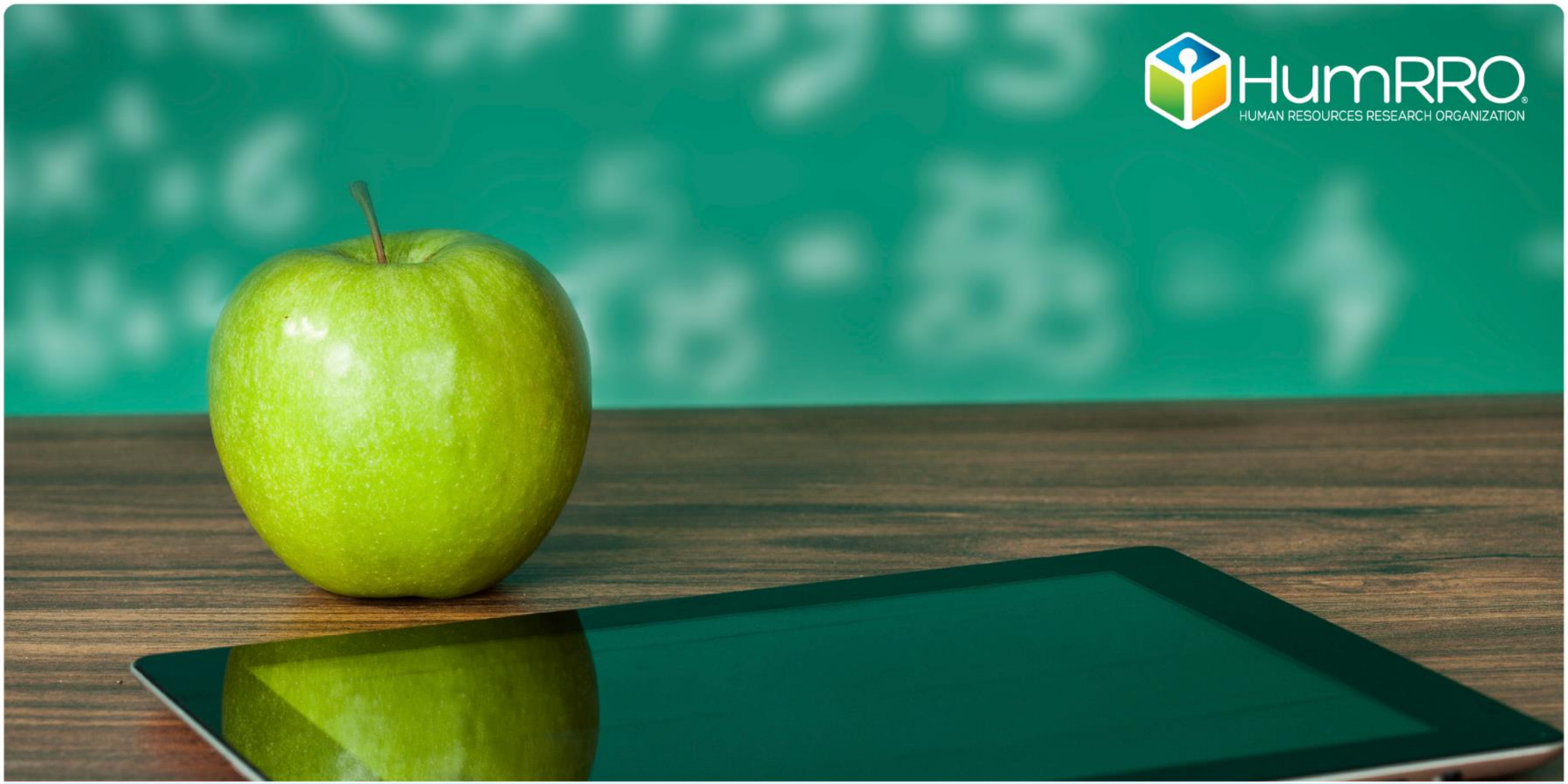


2018 No. 007

South Carolina Assessment Evaluation Report #2
Part I: Technical Evaluation &
Part II: Legal Evaluation

Prepared for:	South Carolina Education Oversight Committee (EOC) 1205 Pendleton Street Room 502 Brown Building Columbia, SC 29201	**Prepared under:**	Contract Number: 4400014645
Editors:	Andrea L. Sinclair Arthur Thacker	**Date:**	February 23, 2018

South Carolina Assessment Evaluation
Report #2

Part 1: Technical Evaluation - 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 second, and most extensive, of three reports detailing the findings from the evaluation. This report serves as the final analysis of the South Carolina College- and Career-Ready (SC READY) assessments and the End-of-Course Examination Program (EOCEP) assessments for Biology 1 and Algebra 1. The third and final report, to be delivered in June 2018, will constitute the final technical evaluation for the EOCEP English 1 assessment, for which text dependent analysis (TDA) items are operational for the first time in 2017-18.

This report is separated into two sections—Part I and Part II. Part I constitutes the technical evaluation of the South Carolina assessments (SC READY, English 1, Biology 1, and Algebra 1) as required by Section III, parts 'a' – 'e' in the Request for Proposal (RFP) (pgs. 15-16). Part II constitutes the legal evaluation of the SC READY assessments as required by Section III, part 'f' in the RFP (pgs. 16 -17). The technical evaluation (Part I) is an evaluation of the assessments' compliance with industry standards of test development and testing practices as described in *The Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014; hereafter referred to as the *Test Standards*). The *Test Standards* provide guidelines for assessing the validity of interpretations of test scores for the intended test uses. The *Test Standards* is *not* a statement of legal requirements (AERA, APA, & NCME, 2014, p. 1). Consequently, HumRRO contracted with an expert consultant, a nationally recognized expert in assessment law, Dr. S. E. Phillips, PhD, JD, to conduct an evaluation of the minimum legal requirements of the SC READY assessments specified in Section 59-18-325 of the South Carolina Code of Laws. The legal evaluation of the SC READY assessments is presented in Part II of this report.

Overall, the technical evaluation of the SC READY assessments and the EOCEP assessments indicates that the assessments adhere to industry best practices with some areas noted for improvement. We outline here the areas of strength for each assessment, and offer some recommendations where further improvements can be made. Each recommendation is accompanied by a priority rating. The table below presents the classification schema applied to the recommendations.

Priority Rating Codes for 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 and resources allow.

SC READY[1]

Review of Item Development Processes (Task 1)[2]

We evaluated the extent to which the evidence on item development processes complies with the *Test Standards*. We evaluated the strength of evidence for four *Test Standards* pertaining to item development. On a 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), three *Test Standards* received a rating of 4 and one received a rating of 5 (M = 4.25, SD = 0.43). This indicates that the evidence mostly or fully covers these *Test Standards*. Thus, overall, we found that the processes used to develop items for the SC READY assessments adhere to industry best practices.

The documentation on item development processes was the same for the SC READY assessments for ELA and math. Thus, the summary of findings for Task 1 apply to both ELA and math.

Areas of Strength

- Test developers clearly described the purposes and uses of the tests.
- Item writers are carefully selected and trained.
- Item development processes follow well-established industry procedures. Items undergo multiple rounds of reviews from various perspectives, such as content, bias, fairness and sensitivity, and accommodations. Readability and grade level appropriateness are considered during the item development processes. Quality assurance procedures are in place to oversee the entire process and identify potential issues.
- A comprehensive review of item development, from start to finish, for a sample of items revealed that the items adhere to item quality guidelines, and that feedback from each round of review was incorporated to improve item quality.

Recommendations for Improvement

We requested 13 different sources of information pertaining to South Carolina's SC READY item development processes from South Carolina's test vendor, the Data Recognition Corporation (DRC). Documentation was provided that addressed each of our requests, suggesting that DRC generally documents steps taken during the item development process. However, we noted some of these documents could be improved by including additional information or details about certain aspects of the item development process.

- As mentioned in our first report (Dickinson, Chen, & Swain, 2017), we found that item review guidelines and checklists vary in their comprehensiveness across documents. For instance,

[1] Not surprisingly, some similar recommendations are provided in the Part I Technical Evaluation and Part II Legal Evaluation. There may be instances of slightly different priority ratings for similar recommendations. Such differences stem from the fact that the Technical Evaluation is making recommendations from a psychometric perspective and the Legal Evaluation is making recommendations from the perspective of compliance with the legal requirements specified in Section 59-18-325 of the South Carolina Code of Laws.

[2] The Review of Item Development Processes (Task 1) for the SC READY assessments was included in Report #1 (Dickinson, Chen, & Swain, 2017). Subsequent documentation was provided for SC READY as a result of recommendations included in Report #1. The additional documentation is reflected in the findings reported in the current report. If the additional documentation did not change the findings provided in Report #1, then those findings are carried over from Report #1 such that the current report represents the complete and final analysis of SC READY.

the *Item Review Checklist* document provides a brief item review checklist, whereas the Item Writer Training files (*Making Assessments Accessible and Inclusive*) provide a detailed content review checklist. It may be useful to add 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 for quality. ***(Priority Rating: Medium)***

- As mentioned in our first report (Dickinson et al., 2017), universal design principles are referenced, but different documents provide different details on how to fulfill these principles. For example, the *Quality Assurance Procedures for Item Development* document lists five item writing and editing practices to comply with the universal design principles. However, the item writer training files provides a more comprehensive list of actions that should be followed to comply with universal design principles. Because of the inconsistency between the documents, the current practices that DRC takes to ensure the accessibility of items is unclear. Inconsistencies in the guidance to comply with universal design principles should be reconciled. ***(Priority Rating: Medium)***
- As mentioned in our first report (Dickinson et al., 2017), test developers documented the recruitment process for item writers as well as item writers' qualifications and relevant experiences. However, no information was provided about how item review committee members (e.g., reviewers for bias, fairness, and sensitivity; accommodation experts) are selected. Details on how item review committee members are selected should be provided. ***(Priority Rating: Medium)***
- Additional research studies could be conducted to inform and strengthen existing item development processes. For example, studies on pilot and field test data could be conducted to detect aspects of item design, content, and format that might introduce construct irrelevant issues for specific subgroups and individuals. Usability studies could be conducted to examine students' interactions with the items. Cognitive lab studies could be conducted to collect information about students' thinking and reasoning processes. Results from additional research studies such as these could further inform the item development processes and strengthen the reliability, validity, and fairness of items for all examinees. ***(Priority Rating: Low)***

Review of Standards Alignment and Item Quality (Task 2)

Panels of content experts reviewed the item quality and alignment of SC READY items to the South Carolina College-and–Career Ready Standards (SCCCRS). Overall, the content experts found that the items were aligned to the standards and that the items were of high quality. Separate panels of content experts conducted these activities for ELA and math. Consequently, the summary of findings is reported separately for ELA and math.

