 reflect the number of unique items that had no missing data for each statistic. For SC Ready ELA, the spring 2016 metadata appeared more in-line with the target psychometrics than Math 2016 (see Tables 4.9 and 4.10). The mean P-Values are in the middle of the acceptable range and the median point-biserial correlations are above .20. There are some point biserial-correlations below the .20 cutoff in all grades, but they are mostly positive unlike Math 2016.

Table 4.9. 2016 SC Ready (ELA) Item Bank Descriptive Statistics

	P-Values						Point-Biserial Correlations					
Form	k	Min	Max	Mean	SD	Median	k	Min	Max	Mean	SD	Median
Grade 3	54	.06	.92	**.50**	.18	.50	54	.09	.59	.37	.12	**.37**
Grade 4	53	.18	1.22	**.57**	.18	.59	53	.08	.62	.36	.13	**.38**
Grade 5	52	.21	1.39	**.61**	.20	.64	52	.06	.68	.40	.13	**.41**
Grade 6	63	.07	1.35	**.50**	.21	.50	63	.08	.65	.39	.13	**.41**
Grade 7	65	.23	1.54	**.55**	.18	.54	65	-.03	.63	.40	.12	**.42**
Grade 8	62	.15	1.74	**.58**	.23	.62	62	.05	.68	.40	.15	**.41**

Note. Mean P-Value target between 0.30 and 0.85, median point-biserial target >= 0.20.

Table 4.10. 2017 SC Ready (ELA) Item Bank Descriptive Statistics

	P-Values						Point-Biserial Correlations					
Form	k	Min	Max	Mean	SD	Median	k	Min	Max	Mean	SD	Median
Grade 3	61	.18	.83	**.54**	.16	.55	60	.14	.58	.38	.10	**.37**
Grade 4	62	.14	.88	**.62**	.16	.63	62	.11	.61	.38	.10	**.39**
Grade 5	59	.31	.89	**.62**	.13	.64	59	.20	.68	.40	.10	**.39**
Grade 6	62	.34	.79	**.59**	.12	.61	62	.20	.65	.43	.09	**.43**
Grade 7	68	.25	.82	**.58**	.11	.57	68	.25	.57	.41	.09	**.42**
Grade 8	68	.23	.85	**.61**	.15	.63	67	.17	.68	.43	.10	**.43**

Note. Mean P-Value target between 0.30 and 0.85, median point-biserial target >= 0.20.

Mean P-Values, median point-biserial correlations, and the number of items went up in most grades, for spring 2017. The greatest improvement was in the point-biserial correlations in which almost all items surpass the .20 cutoff.

Discussion

We conducted an evaluation of DRC's test construction processes for SC Ready and EOCEP Algebra 1 assessments. Our evaluation is based on available documentation and conversation with key test construction staff from SCDE and DRC. It is likely that we did not capture all important information within the test construction processes. In addition, we acknowledge that our review of the metabank data was conducted prior to DRC's completion of their final data quality checks.[8] We plan to conduct at least one site visit to observe one or more processes related to test construction prior to our second report that will further explore test construction processes relative to the EOCEP Biology and English 1 tests.

Several aspects of the EOCEP Algebra 1 and SC Ready Math and ELA exam forms assembly procedures were reviewed. We generally found the processes used to develop test forms for the SC Ready ELA/math and EOCEP Algebra 1 tests adhere to industry best practices. We found that that some or all of the key aspects of the relevant Joint *Standards* are met. For aspects of the relevant *Joint Standards* that were not met, there were some similarities, across Algebra 1 and SC Ready, in the areas where the *Joint Standards* were not met. Although it was revealed during a phone interview on March 1, 2017 with staff from DRC and SCDE, the documents we received (or looked for online) did not mention procedures for conducting mode DIF between paper and online items; the movement toward universal online test administration will eventually remove the need for mode DIF studies. However, there are still some forms administered using paper-and-pencil. Thus, it is worth consideration that these procedures continue. Regardless of the future use, if mode DIF is conducted currently, these procedures should be outlined in the documentation.

The psychometric guidelines for EOCEP Algebra 1 and SC Ready were similar but perhaps could benefit by becoming more similar. The Algebra 1 target P-Value was a single value, which makes it difficult to determine if a test form "meets" that guideline. In the SC Ready guidelines, the mean P-Value for the form is specified to fall within a range, which makes it easier to determine if the test form is "on target." We suggest providing a range of P-Values for Algebra 1, perhaps around the .65 value already specified. We also suggest adding a minimum point-biserial correlation value to the Algebra 1 specification like there is for SC Ready. Items with very low discrimination parameters do not help differentiate high and low examinees well and should be avoided unless necessary.

[8] Email communication from Shar Moseng at DRC on February 24, 2017. DRC provided metadata to HumRRO prior to its completion of their final data quality checks to ensure that HumRRO could meet the deliverable deadline for this report.

Chapter 5: Summary and Interim Recommendations

This is the first of three reports that HumRRO will produce as part of its comprehensive evaluation of the South Carolina educational assessments. Subsequent reports will address additional and related aspects of test development and implementation, building toward a more complete understanding of the quality of the South Carolina assessments.

Based on the findings from the three initial tasks included in this first interim report, we found that the SC Ready ELA/math and EOCEP Algebra 1 tests generally adhere to industry best practices, with some areas noted for improvement. As a summary, we outline the key findings for each test and offer interim recommendations to improve ongoing processes and procedures. Each interim recommendation is accompanied by a priority rating. Table 5.1 presents the classification schema applied to the interim recommendations. HumRRO will provide final recommendations, summarizing across all tasks, to the EOC in the third and final evaluation report.

Table 5.1. Priority Rating Codes for Interim Recommendations

Priority Rating	Description of Priority Rating
Urgent	Definitely needs to be addressed; should be considered and addressed immediately.
High	Needs to be addressed; should be considered and addressed as soon as possible.
Medium	Should be considered and possibly addressed.
Low	Might be considered if time allows.

Algebra 1

Item Development (Chapter 2)

Finding 1.1. The processes used to develop items for the EOCEP Algebra 1 tests adhere to industry best practices. Items undergo a multi-step process that includes review by expert judges regarding content and cognitive complexity alignment, as well as sensitivity and fairness.

Finding 1.2. Universal design principles are referenced, but different documents provide different details on how to fulfill these principles. Inconsistency and lack of detail was found in the presentation of check points (specific points of guidance for item developers) across documents, with missing check points to address the accessibility needs of all students. We did not see documents that clearly describe how empirical results and expert judgements are appropriately used to review items and scoring guides. It is difficult to judge whether empirical results and expert judgements are appropriately used when reviewing items and scoring criteria.

Finding 1.3. Documentation about the item management system (IDEAS) was not found. No documentation was provided on how items are stored, how item review feedback is saved, or how changes are tracked in the system. Currently, preliminary item information is only obtained from field testing.

Finding 1.4. Item development documentation does not clearly specify the intended uses of the test scores.

> **Interim Recommendation 1.1. Improve item development processes (High).** Item development processes could be improved in several ways. Aspects of the item

development process to improve include expanded background information for item developers/reviewers on the goals of the assessment for which items are developed, and expanded item review checklists with clear guidance for evaluating item content, difficulty, clarity, and accuracy. Record keeping of the item development process should also be uniformly implemented and consistently documented. Cross-referencing should be added to item development documents to ensure easy access to all available information. Processes and documentation should clearly and consistently implement universal design principles. More detailed information about the background and characteristics of expert judges and quality assurance staff should be captured and documented.

Interim Recommendation 1.2. Continue to expand the available documentation describing processes and procedures for item development (High). Standard 7.4 of the *Joint Standards* highlights the importance of detailed documentation of all test development procedures. We found several areas where detailed descriptions were lacking in the available documentation, or where no formal documentation was available. There were also instances where inconsistent guidance was provided across documents. Although we were able to clarify our understanding through web searches and phone interviews with relevant staff, the assessment system could be improved through continued expansion of the formal documentation that is available. We recommend that DRC compile a technical manual that documents all aspects of item development.

Interim Recommendation 1.3. Consider adding item tryouts or cognitive labs to the item development process (Medium). Item tryouts, which use a smaller number of students than field testing, and which occur earlier in item development when changes can be made more easily, should be considered for subsequent item development. This would be particularly useful for developing novel item types.

Content Alignment (Chapter 3)

A content alignment study was conducted on two EOCEP Algebra 1 test forms (Spring 2017 and Winter 2016-17) to investigate how well the items align to the SCCCRS. Independent, external content experts served as panelists for this alignment workshop. The findings and recommendations follow.

Finding 2.1. Overall, the alignment results provide support for the content validity of the EOCEP Algebra 1 test. On average, panelists rated approximately 90% of the items as “fully aligned” to the SCCRS. We also investigated alignment using the Webb alignment methodology (1997, 1999, 2005). The Webb alignment criteria were investigated at the level of the content strand and at the level of the key concept. There was one Webb alignment criterion (categorical concurrence) that received a “partially aligned” rating at the content strand level and a “weakly aligned” rating at the key concept level on the Webb rating scale; *however*, the categorical concurrence criterion is intended to inform the minimum number of items required for each reporting category. Because SCDE does not report scores at the level of the content strand or at the level of the key concept, the lower alignment ratings on Webb’s categorical concurrence criterion should be of no concern for the SCDE. The EOCEP Algebra I test meets the remaining Webb criteria for appropriate item difficulty (depth-of-knowledge) and coverage of the standards (range-of-knowledge correspondence and balance-of-knowledge representation). Finally, at the end of the workshop panelists were asked to provide a final holistic rating of the overall alignment between the EOCEP Algebra 1 test and the SCCCRS. Four of the five panelists (including the nationally recognized content expert) rated the overall alignment as “good.”

Finding 2.2. As indicated in Finding 2.1, Webb's depth-of-knowledge consistency criterion was attained per the Webb rating scale. The depth-of-knowledge consistency criterion indicates whether there is consistency between the complexity of knowledge required by the standards and the complexity of knowledge required to correctly answer the items linked to those standards. Webb's suggested minimum for this criterion is that at least 50% of the items should have complexity ratings at or above the level of the corresponding standard. All the content strands meet this alignment criterion. At the level of the key concept, one of the four key concepts—Structure and Expressions—fell considerably short of meeting this criterion for both the Spring 2017 form and the Winter 2016-17 form. This finding suggests that the cognitive complexity required to correctly answer the items linked to the standards within this key concept is, on average, lower than the cognitive complexity required by the standards.

Finding 2.3. In addition to the Webb alignment criteria, we also compared the mean number of items linked to each content strand by the expert panelists to the target number of items in the test blueprint for each content strand. The mean number of items linked to each content strand was within the range specified in the test blueprint for all content strands, except for the Number and Quantity content strand, for which the mean number of linked items was 4.8 (*SD = 0.45)* and 4.6 (*SD* = 0.55) for the spring and winter forms, respectively, which was slightly below the target of 5 – 9 items specified in the test blueprint.

Finding 2.4. The independent, external reviewers found an overwhelming majority of Algebra 1 items to be free of any issues related to clarity, accuracy, grade-level appropriateness, and biased content/presentation. There were only two items on which at least three of the five panelists expressed concerns about the items' clarity. All panelists' comments on items have been provided to DRC, separately from this report, for their consideration.

Finding 2.5. Three of the five panelists mentioned the limitations of multiple-choice tests such as the EOCEP Algebra 1 test for providing useful information about the South Carolina College- and Career-Ready Mathematical Process Standards, or to support research-based instruction. Four of the five panelists also mentioned that some items might be biased towards students with access to, and familiarity with, graphing calculators, though one panelist stated that this is common to most math tests.

Interim Recommendation 2.1. Monitor the cognitive complexity of the items intended to measure the Building Functions key concept (Medium). Consider enhancing the cognitive complexity required to answer the items intended to measure the Structure and Expressions key concept to ensure that there is consistency between the level of cognitive complexity required by the standards that comprise this key concept and the cognitive complexity required to correctly answer the items that measure this key concept. If any reporting were to be considered at the key concept level, this recommendation would become a higher priority.

Interim Recommendation 2.2. Continue to monitor the content representativeness of the item pool (Medium). All test items are linked to a content standard, and evidence from the alignment study indicates appropriate numbers of items for all content strands, with the possible exception of the Number and Quantity content strand. The SCDE may want to consider including an additional item or two to the measure the Number and Quantity content strand to ensure that the EOCEP Algebra 1 test is meeting the intent of the test blueprint. Also, should changes be made to reporting practices (e.g., reporting subscores), ongoing monitoring of the content standard(s) measured by items will help to ensure that there are sufficient numbers of items for such purposes.

Interim Recommendation 2.3. Consider including additional item types to the Algebra 1 test (Low). Item types other than traditional multiple choice would offer more opportunities for students to demonstrate, for example, relating problems to prior knowledge and identifying multiple paths to a solution. Such opportunities may better reflect the South Carolina College- and Career-Ready Mathematical Process Standards while also better supporting research-based instruction.

Test Construction (Chapter 4)

Finding 3.1. The processes and procedures for creating EOCEP Algebra 1 test forms generally reflect industry best practices as outlined in the *Joint Standards*.

Finding 3.2. Available documentation guiding test construction processes and procedures contains several gaps. For example, there is no mention of internal consistency reliability minimums or if this is considered when creating forms. The origin of item statistics used for test form construction (e.g., estimated during field testing or prior operational use) is not clearly stated, nor is it clear at what stage differential item functioning (DIF) is analyzed. Documentation also appears inconsistent regarding the use of classical test theory (CTT) and/or item response theory (IRT) statistics for forms assembly.

Finding 3.3. Item P-Values and point-biserial correlations associated with Algebra 1 forms administered in 2015-16 are within acceptable ranges. However, within the item bank, approximately 5% of items have P-Values below .2, and a small number of items have negative point-biserial correlations.

Interim Recommendation 3.1. Remove items with P-Values and/or point-biserial correlations outside of the acceptable ranges from the item bank (Urgent). Though item statistics are considered during form construction and previous operational test forms only contained items with CTT statistics within the acceptable ranges, removal of problematic items from the item bank would provide an extra quality assurance step. It would also provide a more accurate depiction of the strength of the available item pool and inform item development.

Interim Recommendation 3.2. Continue to expand the available documentation describing processes and procedures for test form construction (High). The content considerations of the test need to be more explicitly defined (e.g., paper/pencil vs computerized administration, procedures for replacing technology enhanced items on a paper/pencil test). The conditions of administration need to be more clearly specified (time for testing, directions, administration guidelines), and the statistical targets for test development (test length, internal consistency reliability, target P-Values, target point-biserial correlations) need to be better specified. Specifically, we recommend a range of P-Values and a minimum point-biserial correlation be specified. We recommend that DRC compile a technical manual that documents all aspects of test construction, including evidence of all studies to investigate potential sources of construct irrelevant variance.

SC Ready

Item Development (Chapter 2)

Finding 4.1. The processes used to develop items for the SC Ready ELA/math tests adhere to industry best practices. Items undergo a multi-step process that includes review by expert judges regarding content and cognitive complexity alignment, as well as sensitivity and fairness.

Finding 4.2. Universal design principles are referenced, but different documents provide different details on how to fulfill these principles. Inconsistency and lack of detail was found in the presentation of check points (specific points of guidance for item developers) across documents, with missing check points to address the accessibility needs of all students. However, we did not see documents that clearly describe how empirical results and expert judgements are appropriately used to review items and scoring guides. It is difficult to judge whether empirical results and expert judgements are appropriately used when reviewing items and scoring criteria.

Finding 4.3. Documentation about the item management system (IDEAS) was not found. No documentation was provided on how items are stored, how item review feedback is saved, or how changes are tracked in the system. Currently, preliminary item information is only obtained from field testing.

Finding 4.4. HumRRO's evaluation of a sample of items found that items generally adhered to item quality guidelines and various review feedback was incorporated to improve the quality of the items. However, we find readability and grade level appropriateness are specifically considered for the reading passages and related item stimuli as indicated in document 13, but not for math items.

Finding 4.5. Students' responses from field tests are used to refine the scoring rubrics for text dependent analysis writing prompts on the SC Ready ELA assessment. However, it is not clear how empirical data are used to review and improve scoring criteria (e.g., refine scoring guides, build training sets).

> **Interim Recommendation 4.1. Improve item development processes (High).** Item development processes could be improved in several ways. Aspects of the item development process to improve include expanded background information for item developers/reviewers on the goals of the assessment for which items are developed, and expanded item review checklists with clear guidance for evaluating item content, difficulty, clarity, and accuracy. Record keeping of the item development process should also be uniformly implemented and consistently documented. Cross-referencing should be added to item development documents to ensure easy access to all available information. Processes and documentation should clearly and consistently implement universal design principles. More detailed information about the background and characteristics of expert judges and quality assurance staff should be captured.
>
> **Interim Recommendation 4.2. Continue to expand the available documentation describing processes and procedures for item development (High).** Standard 7.4 of the *Joint Standards* highlights the importance of detailed documentation of all test development procedures. We found several areas where detailed descriptions were lacking in the available documentation, or where no formal documentation was available. There were also instances where inconsistent guidance was provided across documents. Although we were able to clarify our understanding through web searches and phone interviews with relevant staff, the assessment system could be improved through continued expansion of the formal documentation that is available. We recommend that DRC compile a technical manual that documents all aspects of item development.
>
> **Interim Recommendation 4.3. Incorporate readability and grade-level appropriateness reviews for mathematics items and associated stimuli (High).** The

reading demand of the math items and associated stimuli may introduce construct irrelevant variance and affect students' performance. Adding these reviews during item development would further support the validity of test scores.