Areas of Strength (ELA)

- There is good alignment between the test items and the SCCCRS for ELA. For all grades, the vast majority of items were rated by the content experts as partially or fully aligned to the SCCCRS.
- There is good alignment between the test items and the test blueprint. For all grades, the number of items linked to each ELA domain was within the target number of items specified in the test blueprint.
- The items are of high quality. For all grades, the vast majority of items were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias.

Recommendations for Improvement (ELA)

- In grades 4 and 6, the depth-of-knowledge (DOK) level of over half the items was at or above the DOK level of the standards; however, for the other ELA grades, the majority of the item DOK levels were lower than the DOK levels of the standards to which they were linked, particularly for grades 5, 7, and 8. The South Carolina Department of Education (SCDE) should consider including target DOK levels in its test blueprints to improve consistency between the DOK levels of the standards and the items developed to assess those standards. ***(Priority Rating: High)***

- For all grades, the content experts felt the test blueprints adequately cover what the students should know and be able to do according to the SCCCRS for ELA. However, the content experts provided some suggestions for revising the test blueprints to further improve representation of the SCCCRS. Those recommendations included (a) removing the inquiry standard from the test blueprints and assessing the inquiry standard via another format (e.g., performance-based assessment), and (b) consider assigning different weights to the standards in grades 6, 7, and 8 to reflect increases in skills across grades. The SCDE should convene a group of South Carolina content experts to consider these recommendations for revising the test blueprints for ELA. ***(Priority Rating: Low)***

Areas of Strength (Math)

- There is good alignment between the test items and the SCCCRS for math. For all grades, the vast majority of items were rated by the content experts as partially or fully aligned to the SCCCRS.

- There is good alignment between the test items and the test blueprint. For all grades, the number of items linked to each math standard was within the target number of items specified in the test blueprint.

- There is good consistency between the DOK levels of the items and the DOK levels of the standards to which they are linked.

- The items are of high quality. For all grades, the vast majority of items were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias.

Recommendations for Improvement (Math)

- For grade 4, the content experts felt the test blueprint adequately covers what the students should know and be able to do according to the SCCCRS for grade 4 math. However, for all other grades, the content experts offered suggestions for revising the test blueprints to better address the SCCCRS for math. For grade 3, they recommended assigning greater weight to the Number Sense and Base Ten and the Number Sense and Operations – Fractions categories, given that they are the "foundation of future math understanding." They also felt there was not enough variety of graphing data items and that there was an overuse of interpreting bar graphs. For grade 5, they felt that there was an over-emphasis of standard 5.G.2 (Geometry, about coordinates), standard 5.G.1 (Geometry, define a coordinate system), and that the items that addressed those standards required low-level thinking. They recommended increasing the weights assigned in the test blueprint for Number Sense and Base Ten, Number Sense and Operations – Fractions, and Algebraic Thinking and Operations to reflect the number of standards and collective complexity of standards within those categories. They also recommended reducing the weights allocated to Geometry and Measurement and Data Analysis to reflect the lower number of standards within those categories. For grades 6 and 7, the content experts recommended that the weight for Ratios and Proportional Relationships should be increased because they felt that category was more important than Geometry and Measurement. They also recommended that Data Analysis and Statistics should be given less weight. They recommended that the Number System, Expressions, Equations, and Inequalities, and Ratios and Proportional Relationships categories should each be weighted 25%, while the Geometry and Measurement and Data Analysis and Statistics categories should each be weighted 12.5%. For grade 8, they suggested the Number System and Data Analysis, Statistics, and Probability categories should have less weight, and the weight for Functions, Geometry and Measurement, and Expressions, Equations, and Inequalities, should be increased. The SCDE should convene a group of South Carolina content experts to consider these recommendations for revising the test blueprints for math. ***(Priority Rating: Low)***

Review of Test Construction Processes (Task 3)[3]

We evaluated the extent to which the evidence on test construction processes complies with eight *Test Standards* pertaining to test form construction. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), three *Test Standards* received a rating of 3, two received a rating of 4, and three received a rating of 5 (M = 4.00, SD = 0.87). Thus, overall, we found that the processes used to construct forms for the SC READY assessments adhere to industry best practices. Moreover, an observation of test form assembly for SC READY indicates that the documented procedural steps are mostly followed during actual forms assembly.

There was considerable overlap in the test construction documentation for the SC READY assessments for ELA and math. Furthermore, the findings and conclusions do not differ across ELA and math. Thus, the summary of findings for Task 3 is combined for ELA and math.

Areas of Strength

- The documentation describes in detail the assembly of test items into forms including item order, item statistics, cueing, answer key repetitions, content specifications and other characteristics. Additional detail on test format and timing is found in the *Test Administration Manual*.
- Item statistics from the item bank demonstrate improvements in the available item pool from 2016 to 2017. Items with statistics outside of the acceptable ranges were removed between 2016 and 2017.
- The design for field testing items is an embedded approach in which field test (FT) items are spread throughout an operational form. This ensures item statistics are field tested using the same population of students who are administered the operational items, which allows for accurate item parameter estimation.
- A mode comparison study conducted on the spring 2016 SC READY assessments indicates that nearly all of the individual test items on the paper-and-pencil and online tests for ELA and math were *not* flagged for differential item functioning (DIF). Furthermore, a comparison of item *p*-values (proportion answering correctly) between paper-and-pencil and online tests

[3] The Review of Test Construction Processes (Task 3) for the SC READY assessments was included in Report #1 (Dickinson, Chen, & Swain, 2017). Subsequent documentation was provided for SC READY as a result of recommendations included in Report #1. The additional documentation is reflected in the findings reported here. If the additional documentation did not change the findings provided in Report #1, then those findings are carried over from Report #1 such that the current report represents the complete and final analysis of SC READY.

indicates only very small differences in item *p*-values between paper-and-pencil and online tests for math; however, for ELA, item *p*-value differences, while mostly small, consistently slightly favored paper-and-pencil examines across most items on the tests, such that the overall raw scores for ELA examinees tended to be slightly lower (see recommendation for ELA below).

Recommendations for Improvement

In response to recommendations included in the first report (Dickinson et al., 2017), DRC created and provided an *Item Development Technical Manual* and an *SC READY 2017 Technical Manual*. This additional documentation addressed several of the recommendations in the first report. Thus, the recommendations that follow stem primarily from our on-site observation of SC READY forms construction.

- Currently, information pertinent to forms construction can be obtained from the SCDE website, *030_Forms Construction Guidelines_E.pdf,* and *SC READY 2017 Technical Report for HumRRO.pdf.* It would be helpful to compile this information in a unified source, which should also contain the rationale for the intended uses of the assessments. ***(Priority Rating: High)***
- If items on the SC READY assessment include items from DRC's college- and career-readiness (CCR) item bank, for which item statistics are based on students in other states (i.e., not South Carolina students), then additional detail should be provided on that population of students to ensure that it is representative of the South Carolina population of students.[4] ***(Priority Rating: High)***
- During the forms construction meeting, the psychometrician appeared to use an Excel macro to compute form statistics. Given the high-stakes nature of the decisions based on form statistics, we recommend quality checks be conducted of the Excel macro to ensure the formulas are accurate. Additionally, the process could be modified to rely less on manual modification of Excel spreadsheets (e.g., copying and pasting of item information from different Excel spreadsheets) as input to the macro. ***(Priority Rating: High)***
- Approximately 25% of items are refreshed each year. However, there does not appear to be a mechanism to track how long an item has been on a form. We recommend the item bank include the year(s) and form(s) on which the item was last used and how many times the item has been used on an operational form. ***(Priority Rating: High)***
- If significant numbers of students continue taking the ELA paper-and-pencil tests, then propensity score matching studies should be conducted to confirm that scores on the paper-and-pencil tests and online tests are comparable and do not warrant statistical adjustment. ***(Priority Rating: High)***
- During forms construction, when participants rejected items for inclusion on a form, the participants' reasons for rejection did not appear to be documented. We recommend including item rejection explanations within the item bank. This information would be useful for editors to correct information or allow staff to immediately exclude these items during future forms assembly. ***(Priority Rating: Medium)***
- The SCDE may want to consider requesting that DRC create a statistical program that assembles forms to satisfy content and psychometric requirements simultaneously. These

[4] 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.

forms would then be reviewed by content specialists to identify concerns and be revised as needed. Enacting such a process would be more efficient by removing some of the manual steps involved in the current forms construction process, while still leveraging the expertise of the content experts in the areas in which they uniquely contribute. ***(Priority Rating: Low)***

- During the forms construction meeting, when the content specialists had difficulty finding items to satisfy certain content standards, they appeared to pull items from DRC's CCR item bank. However, it was necessary to align these items to the SCCCRS before they could be used on a form. We recommend this alignment work be completed in a more thoughtful manner rather than on-the-fly. Alignment work can take time and include deliberation with other content experts. ***(Priority Rating: Low)***
- Not all meeting participants were actively engaged in aspects of forms construction during the forms construction meeting. Some participants had considerable periods of time in which they waited for others to finish a step so they could begin their step. Specifically, the SCDE staff's time was not used consistently during the meeting. Consideration should be given to restructuring the way SCDE content experts participate in the forms construction meeting. One suggestion may be for DRC content specialists to develop drafts of the forms, DRC psychometricians review them, and DRC content specialists revise them, all prior to the in-person forms construction meeting (SCDE could virtually attend this portion of the meeting if desired, which would save travel expenses). The in-person meeting could then begin with SCDE content expert reviews of the forms that DRC created. ***(Priority Rating: Low)***

Review of Test Administration Procedures (Task 4)

We evaluated the extent to which the evidence on SC READY test administration complies with 14 *Test Standards* pertaining to test administration. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), one *Test Standard* received a rating of 3, six received a rating of 4, and seven received a rating of 5 (*M* = 4.43, *SD* = 0.62*).* Thus, overall, we found that the test administration procedures for the SC READY assessments mostly adhere to industry best practices.

The documentation for test administration was the same for the SC READY assessments for ELA and math. Thus, the findings for Task 4 apply to both ELA and math.

Areas of Strength

- Among the key test administration documents (*Test Administration Manual, Administration Directions Manual, Online Tools Training*, and *Tutorial*), policies and procedures were clearly stated, comprehensive, and would likely support standardized administrations across conditions.
- Detailed provisions for testing students with documented disabilities are provided in the *Test Administration Manual* for SC READY. Moreover, DRC's *eDIRECT User Guide* lists all accommodations available for students testing online.
- Permissible variations in test administration conditions are clearly documented in the *Test Administration Manual* and the *Administration Directions Manual*.
- Video tutorials provide clear instructions about how to sign in and how to use basic and advanced tools of the online testing system. Information describing item types, sample items, and scoring rubrics for the writing component are available to test takers before the test date.

- DRC provided appropriate training and documentation so that test administrators understand the standardized procedures to follow. The *Test Administration Manual* includes accepted standardized procedures for determining accommodations, minimum technology requirements, test time limits, test make-up policies, and other acceptable variations in test administration. There are training and pretest workshops for school test coordinators, test administrators, and technology coordinators.
- The *Test Administration Manual* clearly states the appropriate processes to report and document test security violations. Additionally, the training PowerPoint® files include several test security case scenario vignettes to help standardize Test Administrator understanding and implementation of test security policies and procedures.

Recommendations for Improvement

- More clearly organize the *Test Administration Manual* so that all requirements are readily highlighted and known to test administrators. ***(Priority Rating: High)***
- We saw little indication of a Help Desk available for preparation and during actual test administration. We recommend making a Help desk available to assist with technical difficulties during the assessment. ***(Priority Rating: High)***
- More clearly describe appropriate procedures for operationally preparing student test tickets and entering student data. ***(Priority Rating: High)***
- More clearly describe procedures for systematically documenting and reporting changes and disruptions during the assessment. ***(Priority Rating: High)***
- Include information from usability studies or empirical research related to test administration to ensure that the test materials are clear and usable for all grade levels and subjects, specifically the SC READY ELA Tutorial and passage interface. This could help to elucidate the concerns surrounding potential mode differences between paper-and-pencil and online administrations noted above (i.e., for Task 3). ***(Priority Rating: High)***

  More clearly identify (a) qualifications of Test Administrators to administer accommodations, and (b) procedures to monitor the implementation of the accommodations. ***(Priority Rating: Medium***
- The Tutorial may use language that is too advanced for younger students. For example, "The ELA test will be a two-day test. For ELA Session 1, the extended response item will be a text dependent analysis or TDA item." Simpler language or more teacher-guided direction should be provided for younger students. ***(Priority Rating: Medium)***
- Provide practice materials in formats that can be accessed by all test takers (e.g., provide practice materials with accommodations that can be accessed by students with disabilities). ***(Priority Rating: Low)***

Review of Scaling, Equating, and Scoring Processes (Task 5)

We evaluated the extent to which the evidence on scaling, equating, and scoring processes complies with 16 *Test Standards* relevant to these topics. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), 13 *Test Standards* received a rating of 4, and three received a rating of 5 (M = 4.19, SD = 0.39*).* Thus, overall, we found that the scaling, equating, and scoring processes for the SC READY assessments mostly adhere to industry best practices.

The documentation for scaling, equating, and scoring processes was similar for the SC READY assessments for ELA and math. Thus, these findings apply to both ELA and math.

Areas of Strength

- The *Technical Report* and *Score Report Users' Guide* clearly outline the purpose of the test. The *Score Report Users' Guide* includes information on the score levels, types of items, and the set of generated reports with descriptions of how reported data should be interpreted and used. The SC READY individual student reports include scale scores and information about score precision and related performance levels and performance level descriptors (PLDs).
- The *Score Report Users' Guide* includes multiple reports tailored to the needs and interests of different stakeholder groups—for example, students, teachers, and school administrators. The *Guide* includes interpretation material and is revised annually.
- The *Standard Setting Technical Report* and *Addenda* are very thorough. DRC used the Bookmark Method, a common item mapping method for setting defensible cut scores. The method is appropriate to the assessments and attends to how the results are used. The technical report clearly describes the discussion of test impact data with panelists after the second round of ratings, and the *Addenda* clearly describes policy-based adjustments to the recommended cut scores. In the post workshop survey, the standard setting panelists generally indicated that training was clear and that they were at least partially confident in their bookmark placement. These processes indicate that panelists had a sound basis for making their judgements and were familiar with the skills and knowledge of students just transitioning into the higher achievement level.

Recommendations for Improvement

- A vertical scale was developed for the 2016-17 SC READY assessments. The vertical scale could be potentially confusing to some stakeholders, including teachers, parents, and students. To help guard against erroneous interpretations, the *Score Report Users' Guide* and supporting communications should more clearly explain interpretations of the vertical scale and their limitations. ***(Priority Rating: High)***
- In light of the changes to the 2016-17 scale, SCDE should conduct a study to verify that scores are correctly interpreted by stakeholders. ***(Priority Rating: High)***
- Currently, there is only one on-line test form and one paper-and-pencil test form with over 90% of the items in common. Creation of back-up forms would help to mitigate concerns with item exposure and test compromise. ***(Priority Rating: High)***
- The *Technical Report* (see Section 7.3) indicates that all students who attempted the test are included in the calibration sample, whereas the *SC READY Horizontal Linking Process* document includes a statement that the "SCDE requests a sample of at least 20,000 records" (p.1). This appears to be a discrepancy and should be resolved. ***(Priority Rating: High)***
- The *Guidelines for Item Analysis and Form Construction* document provides differential item functioning (DIF) information that the content and statistical characteristics of the anchor set reflect the test, but specific information is not provided. More detailed information about how the content and statistical characteristics of the anchor set reflect the test should be provided. ***(Priority Rating: High)***

- The *Technical Report* states that these ordinal categories for the diagnostic reporting categories within ELA and math do not directly correspond to the overall student performance levels (although the diagnostic category scores and overall scores are still correlated). This statement could also be included on the score report or in the *Score Report Users' Guide*. ***(Priority Rating: High)***
- Student reports include normative information with the inclusion of percentile ranks based on the subset of items from DRC's college- and career-readiness (CCR) item bank. Additional detail should be provided on the population of students on which the percentile ranks are based to verify that the population is representative of South Carolina students. ***(Priority Rating: High)***
- Provide information or reference links to the subscale Reading PLDs on the student report. ***(Priority Rating: High)***
- The SC READY tests in grades 3-8 math and grades 4-8 ELA are post-equated. The grade 3 ELA test is pre-equated. This information is not readily available in the *Technical Report*. Specific information regarding the grade 3 ELA test should be included in the *Technical Report*. ***(Priority Rating: Medium)***
- Per *Test Standard* 5.23, cut scores should be informed by empirical data concerning the relation of test performance to relevant criteria. As such, we recommend conducting a study to empirically validate whether attaining the cut score (or above) on each grade level SC READY test predicts success in the next grade level. ***(Priority Rating: Medium)***
- Scoring rubrics, procedures, and criteria are described in *SC READY Scorer Training Materials* and in the *Item Scoring and Quality Control* file. Rater qualifications for scoring are specified, but are not well documented. Rater qualifications should be further documented as should information on procedures for calibrating raters. ***(Priority Rating: Medium)***
- As described in the *SC READY Scorer Accuracy and Consistency* document, rater accuracy is monitored by back reading, inter-rater reliability, and validity papers. We did not find information about a rescoring policy if the inter-rater agreement levels are low. Documentation should include information on rescoring policies. ***(Priority Rating: Medium)***

Review of Psychometric Processing and Item Parameters (Task 6)

For this task, HumRRO conducted a review of psychometric processing for the SC READY grade 5 ELA assessment. We also reviewed the item parameters for all grade levels of SC READY ELA and SC READY math. Overall, results indicate that the psychometric processing steps are logical and that the item parameters are acceptable. The summary of findings and recommendations are presented separately for ELA and math.

Areas of Strength (ELA)

- Through our review of available and requested documentation we could follow the logic of DRC's item calibration and scaling processes and procedures.
- We were able to match some of the initial parameter estimates to the fourth decimal place.
- Our review of the item-level data from the 2016-17 administration of the SC READY ELA assessments indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer.

- For the most part, the Rasch item statistics indicate that the 2016-17 operational items measured student achievement in ELA at appropriate levels of difficulty, and that items functioned as intended.

Recommendations for Improvement (ELA)

- The request for the data and documentation required to conduct our psychometric replication did uncover an internal quality control issue for the testing contractor. Specifically, there was an error during the data cleaning process that resulted in duplicate student records being output into the student data file used to calibrate item parameters. Although DRC concluded that this error did not have any impact on item parameter estimation, it does highlight the benefit of having quality control mechanisms in place during operational psychometric processing. SCDE may want to request expanded internal quality procedures from their testing contractor to minimize the potential for errors during operational psychometric processing. This might include multiple staff members conducting the same analyses concurrently and then comparing at predefined points in the process. If some amount of duplicating is already in place, DRC should clearly document it and consider expanding upon it. ***(Priority Rating: High)***
- SCDE should consider requiring the testing contractor to coordinate with a third-party to independently replicate scaling, equating, and scoring (i.e., the production of scoring tables) to help further ensure accuracy in scores. ***(Priority Rating: Medium)***
- Even if third-party replication is not adopted, SCDE should consider requesting that DRC combine existing psychometric processing documentation into a single, streamlined technical document. This document should include expanded detail about psychometric processing steps. ***(Priority Rating: Medium)***
- Analysis of Rasch IRT statistics did reveal a pattern in which non-traditional item types (e.g., multiple-select, evidence-based) at the middle school level had more items flagged for difficulty parameters that fell outside of the ideal range. We recommend that DRC take a closer look at items flagged for high levels of difficulty to determine if there were any characteristics of these items that may have influenced student responses. At minimum, further scrutiny of these items could inform subsequent item development activities. ***(Priority Rating: Medium)***

Areas of Strength (Math)

- Our review of the item-level data from the 2016-17 administration of the SC READY math assessments indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer.

Recommendations for Improvement (Math)

- Analysis of Rasch IRT statistics did reveal a pattern in which non-traditional item types (e.g., multiple-select, technology enhanced) were more frequently flagged for difficulty parameters that fell outside of the ideal range. We recommend that DRC take a closer look at items flagged for high levels of difficulty to determine if there were any characteristics of these items that may have influenced student responses. At minimum, further scrutiny of these items could inform subsequent item development activities. ***(Priority Rating: Medium)***

EOCEP

Review of Item Development Processes (Task 1)[5]

We evaluated the extent to which the evidence on item development processes complies with four *Test Standards* pertaining to item development. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), three *Test Standards* received a rating of 4 and one received a rating of 5 (M = 4.25, SD = 0.43). This indicates that the evidence mostly or fully covers these *Test Standards*. Thus, overall, we found that the processes used to develop items for the EOCEP assessments adhere to industry best practices.