Interim Recommendation 4.4. Consider adding item tryouts or cognitive labs to the item development process (Medium). Item tryouts, which use a smaller number of students than field testing, and which occur earlier in item development when changes can be made more easily, should be considered for subsequent item development. This would be particularly useful for developing novel item types.

Test Construction (Chapter 4)

Finding 5.1. The processes and procedures for creating test forms generally reflect industry best practices as outlined in the *Joint Standards*.

Finding 5.2. Available documentation guiding test construction processes and procedures contains some gaps. For example, we found no guidelines surrounding issues of item parameter drift or non-convergence that might occur during the post-equating process. We also found no description of how comparisons between paper and computer-based item-level data are conducted, nor mention of forms-level comparisons between paper and computer forms.

Finding 5.3. Item statistics from the item bank demonstrate improvements in the available item pool over time. Items with statistics outside of the acceptable ranges were removed between 2016 and 2017.

Interim Recommendation 5.1. Continue to expand the available documentation describing processes and procedures for test form construction (High). Documentation should be expanded to ensure complete information is available for understanding how issues such as item parameter drift and non-convergence are evaluated and addressed. We recommend that DRC compile a technical manual that documents all aspects of test construction.

Interim Recommendation 5.2. Consider continuing the analysis of mode DIF and expand the available documentation describing these procedures (Medium). Although there is a movement toward near universal online test administration, if there are paper forms administered then the analysis of any differences between paper and online forms should be conducted. Any such analyses should be described in detail in the technical documentation.

References

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

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, D.C.: Council of Chief State Schools Officers. Available: http://www.ccsso.org/Documents/TILSA%20Evaluating%20Alignment%20in%20Large-Scale%20Standards-Based%20Assessment%20Systems%20-%20FINAL.pdf.

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform. http://files.eric.ed.gov/fulltext/ED532068.pdf

Landers, R. N. (2015). Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. *The Winnower 2*:e143518.81744. DOI: 10.15200/winn.143518.81744

Porter, A. C. (2002, October). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, pp. 420-428. http://dx.doi.org/10.1037/0033-2909.86.2.420

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. (1999). *Alignment of Science and Science standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in Science and Science education*. Washington, D.C.: Council of Chief State Schools Officers.

Webb, N. L. (2005). *Webb alignment tool: Training manual*. Madison, WI: Wisconsin Center for Education Research. Available: http://www.wcer.wisc.edu/WAT/index.aspx

Appendix A: Interview Questions for Item Development Process Review

1. Usually how many rounds/levels of reviews that each item goes through? What is each round of review mainly about?

   Possible follow up questions:
   - It's mentioned in the ID document that after an item writer submit an item, an editor decides whether the item move forward to first-level editing. What the criteria are for accept, revision, or rejection?
   - It's also mentioned that the content director gives the item another review before it is submitted to the SCDE for review. What is this review mainly about?
   - The SCDE staff have the opportunity to review all EOCEP items and provide DRC with revisions prior to the content review meetings. What is SCDE staff's review mainly about? What are the revisions that provided by SCDE usually?
   - Are there external reviews?
   - Have you conducted or have any supporting research studies planned? What are these?
   - What are the internal processes to verify an item's coverage of the standards and sub- standard?
   - What are the internal processes to verify an item's difficulty and DOK?

2. In all these rounds of reviews, how feedback and suggestions are tracked and used to improve items?

   Possible follow up questions:
   - Are all review notes kept with the item? On the item card provided, there were not areas to track changes or comments from the item's review for accessibility, readability, or sensitivity. Are these maintained with item statistics?

3. During item review process, Items are reviewed for bias, fairness and sensitivity. Could you provide some details about the bias, fairness and sensitivity reviews?

   Possible follow up questions:
   - How does this bias/fairness/sensitivity review ensure that assessment materials are appropriate for students with various background and characteristics?
   - How bias, fairness and sensitivity issues are considered for subgroups? Is there documentation?
   - How are reasonably anticipated potential areas of unfairness addressed?
   - How are accommodations provided for students with disabilities checked in the system?

4. What are the criteria used to determine 1) whether an item will be accepted for the field test and 2) whether an item will be accepted for the operational test?

5. Are constructed-response items developed together with draft scoring guides in the item development process? If so, how the scoring guides are developed in the ID process?

   Possible follow up questions:
   - How are field assessment data used to refine scoring guides?
   - When are scoring validity and check sets developed? What are the processes to develop them?

6. Do some items require pretesting (e.g., cog labs, tryouts) before the field test? If so, how is that conducted?

7. How are empirical results from field tests used in the item review process? What are the criteria to modify or drop items from the operational test?

8. How are students sampled for field test or pretesting?

   Possible follow up questions:
   - Is field testing done during the operational test window?
   - Is any done outside of the window or outside of South Carolina?

9. How item writers are recruited and selected? From your document, the selection is based on qualifications such as education degree, understanding and knowledge of assessment development, and participation in the assessment-specific training workshop.

   Possible follow up questions:
   - Are the qualification, relevant experiences, and demographic characteristics of item writers or expert judges documented?
   - When are item writers removed from the item writing pool?

10. Could you provide more information about the trainings that item writers received? The documentation stated that item writing training occurs individually. Would you confirm this?

    Possible follow up questions:
    - How long is the training? Is it conducted in-person or via webex? What documents are available to item writers to support their work?
    - What are the training materials that item writers receive?
    - Are there any checks about item writers' understanding and knowledge after the training?
    - Are there any additional materials from the item writer training that you could provide (e.g., Bias, Fairness, and Sensitivity Guidelines, Principles of Universal Design)?

Appendix B: Comparison of the SCCCRS for and the Kentucky Academic Standards for Algebra 1

South Carolina Standard	Kentucky Standard
A1.AAPR.1 Add, subtract, and multiply polynomials and understand that polynomials are closed under these operations. (Limit to linear; quadratic.)	**A.APR.1** Understand that polynomials form a system analogous to the integers, namely, they are closed under the operations of addition, subtraction, and multiplication; add, subtract, and multiply polynomials.
A1.ACE.1 Create and solve equations and inequalities in one variable that model real-world problems involving linear, quadratic, simple rational, and exponential relationships. Interpret the solutions and determine whether they are reasonable. (Limit to linear; quadratic; exponential with integer exponents.)	**A.CED.1** Create equations and inequalities in one variable and use them to solve problems. *Include equations arising from linear and quadratic functions, and simple rational and exponential functions.*
A1.ACE.2 Create equations in two or more variables to represent relationships between quantities. Graph the equations on coordinate axes using appropriate labels, units, and scales. (Limit to linear; quadratic; exponential with integer exponents; direct and indirect variation.)	**A.CED.2** Create equations in two or more variables to represent relationships between quantities, graph equations on a coordinate axes with labels and scales.
A1.ACE.4 Solve literal equations and formulas for a specified variable including equations and formulas that arise in a variety of disciplines.	**A.CED.4** Rearrange formulas to highlight a quantity of interest, using the same reasoning as in solving equations. *For example, rearrange Ohm's law V = IR to highlight resistance R.*
A1.AREI.1 Understand and justify that the steps taken when solving simple equations in one variable create new equations that have the same solution as the original.	**A.REI.1** Explain each step in solving a simple equation as following from the equality of numbers asserted at the previous step, starting from the assumption that the original equation has a solution. Construct a viable argument to justify a solution method.
A1.AREI.3 Solve linear equations and inequalities in one variable, including equations with coefficients represented by letters.	**A.REI.3** Solve linear equations and inequalities in one variable, including equations with coefficients represented by letters.

South Carolina Standard	Kentucky Standard
A1.AREI.4 Solve mathematical and real-world problems involving quadratic equations in one variable. a. Use the method of completing the square to transform any quadratic equation in *x* into an equation of the form $(x - h)^2 = k$ that has the same solutions. Derive the quadratic formula from this form. b. Solve quadratic equations by inspection, taking square roots, completing the square, the quadratic formula and factoring, as appropriate to the initial form of the equation. Recognize when the quadratic formula gives complex solutions and write them as $a + bi$ for real numbers a and b. (Limit to non-complex roots.)	**A.REI.4a** Solve quadratic equations in one variable. a. Use the method of completing the square to transform any quadratic equation in x into an equation of the form (x-p)2=q that has the same solutions. Derive the quadratic formula from this form. b. Solve quadratic equations by inspection (e.g., for x2 = 49), taking square roots, completing the square, the quadratic formula and factoring, as appropriate to the initial form of the equation. Recognize when the quadratic formula gives complex solutions and write them as a ± bi for real numbers a and b.
A1.AREI.5 Justify that the solution to a system of linear equations is not changed when one of the equations is replaced by a linear combination of the other equation.	**A.REI.5** Prove that, given a system of two equations in two variables, replacing one equation by the sum of that equation and a multiple of the other produces a system with the same solutions.
A1.AREI.6 Solve systems of linear equations algebraically and graphically focusing on pairs of linear equations in two variables. a. Solve systems of linear equations using the substitution method. b. Solve systems of linear equations using linear combination.	**A.REI.6** Solve systems of linear equations exactly and approximately (e.g., with graphs), focusing on pairs of linear equations in two variables.
A1.AREI.10 Explain that the graph of an equation in two variables is the set of all its solutions plotted in the coordinate plane.	**A.REI.10** Understand that the graph of an equation in two variables is the set of all its solutions plotted in the coordinate plane, often forming a curve (which could be a line).