The documentation on item development processes did not differ substantively for Biology 1, English 1, and Algebra 1. Thus, the findings for Task 1 apply across all three of the reviewed EOCEP assessments.[6]

Areas of Strength

- Test developers clearly described the purposes and uses of the tests.
- Item writers are carefully selected and trained.
- Item development processes follow well-established industry procedures. Items undergo multiple rounds of reviews from various perspectives, such as content, bias, fairness and sensitivity, and accommodations. Readability and grade level appropriateness are considered during the item development processes. Quality assurance procedures are in place to oversee the entire process and identify potential issues.
- A comprehensive review of item development, from start to finish, for a sample of items revealed that the items adhere to item quality guidelines, and that feedback from each round of review was incorporated to improve item quality.

Recommendations for Improvement

We requested 13 different sources of information/documentation pertaining to South Carolina's EOCEP item development processes. Information/documentation was provided that addressed each of our requests, suggesting that DRC generally documents steps taken during the item development process. However, we noted some of these documents could be improved by including additional information or details about certain aspects of the item development process.

- As mentioned in our first report (Dickinson et al., 2017), we found that item review guidelines and checklists vary in their comprehensiveness across documents. For instance, the *Item Review Checklist* document provides a brief item review checklist, whereas the item writer training files provide a detailed content review checklist. It may be useful to add references

[5] The Review of Item Development Processes (Task 1) for the Algebra 1 assessment was included in Report #1 (Dickinson, Chen, & Swain, 2017). Subsequent documentation was provided for Algebra I as a result of recommendations included in Report #1. The additional documentation is reflected in the findings reported here. If the additional documentation did not change the findings provided in Report #1, then those findings are carried over from Report #1 such that the current report represents the complete and final analysis of Algebra I.

[6] The item development documentation did not differ substantively for the SC READY assessments and the EOCEP assessments. Thus, the Task 1 findings summarized here for the EOCEP assessments are the same as those summarized for the SC READY assessments.

to detailed guidelines and checkpoints in all documents so that item writers or reviewers can use all available information to review items and check for quality. ***(Priority Rating: Medium)***

- As mentioned in our first report (Dickinson et al., 2017), universal design principles are referenced, but different documents provide different details on how to fulfill these principles. For example, the *Quality Assurance Procedures for Item Development* document lists five item writing and editing practices to comply with the universal design principles. However, the item writer training files (*Making Assessments Accessible and Inclusive*) provide a more comprehensive list of actions that should be followed to comply with universal design principles. Because of the inconsistency between the documents, the current practices that DRC takes to ensure the accessibility of items is unclear. Inconsistencies in the guidance to comply with universal design principles should be reconciled. ***(Priority Rating: Medium)***
- As mentioned in our first report (Dickinson et al., 2017), test developers documented the recruitment process for item writers as well as item writers' qualifications and relevant experiences. However, no information was provided about how item review committee members (e.g., reviewers for bias, fairness and sensitivity; accommodation experts) are selected. Details on how item review committee members are selected should be provided. ***(Priority Rating: Medium)***
- Additional research studies could be conducted to inform and strengthen existing item development processes. For example, studies on pilot and field test data could be conducted to detect aspects of item design, content, and format that might introduce construct irrelevant issues for specific subgroups and individuals. Usability studies could be conducted to examine students' interactions with the items. Cognitive lab studies could be conducted to collect information about students' thinking and reasoning processes. Results from additional research studies such as these could further inform the item development processes and strengthen the reliability, validity, and fairness of items for all examinees. ***(Priority Rating: Low)***

Review of Standards Alignment and Item Quality (Task 2)

Panels of content experts reviewed the item quality and the alignment of EOCEP items for English 1, Biology 1, and Algebra 1 to the respective South Carolina standards. Overall, the content experts found that the items were aligned to the standards and that the items were of high quality.

Separate panels of content experts conducted these activities for English 1, Biology 1 and Algebra 1. Consequently, the summary of findings is reported separately for these three assessments.

Areas of Strength (English 1)

- There is good alignment between the test items and the standards for English 1. For the 2016-17 fall/winter form, the majority of items were rated by the content experts as partially or fully aligned to the standards, and for the spring 2017 form nearly all of the items were rated as partially or fully aligned to the standards.
- There is good alignment between the test items and the test blueprint with one exception. The Writing standard had slightly fewer items linked to it than the target number of items specified in the test blueprint. This was the case for both the fall/winter and spring forms.

- The test blueprint adequately reflects what students should know and be able to do per the standards for English 1.
- The items are of high quality. For both the fall/winter and spring forms, the vast majority of English 1 items were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias.

Recommendations for Improvement (English 1)

- The DOK levels of the items on the fall/winter form tended to be slightly lower than DOK levels of the standards to which they were linked. The SCDE should consider including target DOK levels in its test blueprints to improve consistency between the DOK levels of the standards and the items developed to assess those standards. ***(Priority Rating: High)***
- The number of items linked to the Writing standard on the fall/winter and spring forms was slightly below the target number specified on the test blueprint. The SCDE should consider adding one or two more Writing items to the English 1 EOCEP. ***(Priority Rating: Low)***
- There were some minor differences between the fall/winter form and the spring form in alignment ratings, DOK ratings, and item quality ratings. The SCDE should consider having South Carolina content experts review the fall/winter and spring forms for consistency. ***(Priority Rating: Low)***

Areas of Strength (Biology 1)

- There is good alignment between the test items and the standards for Biology 1. For both the 2016-17 fall/winter form and 2017 spring form, the vast majority of items were rated by the content experts as partially or fully aligned to the standards for Biology 1.
- There is good alignment between the test items and the test blueprint for Biology 1 with one possible exception. Standard HB.3 on the spring form was one item short of meeting the target number of items specified in the test blueprint.
- The test blueprint adequately reflects what students should know and be able to do per the standards for Biology 1.
- The items are of high quality. On both the fall/winter and spring forms, the vast majority of items were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias.

Recommendations for Improvement (Biology 1)

- Item DOK levels tended to be lower than the DOK levels of the standards to which they were linked on both the fall/winter and spring forms. The SCDE should consider including target DOK levels in its test blueprints to improve consistency between the DOK levels of the standards and the items developed to assess those standards. ***(Priority Rating: High)***
- The number of items linked to the Standard HB.3 on the spring form was one item short of meeting the target number of items specified on the test blueprint. The SCDE may want to consider adding one more item to assess Standard HB.3 on the spring form. ***(Priority Rating: Low)***

Areas of Strength (Algebra 1)[7]

- There is good alignment between the test items and the standards for Algebra 1. For both the 2016-17 fall/winter form and 2017 spring form, the vast majority of items were rated by the content experts as partially or fully aligned to the standards for Algebra 1.
- Overall, there is good consistency between the DOK levels of the items and the DOK levels of the standards to which they are linked.
- The test blueprint adequately reflects what students should know and be able to do per the standards for Algebra 1.
- The items are of high quality. On both the fall/winter and spring forms, the vast majority of items were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias.

Recommendations for Improvement (Algebra 1)

- 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. Adding DOK levels to test blueprints (see recommendations above pertaining to Task 2) may also help to resolve this issue. ***(Priority Rating: 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. ***(Priority Rating: Low)***
- Consider including additional item types to the Algebra 1 test. 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. ***(Priority Rating: Low)***

Review of Test Construction Processes (Task 3)[8]

We evaluated the extent to which the evidence on test construction processes for the EOCEP assessments complies with eight *Test Standards* pertaining to test form construction. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), three *Test Standards* received a rating of 3, two received a rating of 4, and three received a rating of 5 (M = 4.00, SD = 0.87*)*. Thus, overall, we found that

[7] The alignment and item quality workshop for Algebra I was included in Report #1 (Dickinson, et al., 2017). For ease of reference and completeness, that summary of findings is also included here.

[8] The Review of Test Construction Processes (Task 3) for Algebra 1 was included in Report #1 (Dickinson, Chen, & Swain, 2017). Subsequent documentation was provided for Algebra 1 as a result of recommendations included in Report #1. The additional documentation is reflected in the findings reported here. If the additional documentation did not change the findings provided in Report #1, then those findings are carried over from Report #1 such that the current report represents the complete and final analysis of Algebra 1.

the processes used to construct forms for the EOCEP assessments mostly adhere to industry best practices.

There was considerable overlap in the data and documentation provided for English 1, Biology 1, and Algebra 1. Moreover, the overall findings and conclusions did not differ across these assessments. Thus, the summary of findings is combined for English 1, Biology 1, and Algebra 1.

Areas of Strength

- The DRC *Item Development Technical Manual* provides a detailed description of the life cycle of an item. The procedure for selecting field test (FT) items is well-documented in terms of number of items, their placement, and statistics.
- The practice of SCDE reviewing a composed form (operational and field test items) is a wise practice, as a review of the pool of operational items would not provide a complete picture from an examinee's perspective.
- The design for field testing items is an embedded approach in which FT items are spread throughout an operational form. This ensures item statistics are field tested using the same population of students who are administered the operational items, which allows for accurate item parameter estimation.

Recommendations for Improvement

- Currently, information pertinent to forms construction can be obtained from the SCDE website, *030_Forms Construction Guidelines_E.pdf,* and *2016–17 EOCEP Technical Report for HumRRO.pdf.* It would be helpful to compile this information in a unified source, which should also contain the rationale for the intended uses of the assessments. ***(Priority Rating: High)***
- The documentation clearly refers to use of a Rasch model to calibrate new item parameters and equate them to a common scale. These parameters are used to generate form-level difficulty estimates and make comparisons across forms. Our review revealed a disconnect between the use of a Rasch model to calibrate and equate items and the use of classical test theory (CTT) parameters to assemble forms. We are unclear as to how forms can be pre-equated when CTT parameters are used to assemble forms rather than the equated Rasch difficulties. This should be clarified in the documentation. ***(Priority Rating: Medium)***
- The documentation mentions items are screened for DIF using the ETS Delta method. The documentation does not specify when DIF is evaluated—FT or operational, or after every administration. This should be clarified. ***(Priority Rating: Medium)***
- The vast majority of students complete the on-line EOCEP assessments (98%) as opposed to the paper-and-pencil versions. Nonetheless, the 2% who complete the PBT version could be matched (via propensity score matching) to conduct mode comparability analyses to verify that there are equivalent forms and comparable scores (i.e., no mode differences). ***(Priority Rating: Low)***

Review of Test Administration Procedures (Task 4)

We evaluated the extent to which the evidence on EOCEP test administration complies with 14 *Test Standards* pertaining to test administration. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), one *Test Standard* received a rating of 3, six received a rating of 4, and seven received a rating of 5 (*M* = 4.43, *SD* = 0.62*).* Thus, overall, we found that the test administration procedures for the EOCEP assessments mostly adhere to industry best practices.