South Carolina Standard	Kentucky Standard
A1.AREI.11 Solve an equation of the form $f(x) = g(x)$ graphically by identifying the x-coordinate(s) of the point(s) of intersection of the graphs of $y = f(x)$ and $y = g(x)$. (Limit to linear; quadratic; exponential.)	**A.REI.11** Explain why the x-coordinates of the points where the graphs of the equations $y = f(x)$ and $y = g(x)$ intersect are the solutions of the equation $f(x) = g(x)$; find the solutions approximately, e.g., using technology to graph the functions, make tables of values, or find successive approximations. Include cases where $f(x)$ and/or $g(x)$ are linear, polynomial, rational, absolute value, exponential, and logarithmic functions.* (Modeling standard)
A1.AREI.12 Graph the solutions to a linear inequality in two variables.	**A.REI.12** Graph the solutions to a linear inequality in two variables as a half-plane (excluding the boundary in the case of a strict inequality), and graph the solution set to a system of linear inequalities in two variables as the intersection of the corresponding half-planes.
A1.ASE.1 Interpret the meanings of coefficients, factors, terms, and expressions based on their real-world contexts. Interpret complicated expressions as being composed of simpler expressions. (Limit to linear; quadratic; exponential.)	**A.SSE.1a** Interpret expressions that represent a quantity in terms of its context.*(*Modeling standard) a. Interpret parts of an expression, such as terms, factors, and coefficients. b. Interpret complicated expressions by viewing one or more of their parts as a single entity. For example, interpret as the product of P and a factor not depending on P.
A1.ASE.2 Analyze the structure of binomials, trinomials, and other polynomials in order to rewrite equivalent expressions.	**A.SSE.2** Use the structure of an expression to identify ways to rewrite it. *For example, see x4 – y4 as (x2)2 – (y2)2, thus recognizing it as a difference of squares that can be factored as (x2 – y2)(x2 + y2).*

South Carolina Standard	Kentucky Standard
A1.ASE.3 Choose and produce an equivalent form of an expression to reveal and explain properties of the quantity represented by the expression. a. Find the zeros of a quadratic function by rewriting it in equivalent factored form and explain the connection between the zeros of the function, its linear factors, the *x*-intercepts of its graph, and the solutions to the corresponding quadratic equation.	**A.SSE.3a** Choose and produce an equivalent form of an expression to reveal and explain properties of the quantity represented by the expression.*(Modeling standard) a. Factor a quadratic expression to reveal the zeros of the function it defines.
A1.FBF.3 Describe the effect of the transformations $kf(x)$, $f(x) + k$, $f(x + k)$, and combinations of such transformations on the graph of $y= f(x)$ for any real number k. Find the value of k given the graphs and write the equation of a transformed parent function given its graph. (Limit to linear; quadratic; exponential with integer exponents; vertical shift and vertical stretch.)	**F.BF.3** Identify the effect on the graph of replacing $f(x)$ by $f(x) + k$, $k f(x)$, $f(kx)$, and $f(x + k)$ for specific values of k (both positive and negative); find the value of k given the graphs. Experiment with cases and illustrate an explanation of the effects on the graph using technology. *Include recognizing even and odd functions from their graphs and algebraic expressions for them.*
A1.FIF.1 Extend previous knowledge of a function to apply to general behavior and features of a function. a. Understand that a function from one set (called the domain) to another set (called the range) assigns to each element of the domain exactly one element of the range. b. Represent a function using function notation and explain that $f(x)$ denotes the output of function f that corresponds to the input x. c. Understand that the graph of a function labeled as f is the set of all ordered pairs (x,y) that satisfy the equation $y = f(x)$.	**F.IF.1** Understand that a function from one set (called the domain) to another set (called the range) assigns to each element of the domain exactly one element of the range. If f is a function and x is an element of its domain, then $f(x)$ denotes the output of f corresponding to the input x. The graph of f is the graph of the equation $y = f(x)$.
A1.FIF.2 Evaluate functions and interpret the meaning of expressions involving function notation from a mathematical perspective and in terms of the context when the function describes a real-world situation.	**F.IF.2** Use function notation, evaluate functions for inputs in their domains, and interpret statements that use function notation in terms of a context.

South Carolina Standard	Kentucky Standard
A1.FIF.4 Interpret key features of a function that models the relationship between two quantities when given in graphical or tabular form. Sketch the graph of a function from a verbal description showing key features. Key features include intercepts; intervals where the function is increasing, decreasing, constant, positive, or negative; relative maximums and minimums; symmetries; end behavior and periodicity. (Limit to linear; quadratic; exponential.)	**F.IF.4** For a function that models a relationship between two quantities, interpret key features of graphs and tables in terms of the quantities, and sketch graphs showing key features given a verbal description of the relationship. *Key features include: intercepts; intervals where the function is increasing, decreasing, positive, or negative; relative maximums and minimums; symmetries; end behavior; and periodicity.**(*Modeling standard)
A1.FIF.5 Relate the domain and range of a function to its graph and, where applicable, to the quantitative relationship it describes. (Limit to linear; quadratic; exponential.)	**F.IF.5** Relate the domain of a function to its graph and, where applicable, to the quantitative relationship it describes. *For example, if the function h(n) gives the number of person-hours it takes to assemble n engines in a factory, then the positive integers would be an appropriate domain for the function.**(*Modeling standard)
A1.FIF.6 Given a function in graphical, symbolic, or tabular form, determine the average rate of change of the function over a specified interval. Interpret the meaning of the average rate of change in a given context. (Limit to linear; quadratic; exponential.)	**F.IF.6** Calculate and interpret the average rate of change of a function (presented symbolically or as a table) over a specified interval. Estimate the rate of change from a graph.*(Modeling standard)
A1.FIF.7 Graph functions from their symbolic representations. Indicate key features including intercepts; intervals where the function is increasing, decreasing, positive, or negative; relative maximums and minimums; symmetries; end behavior and periodicity. Graph simple cases by hand and use technology for complicated cases. (Limit to linear; quadratic; exponential only in the form $y = a^x + k$.)	**F.IF.7a** Graph functions expressed symbolically and show key features of the graph, by hand in simple cases and using technology for more complicated cases.*(Modeling standard) a. Graph linear and quadratic functions and show intercepts, maxima, and minima.

South Carolina Standard	Kentucky Standard
A1.FIF.8 Translate between different but equivalent forms of a function equation to reveal and explain different properties of the function. (Limit to linear; quadratic; exponential.) a. Use the process of factoring and completing the square in a quadratic function to show zeros, extreme values, and symmetry of the graph, and interpret these in terms of a context.	**F.IF.8a** Write a function defined by an expression in different but equivalent forms to reveal and explain different properties of the function. a. Use the process of factoring and completing the square in a quadratic function to show zeros, extreme values, and symmetry of the graph, and interpret these in terms of a context. b. Use the properties of exponents to interpret expressions for exponential functions. For example: identify percent rate of change in functions such as y= (1.02)t, y=(.97)t, y=(1.01)12t, y=(1.2)t/10, and classify them as representing exponential growth or decay.
A1.FIF.9 Compare properties of two functions given in different representations such as algebraic, graphical, tabular, or verbal. (Limit to linear; quadratic; exponential.)	**F.IF.9** Compare properties of two functions each represented in a different way (algebraically, graphically, numerically in tables, or by verbal descriptions). *For example, given a graph of one quadratic function and an algebraic expression for another, say which has the larger maximum.*

South Carolina Standard	Kentucky Standard
A1.FLQE.1 Distinguish between situations that can be modeled with linear functions or exponential functions by recognizing situations in which one quantity changes at a constant rate per unit interval as opposed to those in which a quantity changes by a constant percent rate per unit interval. a. Prove that linear functions grow by equal differences over equal intervals and that exponential functions grow by equal factors over equal intervals.	**F.LE.1a** Distinguish between situations that can be modeled with linear functions and with exponential functions. a. Prove that linear functions grow by equal differences over equal intervals; and that exponential functions grow by equal factors over equal intervals. b. Recognize situations in which one quantity changes at a constant rate per unit interval relative to another. c. Recognize situations in which a quantity grows or decays by a constant percent rate per unit interval relative to another.
A1.FLQE.2 Create symbolic representations of linear and exponential functions, including arithmetic and geometric sequences, given graphs, verbal descriptions, and tables. (Limit to linear; exponential.)	**F.LE.2** Construct linear and exponential functions, including arithmetic and geometric sequences, given a graph, a description of a relationship, or two input-output pairs (include reading these from a table).
A1.FLQE.3 Observe using graphs and tables that a quantity increasing exponentially eventually exceeds a quantity increasing linearly, quadratically, or more generally as a polynomial function.	**F.LE.3** Observe using graphs and tables that a quantity increasing exponentially eventually exceeds a quantity increasing linearly, quadratically, or (more generally) as a polynomial function.
A1.FLQE.5 Interpret the parameters in a linear or exponential function in terms of the context. (Limit to linear.)	**F.LE.5** Interpret the parameters in a linear or exponential function in terms of a context.
A1.NQ.1 Use units of measurement to guide the solution of multi-step tasks. Choose and interpret appropriate labels, units, and scales when constructing graphs and other data displays.	**N.Q.1** Use units as a way to understand problems and to guide the solution of multi-step problems; choose and interpret units consistently in formulas; choose and interpret the scale and the origin in graphs and data displays.
A1.NQ.2 Label and define appropriate quantities in descriptive modeling contexts.	**N.Q.2** Define appropriate quantities for the purpose of descriptive modeling.