Areas of Strength

- Among the key test administration documents (*Test Administration Manual*, *Online Tools Training*, and *Tutorial*), policies and procedures were clearly stated, comprehensive, and would likely support standardized administrations across conditions.
- Detailed provisions for testing students with documented disabilities are provided in the *Test Administration Manual* for the EOCEP assessments. Moreover, DRC's *eDIRECT User Guide* lists all accommodations available for students testing online.
- Permissible variations in test administration conditions are clearly documented in the *Test Administration Manual*.
- Video tutorials provide clear instructions about how to sign in and how to use basic and advanced tools of the online testing system. Information such as item types, sample items, and scoring rubrics for the writing component are available to test takers before the test date.
- DRC provided appropriate training and documentation so that test administrators understand the standardized procedures to follow. The *Test Administration Manual* includes accepted standardized procedures for determining accommodations, minimum technology requirements, test time limits, test make-up policies, and other acceptable variations in test administration. There are training and pretest workshops for school test coordinators, test administrators, and technology coordinators.
- The *Test Administration Manual* clearly states the appropriate processes to report and document test security violations. Additionally, the training PowerPoint files include several test security case scenario vignettes to help standardize test administrator understanding and implementation of test security policies and procedures.

Recommendations for Improvement

- More clearly organize the *Test Administration Manual* so that all requirements are readily highlighted and known to test administrators. ***(Priority Rating: High)***
- More clearly describe appropriate procedures for operationally preparing student test tickets and entering student data. ***(Priority Rating: High)***
- More clearly describe procedures for systematically documenting and reporting changes and disruptions during the assessment. ***(Priority Rating: High)***
- More clearly identify (a) qualifications of test administrators to administer accommodations, and (b) procedures to monitor the implementation of the accommodations. ***(Priority Rating: Medium)***
- Information about accommodations is primarily provided in the *Test Administration Manual*, which is less accessible for test takers. We recommend providing a list of online and paper-and-pencil testing accommodations for the EOCEP assessments that are designed specifically for students rather than test administrators. This list could be similar to what is provided for the SC READY assessments (see the *SC READY Online and Paper/Pencil*

Tools and Supports file).[9] Also, a FAQ list could be provided to students to address common questions about accommodations and accessibility. ***(Priority Rating: Medium)***

- Provide practice materials in formats that can be accessed by all test takers (e.g., provide practice materials with accommodations that can be accessed by students with disabilities). ***(Priority Rating: Low)***
- Include information from usability studies or empirical research related to test administration to ensure that the test materials are clear and usable for all grade levels and subjects. ***(Priority Rating: Low)***

Review of Scaling, Equating, and Scoring Processes (Task 5)

We evaluated the extent to which the evidence on scaling, equating, and scoring processes complies with 10 *Test Standards* relevant to the EOCEP assessments. On the 5-point rating scale (where 1 = No evidence of the Standard found in materials and 5 = Evidence in materials fully covers the Standard), six *Test Standards* received a rating of 4, and four received a rating of 5 (M = 4.40, SD = 0.49*).* Thus, overall, we found that the scaling, equating, and scoring processes for the EOCEP assessments mostly or fully adhere to industry best practices.

The documentation for scaling, equating, and scoring processes was similar for English 1, Biology 1, and Algebra 1. Thus, the summary of findings for Task 5 are presented across these three EOCEP assessments.

Areas of Strength

- The *Technical Report* and *Score Report Users' Guide* clearly outline the purpose of the test. The *Score Report Users' Guide* includes information on the score levels, types of items, and the set of generated reports with descriptions of how reported data should be interpreted and used at the summary and individual level.
- The *Score Report Users' Guide* includes multiple reports tailored to the needs and interests of different stakeholder groups—for example, students, teachers, and school administrators. The *Guide* includes interpretation material and is revised annually.
- The *Standard Setting Technical Report* and *Addenda* are very thorough. DRC used the Bookmark Method, a common item mapping method for setting defensible cut scores. The method is appropriate to the assessments and attends to how the results are used. In the post workshop survey, the Standard Setting panelists generally indicated that training was clear and that they were at least partially confident in their bookmark placement. These processes indicate that panelists had a sound basis for making their judgements and were familiar with the skills and knowledge of students just transitioning into the higher achievement level.

Recommendations for Improvement

- Creation of back-up forms would help to mitigate concerns with item exposure and test compromise. ***(Priority Rating: High)***
- The *EOCEP Technical Report* briefly mentioned that the prior test vendor conducted field tests with a sufficient number of items to create pre-calibrated item pools and to construct pre-equated operational-test forms for all tests. We did not find detailed documentation of

[9] https://ed.sc.gov/scdoe/assets/File/tests/middle/scready/SC%20Ready%20Accommodations%20Charts_12_31_15.pdf

the item calibration process and evaluations of the adequacy of the equating functions following operational administration. No post-equating checks are presented in the *EOCEP Technical Report*. The equating process should be more thoroughly documented. ***(Priority Rating: Medium)***

- The student report for EOCEP does not provide information about score precision. For example, there are no error bands that would indicate that the score is an estimate based on the test form. This detail should be included in the score reports. ***(Priority Rating: Medium)***
- Research should be conducted to verify that score reports are correctly interpreted by users. ***(Priority Rating: Medium)***

Review of Psychometric Processing and Item Parameters (Task 6)

For this task, HumRRO conducted a review of the item parameters for the English 1, Biology 1, and Algebra 1 EOCEP assessments. The findings did not differ across these three assessments. Thus, the findings are summarized across all three EOCEPs.

Areas of Strength

- The review of item-level data from the fall/winter 2016-17 and spring 2017 administrations of the English 1, Biology 1, and Algebra 1 assessments indicates that, overall, items are appropriately difficult.
- The review of item-level data from the fall/winter 2016-17 and spring 2017 administrations of the English 1, Biology 1, and Algebra 1 assessments indicates that, overall, items discriminate among student ability levels.
- The review of item-level data from the fall/winter 2016-17 and spring 2017 administrations of the English 1, Biology 1, and Algebra 1 assessments indicates that, overall, items were not written in such a way as to enable students to easily guess the correct answers.

Areas for Improvement

We have no recommendations for improving the Biology 1, English 1, and Algebra 1 assessments based on the results of Task 6.

Conclusion

Overall, the findings from Tasks 1-6 indicate that the South Carolina assessments mostly adhere to sound testing practices as described in *The Standards for Educational and Psychological Testing*, and thereby support the validity of the test scores for their intended uses and purposes. No critical concerns were identified from the technical evaluation of the South Carolina assessments. Nonetheless, several recommendations are provided in Part I of this report to further strengthen and improve the quality of the assessments. We applaud South Carolina for securing an external evaluation of its assessments to help ensure their quality. Periodic evaluations of testing practices will help to ensure their continued technical soundness.

The evaluation included in Part I does not constitute a statement on the legal requirements of the South Carolina assessments, as compliance with the *Test Standards* is not synonymous with compliance with legal requirements. Part II of this report (Task 7) provides an evaluation of the minimum legal requirements of the SC READY assessments specified in Section 59-18-325 of the South Carolina Code of Laws.

South Carolina Assessment Evaluation
Report #2

Part II: Legal Evaluation - Executive Summary

In its Request for Proposals for an assessment system evaluation, the Education Oversight Committee (EOC) included a requirement that the responder evaluate the minimum statutory requirements for the SC READY assessments after the 2017 administration. SC READY is a system of assessments that measure student achievement of the South Carolina state content standards in English language arts (ELA) and Mathematics in Grades 3 through 8.

In response, HumRRO contracted with Dr. S. E. Phillips, PhD, JD, a nationally recognized assessment law expert, for consultation on this legal evaluation (Task 7). The legal evaluation was completed following the 2017 administration of the SC READY assessments and consisted of three phases: review of written materials, follow-up inquiries to key personnel, and analysis and evaluation of the collected evidence. This final report for Task 7 details the findings of the legal evaluation, determines whether the minimum requirements of Section 59-18-325 of the South Carolina Code of Laws have been met, and makes recommendations for strengthening the legal and psychometric defensibility of the SC READY assessment system in the future.

Task 7: Results

The results of the legal evaluation are presented by criterion in the order in which the eight criteria appear in Section 59-18-325. After stating each criterion, relevant SC READY evidence supporting that criterion is presented followed by evaluative commentary on the quality and sufficiency of that evidence.

1. Comparison of Student SC READY Performance to Score Scales of Assessments of Comparable Standards in Other States

Evidence. SC READY comparison scores include user percentile ranks from “other states with comparable standards” and MetaMetrics’® lexile®/quantile® scores. Evidence relevant to Legislative Criterion 1 includes an Achieve Report discussing the comparability of South Carolina ELA and Mathematics content standards to the Common Core State Standards (Common Core) and other states’ college and career readiness (CCR) content standards adapted after an original adoption of the Common Core, the composition of the user group contributing data for the “other states” percentile ranks, and linking studies used to map SC READY scores to the lexile® and quantile® frameworks.

Evaluation. The comparability of the content standards and representativeness of the three user states contributing data for the “other states” percentile ranks is unclear because no demographic or concordance information has been documented. Although the lexile®/quantile® user sample of over 3.5 million students is much larger and more geographically diverse, it still may not be representative of students nationally and no claim is made about the similarity of users’ content standards. In sum, comparative information is available for two volunteer user groups from two different contractors. Limited information about the composition of these user samples makes it difficult to judge their comparability or representativeness. On the other hand, these data may be the best available and do provide some useful comparative information.