South Carolina Standard	Kentucky Standard
A1.NQ.3 Choose a level of accuracy appropriate to limitations on measurement when reporting quantities in context.	**N.Q.3** Choose a level of accuracy appropriate to limitations on measurement when reporting quantities.
A1.NRNS.1 Rewrite expressions involving simple radicals and rational exponents in different forms.	**N.RN.2** Rewrite expressions involving radicals and rational exponents using the properties of exponents.
A1.NRNS.2 Use the definition of the meaning of rational exponents to translate between rational exponent and radical forms.	**N.RN.1** Explain how the definition of the meaning of rational exponents follows from extending the properties of integer exponents to those values, allowing for a notation for radicals in terms of rational exponents. *For example, we define 51/3 to be the cube root of 5 because we want (51/3)3 = 5(1/3)3 to hold, so (51/3)3 must equal 5.*
A1.NRNS.3 Explain why the sum or product of rational numbers is rational; that the sum of a rational number and an irrational number is irrational; and that the product of a nonzero rational number and an irrational number is irrational.	**N.RN.3** Explain why the sum or product of two rational numbers is rational; that the sum of a rational number and an irrational number is irrational; and that the product of a nonzero rational number and an irrational number is irrational.
A1.SPID.6 Using technology, create scatterplots and analyze those plots to compare the fit of linear, quadratic, or exponential models to a given data set. Select the appropriate model, fit a function to the data set, and use the function to solve problems in the context of the data.	**S.ID.6a** Represent data on two quantitative variables on a scatter plot, and describe how the variables are related. a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear and exponential models. b. Informally assess the fit of a function by plotting and analyzing residuals. c. Fit a linear function for a scatter plot that suggests a linear association.
A1.SPID.7 Create a linear function to graphically model data from a real-world problem and interpret the meaning of the slope and intercept(s) in the context of the given problem.	**S.ID.7** Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.
A1.SPID.8 Using technology, compute and interpret the correlation coefficient of a linear fit.	**S.ID.8** Compute (using technology) and interpret the correlation coefficient of a linear fit.

South Carolina Alignment Study
Panelist Instructions

	Rating Task	Documents Needed	File Format
1	Standards Ratings - Consensus	(1) Panelist Instructions	Print copy
		(2) SCCCR Algebra 1 Test Blueprint	Print copy
		(3) SCCCR Standards for High School	Print copy
		(4) Algebra 1 Standards Rating Form	Print copy
		(5) DOK Definitions for Mathematics	Print copy
		(6) Hess' Cognitive Rigor Matrix	Print copy
		(7) Algebra 1 Standards Rating Form	Excel (Group lead only)
2	Item Ratings - Independent	(1) Panelist Instructions	Print copy
		(2) SCCCR Algebra 1 Test Blueprint	Print copy
		(3) Algebra 1 EOCEP Items	Print copy
		(4) SC Item Rating Form- Algebra 1	Excel
		(5) DOK Definitions for Mathematics	
		(6) Hess' Cognitive Rigor Matrix	Print copy
3	Debriefing/Evaluation	(1) Debriefing/Evaluation Form	Print copy

Prior to alignment steps:

1. Introductions
2. Review all of the materials:
    a. Panelist Instructions
    b. SCCCR Algebra 1 Test Blueprint
    c. SCCCR Standards for High School
    d. Algebra 1 Standards Rating Form
    e. DOK Definitions for Mathematics
    f. Hess' Cognitive Rigor Matrix
    g. SC Item Rating Form- Algebra 1
3. Additional documents will be handed out as needed
    a. Non-disclosure agreement
    b. Debriefing and Evaluation form

Task 1 SCCCR Algebra 1 Standards Rating (Consensus)

Task preparation:

1. Facilitator will introduce the task.
2. Documents needed are:
    a. Panelist Instructions
    b. SCCCR Algebra 1 Test Blueprint
    c. SCCCR Standards for High School
    d. Algebra 1 Standards Rating Form
    e. DOK Definitions for Mathematics
    f. Hess' Cognitive Rigor Matrix

Conduct Task:

1. The facilitator will ask for a volunteer from your group to help them move the group through the task and enter the group's consensus rating for each content objective into the electronic spreadsheet.

2. Using the DOK definitions and Hess' Cognitive Rigot Matrix, everyone will rate the depth of knowledge of the first few objectives individually, record their ratings on the paper rating form, and discuss them as a group. One group member will enter the final consensus rating for each objective into the electronic spreadsheet. The rules for reaching consensus are:
    a. **If the group doesn't fully agree, then majority rules.**
    b. **If there is an exact split between group members, then the higher level prevails.**
3. Continue until all objectives for all grades have been completed
4. Review the SCCCR Standards for High School document.
5. Discuss with the group if the standards included in the test blueprint adequately reflect the intent of the South Carolina College- and Career-Ready (SCCCR) Standards for High School and the SCCCR Mathematical Process Standards.
6. Enter the group's consensus rating (Yes or No) and any related comments in the electronic spreadsheet.

Task 2 Rate Algebra 1 Items

Task Preparation:

1. The facilitator will explain the process for this task and have everyone open the rating form on their laptop.
    a. Locate the file, provided by the facilitator, on the desktop, double click to open.
    b. "Save As" file name by first adding **underscore and your 3 initials** to the file name (e.g., SC Item Rating Form- Algebra 1_**ymn**).
2. Rating form review:
    a. The facilitator will talk discuss each column.
        i. Columns A & B include the item sequence and item identifier.
        ii. Column C, enter DOK level that best represents the cognitive demand of the item.
        iii. Column D, specifies the content objective currently linked to the item.
        iv. Column E, determine the level of quality content match between the item and the objective.
        v. Column F, enter an alternative to the content objective listed in column E, or list a secondary content objective if you feel that the item measures another content objective equally well.
        vi. Column G, describe the content that the item measures which is not part of the primary objective indicated.
        vii. Column H, enter 'N' if item is not presented in a clear manner.
        viii. Column I, enter 'N' if item contains inaccurate content.
        ix. Column J, enter 'N' if item is not grade-level appropriate.
        x. Column K, enter 'N' if item does not support research-based instruction.
        xi. Column L, enter 'N' if item reflects bias against particular subgroups in its content or presentation.
        xii. Column M, provide explanation for any 'N' ratings in columns H-L.

Conduct Task:

1. Rate one item independently and then discuss ratings with group. You do NOT need to change your ratings in response to the group discussion, but you may choose to do so.
2. After the group is sufficiently calibrated (2-3 items), you will work independently until the task has been completed for all test items.

Task 3 Debriefing/Evaluation

Conduct Task:

1. The facilitator will hand out the Debriefing/Evaluation Form.
2. Complete the form (front and back) and insert it into the envelope provided by the facilitator.

DOK Definitions for Mathematics

Level 1 (recall): Level 1 includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level.

With regard to items, students may be asked to calculate or solve by a simple formula. A student answering a Level 1 item either knows the answer or does not: that is, the answer does not need to be "figured out" or "solved."

Standards objectives or items at this level may include words such as *recall*, *recognize*, *use, measure, and identify*.

Examples: Solve a one-step problem, represent math relationships in words, pictures, or symbols, or locate points on a grid or number line.

Level 2 (skill/concept): Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in level 1. Students are required to make some decisions as to how to approach the question or problem. These actions imply more than one step. Caution is warranted in interpreting Level 2 as ONLY skills and exclude cognitive processing such as visualization and using probability.

With regard to items, involves students to make some decisions as to how to approach the problem or activity, interpreting information from a simple graph, or classifying/organizing data.

Standards objectives or items at this level may include words such as *estimate, make observations, display, classify*, *organize*, and *collect*, *display*, or *compare data*.

Examples: determine the first step needed to solve this problem or organize or display data in tables, graphs, and charts.

Level 3 (strategic thinking): Level 3 requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity does not result only from possible multiple answers, but because a multi-step task requires more demanding reasoning.