2. Development of a System of Summative, Vertically-Scaled, Benchmarked, Standards-Based Assessments

Evidence. The SC READY assessments are a system of grade level, standards-based assessments administered at the end of the school year. HumRRO evaluations confirmed that the 2017 SC READY assessments demonstrated very good alignment between the content standards, test blueprints and test items for ELA and good to acceptable alignment for Mathematics. Vertical scale scores are reported and the tests are directly benchmarked to performance by students in relatively large and small user norm groups from two contractors.

Evaluation. The lexile® and quantile® trajectories to Grade 12 CCR ranges provide useful evidence for claims of *on track performance for CCR*, particularly for students who *meet expectations*, but the accuracy of such predictions for South Carolina students has not yet been documented. As an alternative, the state might consider using South Carolina data to validate a chain of performance linking each grade level to preparedness for the following grade level with a culminating prediction of sufficient content knowledge in Grade 8 to be prepared for CCR courses in high school that are in turn linked to appropriate CCR measures such as college admissions tests' CCR benchmarks.

Reliability estimates for SC READY were generally high and met the Assessment TAC recommendation of .85 for all subjects, grade levels and groups except students with disabilities in Grades 7 and 8 Mathematics. Similar reliability estimates are not yet available for ELA Reading and some reliability evidence is needed for the reporting category indicator scores.

The 2017 vertical score scale was developed from 2017 data for which lower grade items were administered in adjacent upper grades. A major issue with the 2017 SC READY vertical scale is the potential for confusion and distress when students with equivalent scale scores are compared or negative growth is reported. Alternatively, if one assumed (purely for illustration purposes) that the 2017 vertical scale grade level distributions exhibited the same minimal overlap as the within-grade-level scale scores reported for SC READY in 2016, the potential for misinterpretation and anxiety would be greatly reduced.

3. Creation of SC READY Scores for Achievement of State Standards, Preparation for the Next Grade Level, and Student Growth in ELA (reading, writing) and Mathematics

Evidence. Individual student score reports for the SC READY ELA and Mathematics tests include several different types of scores designed to provide evidence of student achievement of state standards. For the ELA total score, the ELA Reading subscore, and the Mathematics total score, the student receives a performance level designation of *exceeds expectations, meets expectations, approaches expectations, or does not meet expectations* as defined by the South Carolina grade-level content standards and standard setting activities. One might logically conclude that students who score at or above the *meets expectations* performance level cut score on their grade level SC READY ELA or Mathematics tests have sufficient prerequisite knowledge and skills to be adequately prepared for the material covered at the next grade level. Students can demonstrate growth in ELA and Mathematics by maintaining a *meets or exceeds expectations* performance level in the prior and current testing years, exceeding the prior year's lexile® or quantile® scores, or increasing their vertical scale scores.

Evaluation. There is substantial evidence that the SC READY assessments provide appropriate scores indicating achievement of state standards and preparation for the next grade level. The evidence for growth measures is less convincing. It is unfortunate that the 2017 vertical scale score model does not provide traditional growth scores with reasonable interpretations. Its contradictory properties for scores that are supposed to be comparable and potential for reporting negative growth may make its scale scores troublesome for important audiences such as parents, educators and the public.

This leaves only the lexile® and quantile® scores as reasonable measures of growth over time. However, these scores are incomplete growth measures for ELA because they include reading but not writing. Moreover, the samples used to link the SC READY scores to the lexile® and quantile® scales were quite small relative to the student population, and student motivation for the separate linking tests may have been diminished because students likely knew it was a research study with no reporting of individual student scores.

4. Measurement of Student Progress Toward National College- and Career-Ready Benchmarks Derived from Empirical Research and State Standards

Evidence. MetaMetrics® conducted empirical research to develop direct links to lexile® and quantile® CCR ranges by analyzing typical reading texts and mathematical materials used in postsecondary education and the workplace. The reported lexile® and quantile® predicted growth trajectories are selected from among a set of typical student growth curves from a North Carolina norm group that best fit the current (and earlier grade level, if available) point estimate(s). If the estimated growth trajectory ends within the CCR interval, the student is predicted to achieve CCR by the end of Grade 12. If not, the score report provides a recommended growth trajectory that reflects the proportional accelerated improvement across the remaining grades that will be needed to reach the CCR interval by the end of Grade 12.

The vertical moderation procedure used in standard setting for the SC READY assessments provided an indirect link to national CCR standards. Panelists were provided with impact data from students' 2015 ACT Aspire® test series scores linked to the ACT Assessment college admissions test when they made their cut score adjustments.

Evaluation. It is difficult to identify a single, appropriate, national benchmark for CCR. Many states have used college admissions test benchmarks, but they apply only to high school students and are problematic because they assess content that does not align very well with most state content standards. MetaMetrics® has taken a different approach by quantifying the complexity of reading text or mathematical materials typically encountered in entry-level college courses or jobs requiring a high school diploma. The validity data linking SC READY *meets expectations* performance intervals to the lexile® and quantile® *on track for CCR* target ranges provide persuasive evidence that longitudinal data yet to be collected for South Carolina will support current CCR predictions.

5. Establishment of at Least Four Student Achievement Levels

Evidence. Evidence relevant to Legislative Criterion 5 includes the policy definitions and performance level descriptors for four student achievement (performance) levels and the standard setting activities that delimited the test score intervals corresponding to each of the four performance levels for the SC READY ELA and Mathematics assessments in Grades 3-8.

Evaluation. The SC READY assessments include four performance levels, two that signify proficiency and two that do not. Each of the performance levels is described by general policy statements related to the subject matter and by more specific performance level descriptors related to the state content standards. There is good documentation of the standard setting activities that recommended cut scores to delimit the four performance levels on the test score scales.

The consistency with which the SC READY assessments are predicted to classify students in the same performance level if they were to retest under similar conditions is quantified by estimates of decision consistency. Decision consistency estimates for SC READY were high, especially for classifying students into two performance categories (proficient and not proficient).

6. Inclusion of a Variety of Question Types that Test Student Understanding of the Content

Evidence. There are six different question types utilized in the SC READY assessments. Each is designed to address a different type of student understanding of the content. The question types include multiple choice (recognize a correct answer), multi-select (distinguish multiple correct and incorrect answers), evidence-based selected response (use evidence from a text to justify and support an answer), short answer or gridded response (supply a correct answer by typing or blackening ovals in a number grid), technology enhanced (online only: drag and drop, click on a spot, graph, or arrange options correctly) and a text-dependent analysis essay item (written response supported by text evidence) scored holistically by two raters.

Evaluation. The SC READY assessments are composed of a variety of item types that measure student understanding of the content in different ways. For some items, students select a correct answer and for others, the student must produce the answer. Some items require distinguishing multiple correct and incorrect answers and some require identification of evidence that best supports an answer. For students testing online, a few items utilize some of the unique features of the technology. There is also an extended essay item that requires students to combine text analysis, writing skill and use of evidence to support an answer.

Several studies conducted by HumRRO support the quality of the SC READY items.
The evidence for the content validity, alignment, differential functioning, reliability and quality control all supports the appropriateness and quality of the SC READY items and test forms.
No indicators of text complexity, such as readability indices or passage/form word lengths, are reported for the SC READY assessments.

DIF statistics are within normal limits for a standards-based achievement test but ethnic DIF is reported only for African-Americans. There appear to be enough Hispanic students to also calculate DIF statistics for that group. Psychometric best practice is to ask the fairness/sensitivity committee to re-evaluate items exhibiting DIF to determine if the committee members can identify anything about the items likely to have caused the DIF. If yes, the item is revised; if not, it is assumed the result occurred by chance and the item is retained for use if needed to satisfy the test blueprint.

7. Test Administration in Paper-Based and Computer-Based Formats

Evidence. Evidence relevant to Legislative Criterion 7 includes mode administration data, the district waiver policy, test forms, a mode comparability study, separate scale score tables, test accommodations policies, and test security policies.

Overall in 2016 about 35% of students tested online and 65% tested on paper. In 2017, the percent of students testing online improved substantially, ranging from nearly 60% in Grade 3 to almost 85% in Grade 8. Waivers of the requirement to test all students online are granted by the State Board of Education (SBE). In 2017, the SBE granted 55 waivers, primarily for lack of sufficient infrastructure and testing devices.

At the request of SCDE, the contractor completed a mode comparability study for the online and paper/pencil forms using the Spring 2016 field test data. Only two of 449 (about ½%) of the SC READY ELA operational items exhibited mode DIF (one each in Grades 5 and 8). For Mathematics, no mode DIF items were identified. The mode comparability study also examined p-value differences for online and paper/pencil tests. Summed across all the items, the study found an advantage for paper/pencil of about 1½ to 3⅓ raw score points for ELA and .03 to .62 raw score points for Mathematics.

Evaluation. The mode comparability study did not account for overall differences in the ability of online and paper/pencil test takers to manage the logistics of responding to entire test forms. In addition, the observed raw score differences occurred in groups of unequal ability. To evaluate whether there is a true mode advantage for paper/pencil ELA test takers, a linking study using matched samples could be conducted. A useful methodology for doing so annually is to create matched groups by selecting representative samples from the larger group that match the smaller group to create reference and focal groups of equal size and ability.

In other applications, decisions to report mode equated scores have been made when the average difference is more than one raw score point or when differential advantages were observed in specific segments of the test score distribution. The purpose for conducting mode equating when empirical studies detect *practically significant* differential *test form* performance is to be fair to all students and remove any performance incentives for educators to prefer administering paper/pencil tests. Conducting mode comparability equating should remain a priority as long as a considerable number of students continue to be tested via paper/pencil.

Test Administration and Test Security Policies for SC READY are detailed and strict. Reporting of violations is mandatory and the statutory provisions and administrative rules provide clear guidelines for investigations and sanctions for violators.

South Carolina also has a clear and detailed Testing Accommodations Policy. Testing accommodations decisions are made by the student's individualized education program (IEP) team and it is considered a security violation if they are not administered as prescribed. There are appropriate procedures for requesting accommodated testing forms and the online test engine has several useful features available to all students. Testing accommodations have been appropriately classified as standard when the tested skills are congruent with those specified by the content standards and the resulting test scores are comparable to test scores obtained under standardized conditions.

South Carolina has made substantial progress moving schools and districts to online testing, but there are still substantial numbers of students testing paper/pencil in the lower grades. Providing support and incentives for meeting the 100% online goal (except for accommodations) will likely remain a challenge.

8. Information Reported That Can Assist Educators to Align Assessment, Curriculum, and Instruction

Evidence. Educators have several tools available to assist them in using SC READY assessment information to align assessment, curriculum and instruction. Evidence relevant to Legislative Criterion 8 includes the South Carolina ELA and Mathematics content standards, Performance Level Descriptors (PLDs), test blueprints and sample items, SC READY Individual Student Reports (ISRs), District and School Roster Reports and labels, the eDirect Information Portal and Lexile® and Quantile® Score Reports.

Evaluation. The SC READY assessments include informative score reports and user information to aid educators in utilizing the test results to align their curriculum and instruction with the tested content from the state content standards. Appropriate interpretive cautions are also included with the reported scores on the individual student score reports.

Task 7: Ratings

The Task 7 legal review examined and evaluated the available evidence to determine whether the 2017 SC Ready assessment system meets the eight minimum legislative criteria prescribed in Section 59-18-325. Based on this review, the eight legislative criteria were rated using the rating scale presented in Table A.

Table A. Rating Scale for Legislative Criteria

RATING	DESCRIPTION
Meets +	Robustly meets minimum legislative criteria; evidence is extensive for all aspects
Meets	Meets minimum legislative criteria; evidence is adequate for all aspects
Meets –	Barely meets minimum legislative criteria; evidence is limited for some aspects
Does Not Meet	Fails to meet minimum legislative criteria; evidence is missing or inadequate

The ratings of each of the legislative criteria reflect an assessment of the adequacy and strength of the evidence presented and the degree to which the evidence is consistent with professional psychometric standards and supports the legal defensibility of the assessment program. The ratings for each of the eight legislative criteria with key comments are presented in Table B.

Summary: ***Overall, the SC READY ELA and Mathematics assessment system meets all of the eight minimum legislative criteria prescribed in Section 59-18-325.*** Policymakers, educators and the public can have confidence that the scores South Carolina students obtain on the SC READY assessments accurately reflect their current achievement of state standards and provide meaningful guidance about their readiness for the academic content of the next grade level. The assessment system effectively utilizes a variety of item types and a comprehensive development and review process to screen, assemble and analyze items aligned to the state content standards. Psychometrically appropriate standard setting procedures were used to establish four student achievement levels labeled *does not meet expectations*, *approaches expectations*, *meets expectations,* and *exceeds expectations*. Online and paper/pencil Test Administration, Testing Accommodations and Test Security policies are detailed, clear and designed to produce psychometrically valid and reliable student scores. Individual student reports present test information clearly and concisely and contain appropriate caveats for interpreting test scores. The best available evidence links the test performance of South Carolina students to the performance of students in other states and to college- and career-readiness. Useful information is provided for aligning curricula/instruction with the assessments.

Table B. Ratings and Comments for the Eight SC READY Legislative Criteria

Rating	Legislative Criteria Comments
Meets	**1. Linked scales for comparison to other states with comparable standards** comparison groups are best available but may be nationally unrepresentative, of inadequate size, or have insufficiently aligned content standards
Meets	**2. Vertically-scaled, benchmarked, standards-based, summative assessment system** system of grade level, standards-aligned, end-of-year tests with potentially confusing vertical scale scores and *on track for CCR* benchmarks
Meets –	**3. performance against state standards in ELA, reading, writing and mathematics; Preparedness for the next grade; growth** validity studies linking test scores to performance at the next grade level not yet done; vertical scale scores may show negative growth and other growth evidence is indirect; writing is part of ELA but no subscores with achievement levels are reported
Meets –	**4. Progress toward national CCR benchmarks from empirical research and state standards** available CCR evidence is indirect but persuasive; direct CCR predictions for elementary students are ill-advised due to imprecision and unproven validity; inchoate validity studies linking Grade 8 test scores to admissions test CCR benchmarks
Meets +	**5. Establishment of at least four student achievement levels** appropriate and well-documented standard setting procedures and performance level descriptors for 4 levels (*does not meet, approaches, meets*, & *exceeds expectations*)
Meets +	**6. Use of a variety of item types requiring demonstration of content understanding** mixture of item types; multiple-select, evidence-based & text-dependent analysis essay items simulate the type of thinking and analysis typically associated with CCR
Meets	**7. Availability of online and paper/pencil administrations** paper form and easy-to-use online testing platform with appropriate accommodations; online testing goals and capabilities (e.g., TE items; adaptive testing) not yet fully attained
Meets	**8. Reports information to assist educators in aligning curricula with assessments** summative assessments useful for global curricular alignment; reporting categories guide educators to areas for more in-depth evaluation

As with any new testing program, there are many supporting research studies and procedural decisions yet to be finalized for future test administrations to maintain the quality, equivalence, alignment and usefulness of the test forms. The SCDE has a knowledgeable Assessment TAC and experienced contractor staff to aid them in appropriately constructing and analyzing future test forms and in designing and conducting useful research studies. In the spirit of improving and strengthening the assessment program as these future actions are deliberated, the next section provides specific recommendations related to each legislative criterion. Addressing these recommendations and the suggestions provided in prior sections of this report will further support the psychometric and legal defensibility of the SC READY assessment system.

Task 7: Recommendations

Recommendations for improvement are listed below. Each recommendation is associated with one of the eight legislative criteria and has been assigned a priority rating of *urgent*, *high*, *medium* or *low* as described in Table C. In addition to improving legal defensibility, many of these recommendations also support improved psychometric defensibility.

Table C. Priority Ratings for Recommendations

PRIORITY	DESCRIPTION
Urgent	Definitely needs to be considered and addressed now
High	Needs to be considered and addressed as soon as possible
Medium	Should be considered and addressed as time and circumstances permit
Low	Might be considered and addressed as part of long term planning

Urgent Priority

Legislative Criteria 1 & 2: Request that the contractor provide South Carolina with additional validity information about the participating states and the methods used to derive the reported *other states with comparable standards* percentile rank norms. Consider requesting that the contractor organize alignment information similar to a textbook crosswalk (e.g., from the Achieve Report or published state content standards) to confirm the comparability of the other states' standards to those of South Carolina. Also consider exploring the option of reporting percentile ranks for *other states* independent of South Carolina data.

Legislative Criteria 2 & 3: Weigh the advantages against the potential misinterpretations of using the current, vertical scale, and consider adopting a more traditional vertical scale before reporting 2018 SC READY scores to provide reasonable growth score interpretations and avoid the appearance of negative growth. Now is an ideal time to make this change before a second year of comparative data is reported. Score reports for 2018 could report revised 2017 scale scores on the new vertical scale for comparison.

Legislative Criterion 5: Urge the State Board of Education (SBE), with the advice and consent of the Education Oversight Committee (EOC) per Section 59-18-320(D), to officially adopt the SC READY cut scores.

Legislative Criterion 7: Create a backup test form for each grade/subject to be held in reserve in case the operational test form is compromised before all schools have finished testing.

Legislative Criterion 8: Provide additional explanatory text in the Score Report User's Guide identifying the standard error of measurement (SEM) type and size actually used to calculate the scale score ranges reported on the individual student reports, and if necessary, revise the sample reports to be consistent with the actual data.

High Priority____

Legislative Criteria 1-8: Consolidate scattered program documents and information into a single, expanded Technical Manual with summarized material and data, relevant appendices, and references to supporting documents.

Legislative Criterion 2: For the Grades 3-8 ELA Reading subscores, report decision consistency estimates and reliabilities using the same methodology and statistics as for the total ELA scores. Revise, if necessary, when scores become more stable.

Legislative Criterion 2: To be consistent with the 2014 *Test Standards*, report preliminary reliability estimates for the reporting category indicator scores (low, middle, high) now and then revisit and revise them later, as appropriate, when scores are more stable.

Legislative Criterion 4: Consider creating an ELA Writing subscore and reporting performance levels and statistics similar to what is currently being done for ELA Reading.

Legislative Criterion 6: Document the frequency of item usage across years and use this information to target items for replacement based on prior exposure.

Legislative Criterion 6: Calculate ethnic differential item functioning (DIF) for Hispanics which represent about 9% of the South Carolina Grades 3-8 student population. Special rules/procedures for small samples may be appropriate for some grade/subject combinations.

Legislative Criterion 6: Consider routine replication of psychometric processing by an independent third party as an additional quality check. This will require more detailed documentation of procedures.

Legislative Criteria 6 & 7: As long as significant numbers of schools continue to census test with paper/pencil, conduct annual mode equating studies for ELA to ensure comparable scores and deter incentives for avoiding online testing. Also do so at least once for Mathematics to confirm that the differences are too small to warrant adjustment.

Legislative Criterion 7: Reconsider whether oral test administrations of the ELA Reading subtest should continue to be classified as standard accommodations in Grades 4-8 given the skill differences between reading and listening comprehension, the Achieve Report finding that reading fluency skills are included in the state content standards through the upper grades, and the removal of students tested orally from the lexile® linking study calibrations.

Medium Priority

Legislative Criterion 2: Design and conduct empirical research studies to validate CCR benchmarks using South Carolina data.

Legislative Criterion 3: Print numerical values next to point estimates on the lexile® and quantile® score report graphs to make year-to-year growth comparisons easier.

Legislative Criterion 3: Conduct research studies to empirically confirm that SC READY proficiency scores indicate adequate preparation for the next grade level for South Carolina students.

Legislative Criteria 3 & 4: Consider placing error bands around the reported lexile® and quantile® growth trajectories using ± 1 SEM estimated from the longitudinal sample. Also consider strengthening the cautionary statements at the bottom of the score reports. Develop a research plan to collect validity evidence to support CCR claims for South Carolina students.

Legislative Criterion 5: For future standard settings, select a wider representation of stakeholders to serve on the vertical moderation panels.

Legislative Criterion 6: Use an index of readability or total word counts to track the reading load for ELA passages and ELA and Mathematics test forms within and across grade levels.

Legislative Criterion 6: Ask the fairness/sensitivity educator committee to re-examine items with gender or ethnic DIF when deciding whether to retain or revise them.

Legislative Criterion 6: Report demographic information for fairness/sensitivity and content review committees similar to that reported for standard setting committees.

Legislative Criterion 7: Expand the number of annual site visits to increase coverage and deterrence. Develop a site visit plan and seek Assessment TAC advice. Select schools where violations are suspected and randomly select others so each District receives at least one unannounced visit over a several year period.

Low Priority

Legislative Criteria 2 & 6: Consider convening an experienced educator panel to reconsider the assessment of inquiry skills for ELA and blueprint weights for Mathematics.

Legislative Criterion 6: Consider specifying target depth of knowledge (DOK) levels in the test blueprints to support greater consistency with the content standards, especially for ELA where the greatest variability was observed.