With regard to items, requiring students to explain their thinking is at Level 3 such as an activity that has more than one possible answer and requires students to justify the response they give.

Standards objectives or items at this level may include words such as *interpret*, *analyze*, *verify*, *justify*, and *cite evidence*.

Examples: solve non-routine problems, determine which data should be used from this graph to solve problem, describe what the data in the graph indicate.

Level 4 (extended thinking): Level 4 requires complex reasoning, experimental design and planning, and could require an extended period of time for carrying out the multiple steps of an assessment item. The cognitive demand is high and complex by making several connections.

Standards objectives or items at this level may include words such as *analyze*, *synthesize*, and *evaluate*.

Examples: analyze and apply multisource information to explain the results of the data displayed, critique how similar problems were solved using different approaches, design a mathematical model to inform and solve a practical or abstract situation.

Algebra 1 Standards Rating Form			
Standard Code	DOK	Do the standards included in the test blueprint adequately reflect the intent of the South Carolina College- and Career-Ready (SCCCR) Standards for High School and the SCCCR Mathematical Process Standards?	Comment
A1.AAPR.1			
A1.ACE.1			
A1.ACE.2			
A1.ACE.4			
A1.AREI.1			
A1.AREI.3			
A1.AREI.4			
A1.AREI.5			
A1.AREI.6			
A1.AREI.10			
A1.AREI.11			
A1.AREI.12			
A1.ASE.1			
A1.ASE.2			
A1.ASE.3			
A1.FBF.3			
A1.FIF.1			
A1.FIF.2			
A1.FIF.4			
A1.FIF.5			
A1.FIF.6			
A1.FIF.7			
A1.FIF.8			
A1.FIF.9			
A1.FLQE.1			
A1.FLQE.2			
A1.FLQE.3			
A1.FLQE.5			
A1.NQ.1			
A1.NQ.2			
A1.NQ.3			
A1.NRNS.1			
A1.NRNS.2			
A1.NRNS.3			
A1.SPID.6			
A1.SPID.7			
A1.SPID.8			

South Carolina Item Rating Sheet: Algebra 1										
Item Number	**Enter Depth of Knowledge (DOK) Rating**	**South Carolina Standard (Primary)**	**Quality of Content Match**	**South Carolina Standard (Secondary)**	**Explanation**	**Clarity of Presentation**	**Accuracy of Content**	**Grade-Level Appropriateness**	**Unbiased Content/Presentation**	**Explanation**
	1 - Recall 2 - Skill/Concept 3 - Strategic Thinking 4 - Extended Thinking	Content objective currently linked to item.	0 - No match 1 - Partially matched 2 - Fully matched	List an alternate or secondary content objective, if appropriate.	If Quality of Match rating is '0', decribe content in the item that is not found in any standard.	Enter 'N' if the item is not presented in a clear manner.	Enter 'N' if the item contains inaccurate content.	Enter 'N' if the item is not grade-level appropriate.	Enter 'N' if the item reflects bias against particular subgroups in its content or presentation.	If 'N' was entered into any column H-L, provide an explanation of why for each 'N' rating.
Winter										
1		A1.SPID.7								
2		A1.AREI.3								
3		A1.ASE.2								
4		A1.FLQE.1								
5		A1.ASE.1								
6		A1.AREI.1								
7		A1.AREI.5								
8		A1.FIF.4								
9		A1.NRNS.1								
10		A1.FLQE.2								
11		A1.NQ.3								
12		A1.AREI.6								
13		A1.FIF.8a								
14		A1.FLQE.1								
15		A1.FLQE.2								
16		A1.FLQE.5								
17		A1.FIF.7								
18		A1.AREI.6								
19		A1.ACE.1								
20		A1.AREI.4a								
21		A1.ASE.3								
22		A1.NRNS.1								
23		A1.NRNS.3								
24		A1.AREI.6a								
25		A1.SPID.6								

Debriefing: South Carolina Alignment Study

Did the items you reviewed generally represent the content in the objectives to which they were linked? Are there elements of the blueprint that you feel were not adequately reflected in the test items?

Did the items generally reflect a range of cognitive/performance expectations appropriate for students at the test grade level? If not, did item DOK levels tend to be too high or too low?

Were the items you reviewed generally clear, accurate, grade-level appropriate, supportive of research-based instruction, and free of biased content/presentation? If not, please briefly summarize your concerns about item quality.

What is your general opinion of the alignment between the Algebra 1 items and content objectives?

- ☐ Perfect alignment
- ☐ Good alignment
- ☐ Needs some improvement
- ☐ Needs major improvement (please explain specifically what that would be)
- ☐ Not aligned in any way (please explain and provide some examples)

Comments:

Evaluation: Alignment Review Training and Procedures

Please indicate your agreement by marking an 'X' in the appropriate box for each statement.

After training, I felt prepared to be a panelist.

☐ Strongly Disagree ☐ Disagree ☐ Somewhat Disagree ☐ Somewhat Agree ☐ Agree ☐ Strongly Agree

If Disagree or Strongly Disagree, suggest how it could be improved:_____

_____

HumRRO staff seemed knowledgeable of the alignment process.

☐ Strongly Disagree ☐ Disagree ☐ Somewhat Disagree ☐ Somewhat Agree ☐ Agree ☐ Strongly Agree

If Disagree or Strongly Disagree, suggest how it could be improved:_____

_____

The instructions and support documentation were clear, understandable, and useful in performing the alignment steps.

☐ Strongly Disagree ☐ Disagree ☐ Somewhat Disagree ☐ Somewhat Agree ☐ Agree ☐ Strongly Agree

If Disagree or Strongly Disagree, suggest how it could be improved:_____

_____

The paper and Excel forms were relatively easy to use to enter data.

☐ Strongly Disagree ☐ Disagree ☐ Somewhat Disagree ☐ Somewhat Agree ☐ Agree ☐ Strongly Agree

If Disagree or Strongly Disagree, suggest how it could be improved:_____

_____

Please provide any additional comments:_____

_____

_____

Appendix D: Categorical Concurrence Means and Standard Deviations

Table D-1. Categorical Concurrence: Mean Number of Items per Key Concept (Spring 2017 Form)

Key Concept	Number of Items per Key Concept		At Least Six Items per Key Concept
	Mean Items Matched	*SD*	
Arithmetic with Polynomials and Rational Expressions	2.00	0.00	No
Building Functions	1.00	0.00	No
Creating Equations	4.40	0.55	No
Interpreting Data	2.00	0.00	No
Interpreting Functions	12.00	0.00	Yes
Linear, Quadratic, and Exponential	7.00	0.00	Yes
Quantities	1.00	0.00	No
Real Number System	4.00	0.00	No
Reasoning with Equations and Inequalities	12.00	0.00	Yes
Structure and Expressions	3.40	0.55	No
Percentage of key concepts with at least six items: 30%			

Table D-2. Categorical Concurrence: Mean Number of Items per Key Concept (Winter 2016-17 Form)

Key Concept	Number of Items per Key Concept		At Least Six Items per Key Concept
	Mean Items Matched	*SD*	
Arithmetic with Polynomials and Rational Expressions	2.00	0.00	No
Building Functions	1.00	0.00	No
Creating Equations	4.00	0.00	No
Interpreting Data	2.00	0.00	No
Interpreting Functions	12.00	0.00	Yes
Linear, Quadratic, and Exponential	7.00	0.00	Yes
Quantities	1.00	0.00	No
Real Number System	3.60	0.55	No
Reasoning with Equations and Inequalities	12.40	0.55	Yes
Structure and Expressions	4.40	0.55	No
Percentage of key concepts with at least six items: 30%			

Table D-3. Categorical Concurrence: Mean Number of Items per Key Content Strand (Spring 2017 Form)

Content Strand	Number of Items per Strand		At Least Six Items per Strand
	Mean Items Matched	*SD*	
Algebra	21.80	0.84	Yes
Functions	20.00	0.00	Yes
Number and Quantity	4.80	0.45	No
Statistics and Probability	2.00	0.00	No
Percentage of content strands with at least six items: 50%			

Table D-4. Categorical Concurrence: Mean Number of Items per Key Content Strand (Winter 2016-17 Form)

Content Strand	Number of Items per Strand		At Least Six Items per Strand
	Mean Items Matched	*SD*	
Algebra	22.80	0.45	Yes
Functions	20.00	0.00	Yes
Number and Quantity	4.60	0.55	No
Statistics and Probability	2.00	0.00	No
Percentage of content strands with at least six items: 50%			

Appendix E: Depth of Knowledge Consistency Means and Standard Deviations

Table E-1. Depth of Knowledge: Mean Percent of Items per Key Concept with DOK Below, At, and Above DOK Level of Standards (Spring 2017 Form)

Key concept	Mean Items per Key Concept	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Items Below		% Items Same Level		% Items Above		
		M	*SD*	*M*	*SD*	*M*	*SD*	
Arithmetic with Polynomials and Rational Expressions	2.0	0.0	0.0	100.0	0.0	0.0	0.0	Yes
Building Functions	1.0	60.0	54.8	40.0	54.8	0.0	0.0	No
Creating Equations	4.4	35.0	15.4	60.0	15.4	5.0	11.2	Yes
Interpreting Data	2.0	0.0	0.0	60.0	22.4	40.0	22.4	Yes
Interpreting Functions	12.0	26.7	16.0	56.7	13.7	16.7	8.3	Yes
Linear, Quadratic, and Exponential	7.0	45.7	18.6	40.0	18.6	14.3	0.0	Yes
Quantities	1.0	0.0	0.0	50.0	57.7	50.0	57.7	Yes
Real Number System	4.0	0.0	0.0	65.0	13.7	35.0	13.7	Yes
Reasoning with Equations and Inequalities	12.0	28.3	4.6	56.7	14.9	15.0	10.9	Yes
Structure and Expressions	3.4	73.3	25.3	26.7	25.3	0.0	0.0	No
Percentage of key concepts with 50% of item DOK at or above standard DOK: 80%								

Table E-2. Depth of Knowledge: Mean Percent of Items per Key Concept with DOK Below, At, and Above DOK Level of Standards (Winter 2016-17 Form)

Key concept	Mean Items per Key Concept	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Items Below		% Items Same Level		% Items Above		
		M	*SD*	*M*	*SD*	*M*	*SD*	
Arithmetic with Polynomials and Rational Expressions	2.0	0.0	0.0	50.0	35.4	50.0	35.4	Yes
Building Functions	1.0	40.0	54.8	60.0	54.8	0.0	0.0	Yes
Creating Equations	4.0	50.0	0.0	40.0	13.7	10.0	13.7	Yes
Interpreting Data	2.0	40.0	22.4	30.0	44.7	30.0	27.4	Yes
Interpreting Functions	12.0	13.3	7.5	60.0	10.9	26.7	9.1	Yes
Linear, Quadratic, and Exponential	7.0	22.9	12.8	65.7	16.3	11.4	12.0	Yes
Quantities	1.0	0.0	0.0	60.0	54.8	40.0	54.8	Yes
Real Number System	3.6	0.0	0.0	83.3	15.6	16.7	15.6	Yes
Reasoning with Equations and Inequalities	12.4	22.7	7.2	62.8	4.9	14.5	3.5	Yes
Structure and Expressions	4.4	67.0	21.1	33.0	21.1	0.0	0.0	No

Percentage of key concepts with 50% of item DOK at or above standard DOK: 90%

Table E-3. Depth of Knowledge: Mean Percent of Items per Content Strand with DOK Below, At, and Above DOK Level of Standards (Spring 2017 Form)

Content Strand	Mean Items per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Items Below		% Items Same Level		% Items Above		
		M	*SD*	*M*	*SD*	*M*	*SD*	
Algebra	21.8	34.0	4.1	56.8	7.9	9.2	6.6	Yes
Functions	20.0	35.0	11.7	50.0	11.2	15.0	5.0	Yes
Number and Quantity	4.8	0.0	0.0	62.0	11.0	38.0	11.0	Yes
Statistics and Probability	2.0	0.0	0.0	60.0	22.4	40.0	22.4	Yes

Percentage of strands with 50% of item DOK at or above standard DOK: 100%

Table E-4. Depth of Knowledge: Mean Percent of Items per Content Strand with DOK Below, At, and Above DOK Level of Standards (Winter 2016-17 Form)

Content Strand	Mean Items per Strand	Depth-of-Knowledge Consistency						DOK Consistency Target Met
		% Items Below		% Items Same Level		% Items Above		
		M	*SD*	*M*	*SD*	*M*	*SD*	
Algebra	22.8	34.2	4.5	51.7	6.9	14.1	3.8	Yes
Functions	20.0	18.0	5.7	62.0	5.7	20.0	9.4	Yes
Number and Quantity	4.6	0.0	0.0	78.0	22.8	22.0	22.8	Yes
Statistics and Probability	2.0	40.0	22.4	30.0	44.7	30.0	27.4	Yes

Percentage of strands with 50% of item DOK at or above standard DOK: 100%

Appendix F: Range of Knowledge Means and Standard Deviations

Table F-1. Range-of-Knowledge: Mean Percent of Standards per Key Concept (Spring 2017 Form)

Key Concept	Number of Standards	Mean Items per Key Concept	Range of Standards			Range-of-Knowledge Target Met
			Standards with at Least One Item		% of Total Standards per Key Concept	
			M	*SD*		
Arithmetic with Polynomials and Rational Expressions	1	2.0	1.0	0.0	100.0	Yes
Building Functions	1	1.0	1.0	0.0	100.0	Yes
Creating Equations	3	4.4	3.0	0.0	100.0	Yes
Interpreting Data	3	2.0	2.0	0.0	66.7	Yes
Interpreting Functions	8	12.0	8.0	0.0	100.0	Yes
Linear, Quadratic, and Exponential	4	7.0	4.0	0.0	100.0	Yes
Quantities	3	1.0	1.0	0.0	33.3	No
Real Number System	3	4.0	3.0	0.0	100.0	Yes
Reasoning with Equations and Inequalities	8	12.0	8.0	0.0	100.0	Yes
Structure and Expressions	3	3.4	2.0	0.0	66.7	Yes
Percentage of Key Concepts with 50% of standards linked to at least one item: 90%						

Table F-2. Range-of-Knowledge: Mean Percent of Standards per Key Concept (Winter 2016-17 Form)

Key Concept	Number of Standards	Mean Items per Key Concept	Range of Standards			Range-of-Knowledge Target Met
			Standards with at Least One Item		% of Total Standards per Key Concept	
			M	*SD*		
Arithmetic with Polynomials and Rational Expressions	1	2.0	1.0	0.0	100.0	Yes
Building Functions	1	1.0	1.0	0.0	100.0	Yes
Creating Equations	3	4.0	3.0	0.0	100.0	Yes
Interpreting Data	3	2.0	2.0	0.0	66.7	Yes
Interpreting Functions	8	12.0	8.0	0.0	100.0	Yes
Linear, Quadratic, and Exponential	4	7.0	3.0	0.0	75.0	Yes
Quantities	3	1.0	1.0	0.0	33.3	No
Real Number System	3	3.6	3.0	0.0	100.0	Yes
Reasoning with Equations and Inequalities	8	12.4	8.0	0.0	100.0	Yes
Structure and Expressions	3	4.4	3.0	0.0	100.0	Yes
Percentage of Key Concepts with 50% of standards linked to at least one item: 90%						

Table F-3. Range-of-Knowledge: Mean Percent of Standards per Content Strand (Spring 2017 Form)

Content Strand	Number of Standards	Mean Items per Strand	Range of Standards			Range-of-Knowledge Target Met
			Standards with at Least One Item		% of Total Standards per Strand	
			M	*SD*		
Algebra	15	21.8	14.0	0.0	93.3	Yes
Functions	13	20.0	13.0	0.0	100.0	Yes
Number and Quantity	6	4.8	3.8	0.4	63.3	Yes
Statistics and Probability	3	2.0	2.0	0.0	66.7	Yes
Percentage of Content Strands with 50% of standards linked to at least one item: 100%						

Table F-4. Range-of-Knowledge: Mean Percent of Standards per Content Strand (Winter 2016-17 Form)

Content Strand	Number of Standards	Mean Items per Strand	Range of Standards			Range-of-Knowledge Target Met
			Standards with at Least One Item		% of Total Standards per Strand	
			M	*SD*		
Algebra	15	22.8	15.0	0.0	100.0	Yes
Functions	13	20.0	12.0	0.0	92.3	Yes
Number and Quantity	6	4.6	4.0	0.0	66.7	Yes
Statistics and Probability	3	2.0	2.0	0.0	66.7	Yes
Percentage of Content Strands with 50% of standards linked to at least one item: 100%						

Appendix G: Balance of Knowledge Means and Standard Deviations

Table G-1. Balance-of-Knowledge Representation: Mean Balance Index per Key Concept (Spring 2017 Form)

Key Concept	Standards per Key Concept	Balance-of-Knowledge Representation					Balance Index Target Met
		Mean Standards Linked with Items	Mean Items per Key Concept	Mean % of Items (of total) Linked to Key Concept	Mean Balance Index		
		M	*M*	*M*	*M*	*SD*	
Arithmetic with Polynomials and Rational Expressions	1	1.0	2.0	4.1	100.0	0.0	Yes
Building Functions	1	1.0	1.0	2.1	100.0	0.0	Yes
Creating Equations	3	3.0	4.4	9.0	84.7	1.8	Yes
Interpreting Data	3	2.0	2.0	4.1	100.0	0.0	Yes
Interpreting Functions	8	8.0	12.0	24.7	79.2	0.0	Yes
Linear, Quadratic, and Exponential	4	4.0	7.0	14.4	89.3	0.0	Yes
Quantities	3	1.0	1.0	2.1	100.0	0.0	Yes
Real Number System	3	3.0	4.0	8.2	83.3	0.0	Yes
Reasoning with Equations and Inequalities	8	8.0	12.0	24.7	75.0	0.0	Yes
Structure and Expressions	3	2.0	3.4	7.0	90.0	9.1	Yes
Total	37						
Percentage of Key Concepts with a balance of representation index of 70 or greater: 100%							

Table G-2. Balance-of-Knowledge Representation: Mean Balance Index per Key Concept (Winter 2016-17 Form)

Key Concept	Standards per Key Concept	Balance-of-Knowledge Representation					Balance Index Target Met
		Mean Standards Linked with Items	Mean Items per Key Concept	Mean % of Items (of total) Linked to Key Concept	Mean Balance Index		
		M	*M*	*M*	*M*	*SD*	
Arithmetic with Polynomials and Rational Expressions	1	1.0	2.0	4.0	100.0	0.0	Yes
Building Functions	1	1.0	1.0	2.0	100.0	0.0	Yes
Creating Equations	3	3.0	4.0	8.1	83.3	0.0	Yes
Interpreting Data	3	2.0	2.0	4.0	100.0	0.0	Yes
Interpreting Functions	8	8.0	12.0	24.3	79.2	0.0	Yes
Linear, Quadratic, and Exponential	4	3.0	7.0	14.2	90.5	0.0	Yes
Quantities	3	1.0	1.0	2.0	100.0	0.0	Yes
Real Number System	3	3.0	3.6	7.3	90.0	9.1	Yes
Reasoning with Equations and Inequalities	8	8.0	12.4	25.1	77.9	1.8	Yes
Structure and Expressions	3	3.0	4.4	8.9	79.3	5.5	Yes
Total	37						
Percentage of Key Concepts with a balance of representation index of 70 or greater: 100%							

Table G-3. Balance-of-Knowledge Representation: Mean Balance Index per Content Strand (Spring 2017 Form)

Content Strand	Standards per Strand	Balance-of-Knowledge Representation					Balance Index Target Met
		Mean Standards Linked with Items	Mean Items per Strand	Mean % of Items (of total) Linked to Strand	Mean Balance Index		
		M	*M*	*M*	*M*	*SD*	
Algebra	15	14.0	21.8	44.8	79.2	0.8	Yes
Functions	13	13.0	20.0	41.2	81.2	0.0	Yes
Number and Quantity	6	3.8	4.8	9.9	84.7	0.7	Yes
Statistics and Probability	3	2.0	2.0	4.1	100.0	0.0	Yes
Total	37						
Percentage of Content Strands with a balance of representation index of 70 or greater: 100%							

Table G-4. Balance-of-Knowledge Representation: Mean Balance Index per Content Strand (Winter 2016-17 Form)

Content Strand	Standards per Strand	Balance-of-Knowledge Representation					Balance Index Target Met
		Mean Standards Linked with Items	Mean Items per Strand	Mean % of Items (of total) Linked to Strand	Mean Balance Index		
		M	*M*	*M*	*M*	*SD*	
Algebra	15	15.0	22.8	46.2	79.5	0.8	Yes
Functions	13	12.0	20.0	40.5	80.0	0.0	Yes
Number and Quantity	6	4.0	4.6	9.3	91.0	8.2	Yes
Statistics and Probability	3	2.0	2.0	4.0	100.0	0.0	Yes
Total	37						
Percentage of Content Strands with a balance of representation index of 70 or greater: 100%							

Appendix H: Rationale for Standards used in Test Construction Review

Key			
Not Relevant	Maybe Relevant	Overarching Relevance	Relevant for a Different Task

No.	Standard	Relevant	Rationale
4.0	Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.	No	Too overarching, so it should be mentioned in the report but not in Task 3.
4.1	Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).	Yes	
4.2	In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.	Yes	
4.3	Test developers should document the rationale and supporting evidence for the administration, scoring, and reporting rules used in computer-adaptive, multistage-adaptive, or other tests delivered using computer algorithms to select items. This documentation should include procedures used in selecting items or sets of items for administration, in determining the starting point and termination conditions for the test, in scoring the test, and in controlling item exposure.	No	The Algebra 1 exam is fixed form, not computer-adaptive.
4.4	If test developers prepare different versions of a test with some change to the test specifications, they should document the content and psychometric specifications of each version. The documentation should describe the impact of differences among versions on the validity of score interpretations for intended uses and on the precision and comparability of scores.	Yes	

(continued)

Key			
Not Relevant	Maybe Relevant	Overarching Relevance	Relevant for a Different Task

No.	Standard	Relevant	Rationale
4.5	If the test developer indicates that the conditions of administration are permitted to vary from one test taker or group to another, permissible variation in conditions for administration should be identified. A rationale for permitting the different conditions and any requirements for permitting the different conditions should be documented.	Yes	
4.6	When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.	Yes	Relevant to our current study but not to Task 3.
4.7	The procedures used to develop, review, and try out items and to select items from the item pool should be documented.	Yes	
4.8	The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.	Yes	Relevant to our current study but not to Task 3.
4.9	When item or test form tryouts are conducted, the procedures used to select the sample(s) of test takers as well as the resulting characteristics of the sample(s) should be documented. The sample(s) should be as representative as possible of the population(s) for which the test is intended.	Yes	
4.10	When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented. When model-based methods (e.g., IRT) are used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.	Yes	

(continued)

Key			
Not Relevant	Maybe Relevant	Overarching Relevance	Relevant for a Different Task

No.	Standard	Relevant	Rationale
4.11	Test developers should conduct cross-validation studies when items or tests are selected primarily on the basis of empirical relationships rather than on the basis of content or theoretical considerations. The extent to which the different studies show consistent results should be documented.	No	No items for the Algebra 1 exam are chosen based on "empirical relationships" but only on content and psychometric properties.
4.12	Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.	Maybe?	Task 2
4.13	When credible evidence indicates that irrelevant variance could affect scores from the test, then to the extent feasible, the test developer should investigate sources of irrelevant variance. Where possible, such sources of irrelevant variance should be removed or reduced by the test developer.	Yes	
4.14	For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure.	No	The Algebra 1 and SC Ready exams do not have a time limit.
4.15	The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented.	No	Task 4
4.16	The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.	No	Task 4

(continued)

Key			
Not Relevant	Maybe Relevant	Overarching Relevance	Relevant for a Different Task

No.	Standard	Relevant	Rationale
4.17	If a test or part of a test is intended for research use only and is not distributed for operational use, statements to that effect should be displayed prominently on all relevant test administration and interpretation materials that are provided to the test user.	No	The Algebra 1 exam is not used for research purposes only.
4.18	Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended- response items such as performance tasks, portfolios, and essays.	No	Task 5
4.19	When automated algorithms are to be used to score complex examinee responses, characteristics of responses at each score level should be documented along with the theoretical and empirical bases for the use of the algorithms.	No	Task 5
4.20	The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.	No	Task 5
4.21	When test users are responsible for scoring and scoring requires scorer judgment, the test user is responsible for providing adequate training and instruction to the scorers and for examining scorer agreement and accuracy. The test developer should document the expected level of scorer agreement and accuracy and should provide as much technical guidance as possible to aid test users in satisfying this standard.	No	Task 5
4.22	Test developers should specify the procedures used to interpret test scores and, when appropriate, the normative or standardization samples or the criterion used.	No	Task 5

(continued)

Key			
Not Relevant	Maybe Relevant	Overarching Relevance	Relevant for a Different Task

No.	Standard	Relevant	Rationale
4.23	When a test score is derived from the differential weighting of items or subscores, the test developer should document the rationale and process used to develop, review, and assign item weights. When the item weights are obtained based on empirical data, the sample used for obtaining item weights should be representative of the population for which the test is intended and large enough to provide accurate estimates of optimal weights. When the item weights are obtained based on expert judgment, the qualifications of the judges should be documented.	No	Task 5
4.24	Test specifications should be amended or revised when new research data, significant changes in the domain represented, or newly recommended conditions of test use may reduce the validity of test score interpretations. Although a test that remains useful need not be withdrawn or revised simply because of the passage of time, test developers and test publishers are responsible for monitoring changing conditions and for amending, revising, or withdrawing the test as indicated.	Yes	Although not included in the documentation provided, this standard should be considered for future test development purposes. The test specifications are clear and current but they may not always remain that way. Future development in the best practices of teaching Algebra may necessitate a change to the test specifications, and the items to which they are written to assess.
4.25	When tests are revised, users should be informed of the changes to the specifications, of any adjustments made to the score scale, and of the degree of comparability of scores from the original and revised tests. Tests should be labeled as “revised” only when the test specifications have been updated in significant ways.	Yes	Although not included in the documentation provided, this standard should be considered for future test development purposes. The test specifications are clear and current but they may not always remain that way. Future development in the best practices of teaching Algebra may necessitate a change to the test specifications, and the items to which they are written to assess.