Legislative Criterion 6: Superimpose cut scores on the Rasch item maps and identify the content of the items within each performance level to refine the PLDs and further strengthen the standards-based validity evidence for the SC READY assessment system.

Legislative Criterion 7: Continue to expand the availability of accommodated practice materials. Develop a plan for monitoring the provision of accommodations using school/district testing coordinators and/or site visits.

Legislative Criterion 7: Continue to explore item formats that take full advantage of the technological capabilities of online testing. Consider computer adaptive testing to shorten test lengths and administration times, and speed score reporting while maintaining score accuracy.

South Carolina Assessment Evaluation Report #2
Part I: Technical Evaluation

Table of Contents

List of Tables

Table of Contents (Continued)

Table of Contents (Continued)

List of Figures

South Carolina Assessment Evaluation Report #2
Part I: Technical Evaluation

Introduction

Andrea L. Sinclair

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 second of three reports summarizing 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 part of this process, assessments are evaluated for validity, including alignment with the state standards, level of difficulty, and the ability to differentiate levels of achievement. Based on the evaluation, recommendations for improvements and changes are made. The EOC shares the information and recommendations 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 how it will address the recommendations and 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 and the End-of-Course Examination Program (EOCEP) for high school. The Data Recognition Corporation (DRC) works in coordination with SCDE to develop, administer, and score the tests.

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 math. The EOCEP is administered in ELA, math, science, and social studies to all public school students by the third year of high school. HumRRO's evaluation includes the SC READY for ELA and math at all tested grade levels, as well as the EOCEP assessments for English 1, Biology 1, and Algebra 1.

HumRRO's approach to evaluating South Carolina's assessment system includes a series of separate but related tasks that focus on the key elements of assessment design and implementation. Specifically, HumRRO identified the following seven tasks that address the general requirements listed in Section III (a-f) (pgs. 15-17) in the Request for Proposals (RFP):

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

Each of the above tasks is being conducted for the SC READY 3–8 ELA and math assessments, and for the EOCEP assessments in English 1, Biology 1, and Algebra 1, with one exception. Task 7 pertains only to the SC READY assessments. For Task 7, HumRRO contracted with an expert consultant, a nationally recognized expert in assessment law, Dr. S.E. Phillips, PhD, JD, to evaluate compliance of the SC READY assessments with the minimum legal requirements of Section 59-18-325.

To accomplish the above tasks, HumRRO coordinated with DRC and SCDE to obtain the necessary documentation and data. HumRRO's primary communication is with the Project Manager at DRC, who in turn coordinates with SCDE, as needed, to address our data requests and questions.

The seven tasks are being completed in a staggered fashion and the results presented over a series of three reports. The current report is the second of three reports, and serves as the most comprehensive. The third and final report, to be submitted in June 2018, will include the final technical evaluation of the EOCEP English 1 assessment for which text-dependent analysis (TDA) items became operational for the first time in the 2017–18 academic year. Table 1.0 summarizes the tasks and assessments included in each report.

Table 1.0 Tasks and Assessments Included in each HumRRO Report

Tasks	Report Number			
	SC READY	EOCEP English 1	EOCEP Biology 1	EOCEP Algebra 1
1. Review Item Development Processes	1, 2	2	2	1, 2
2. Review Item to Standards Alignment & Item Quality	2	2	2	1
3. Review Test Construction Processes	1, 2	2	2	1, 2
4. Review Test Administration Procedures	2	2, 3	2	2
5. Review Scaling, Equating, and Scoring Processes	2	2, 3	2	2
6. Review Psychometric Processing & Item Parameters	2	2, 3	2	2
7. Review Minimum Legal Requirements	2	--	--	--

The remaining chapters in Part I of this report describe the evaluation method and present results and related discussion for Tasks 1 – 6. The final chapter, Chapter 8, provides the conclusions for Part I. Part II (Chapter 9) describes the review of the SC READY assessments in view of minimum legal requirements (Task 7).

Chapter 1: Review Item Development Processes (Task 1)

Jing Chen & Hillary Michaels

Task 1: Introduction

For Report #2, HumRRO conducted an evaluation of the item development processes for the End-of-Course Examination Program (EOCEP) for the Biology 1 and English 1 assessments. The purpose of our evaluation was to document the extent to which best practices are employed to ensure the development of high-quality test items. A prior report (Report #1) provided findings of this same review for the Algebra 1 EOCEP and the SC READY ELA and math assessments (Dickinson, Chen, & Swain, 2017); however, we received additional information from DRC about the item development processes for Algebra 1 and the SC READY assessments since the delivery of Report #1. Consequently, in addition to presenting an evaluation of the item development processes for Biology 1 and English 1, we also provide in this chapter updated findings for Algebra 1 and the SC READY assessments.

It is worth noting the evaluation we describe in this chapter focused on the processes and procedures for initial item development and review and, therefore, is qualitative in nature. Subsequent chapters of this report include additional tasks that focus on item-level statistics and other quantitative data to further inform the quality of test items.

Task 1: Method

Our evaluation of the item development processes was conducted in two steps. First, we reviewed all available relevant documents and evaluated the processes described based on industry standards. Second, 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 evaluation methods we used in each step are described in more detail below.

Step I. Document Review

We worked in cooperation with the South Carolina Education Oversight Committee (EOC), the South Carolina Department of Education (SCDE), and the Data Recognition Corporation (DRC), with primary support provided by DRC, to obtain documentation related to South Carolina's item development processes. We also searched the SCDE website to identify additional relevant information. The documents we collected fell into several categories based on their foci (e.g., item writer training materials, item review guidelines, quality assurance procedures). Table 1.1 lists the documents we collected and reviewed. These documents provided useful information about various steps and procedures within the item development processes.

Table 1.1. Documents Reviewed for Task 1 – Item Development

Document Focus	Document File Name	Assessment(s) that the file applies to or comes from			
		Biology 1	English 1	Algebra 1[a]	SC READY[a]
Flowchart of Item Development Process	021_Flowchart Item Dev Process_E.pdf	X	X	X	
Item Review Checklist	023_Item Review Checklist_E.pdf	X	X	X	
Item Review Criteria	024_Criteria to Flag Items for Editing_E.pdf	X	X	X	
	026_Bias Sensitivity Criteria_E.pdf	X	X	X	
Item Writer Qualifications and Training Materials	028_Item Writers_E.pdf	X	X	X	
	012F_EOCEP Training Materials for Item Writers[b]	X	X	X	
Quality Assurance Procedures	022_Quality Assurance Procedures for Item Development_E.pdf	X	X	X	
Guidelines for Selecting/Developing Passages and Other Item Stimuli	029_Guidelines for Passages and Stimuli_E.pdf		X		
Assessment Accommodations	013F_EOCEP Accessibility and Accommodations[b]	X	X	X	
Item Banking System	014F_IDEAS Information[b]	X	X	X	X
Development of Scoring Materials	025_English TDA Scoring Guides Anchor Papers Practice Sets_E.pdf		X		
Technical Reports and Technical Manuals	2016-17 EOCEP Technical Report for HumRRO.pdf	X	X	X	
	SC READY 2017 Technical Report_100917.pdf				X
	DRC Item Development Tech Manual_101817.pdf	X	X	X	X
	EOCEP Forms Construction Guidelines_101817.pdf	X	X	X	
	SC READY Forms Construction Guidelines_101817.pdf				X
Sample Item Full Development Documentation	1_Item Development Documentation_ BIO.pdf	X			
	2_Item Development Documentation_ BIO.pdf	X			
	3_Item Development Documentation_ BIO.pdf	X			
	1_Item Development Documentation_ELA.pdf		X		
	2_Item Development Documentation_ELA.pdf		X		
	3_Item Development Documentation_ELA.pdf		X		

[a] Indicates we received additional materials for these assessments since our first report (Dickinson et al., 2017).

[b] Indicates the folder includes multiple files.

Our 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 *Test Standards*). In our previous evaluation of the SC READY and Algebra 1 assessments (Dickinson et al., 2017), we identified four standards from the *Test Standards* that were directly relevant to item development processes. We developed a rating scale to evaluate the degree to which the evidence for the assessments supports adherence to these *Test Standards*. The rating scale ranged from 1 to 5, with higher scores indicating stronger evidence for compliance with the standard (See Table 1.2). We used the same four standards and rating scale as used in Report #1. In addition, in this current report we updated our ratings for Algebra 1 and SC READY based on new information we received since our first report.

Table 1.2. Rating Scale for Evaluating Strength of Evidence for Test Standards

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

[a] Materials include all documents and data provided, any emails or phone calls with SCDE/DRC staff, as well as information we found online.

For each identified *Test Standard*, two HumRRO researchers independently assigned an overall rating based on the evidence collected. Then, the ratings assigned by the two researchers were compared and discussed to reach a final consensus rating for each standard.

Step II. Item Review

In addition to a document review, we reviewed a targeted sample of items from each assessment. The purpose of the item review was to track a sample of items from initial draft through the item development process to see how they were modified or dropped from operational use. To do this, we collected item cards for the sampled items. The item cards included each iteration of an item through the development process, along with reviewer comments. The item cards provided concrete examples that illustrated the item review and revision procedures. In addition, the item cards identify the targeted standard and sub-standard, or indicator, and a conceptual level of item difficulty (easy, moderate, or hard).

We requested and reviewed all available documentation for a representative sample of items. Six items were selected—three Biology 1 items and three English 1 items.

Task 1: Results

The information we collected from the two steps described above indicates the item development processes for the Biology 1 and English 1 are virtually the same. Because they do not differ substantively across these assessments, we present one set of results for Biology 1 and English 1. In addition, we note any changes to our prior ratings of Algebra 1 and SC READY (Dickinson et al., 2017) based upon the additional information we received since our first report.

Because the item development processes and documentation for Biology 1, English 1, Algebra 1 and SC READY do not differ substantively across these assessments, the results presented in Table 1.3 represent the final analysis of our review of item development processes for all reviewed assessments. Table 1.3 provides an overall rating for each relevant *Test Standard* based on our review of all available information.

Table 1.3. Evaluation of Item Development Processes Based on the Test Standards

Standard Number	Standard Content	Rating[a]
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.	5
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.	4
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

[a] See Table 1.2 for the rating scale.

Next, we discuss the rationales for our ratings in Table 1.3 and explain to what extent the standard was met. We also provide suggestions for further strengthening compliance with the *Test Standards*.

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 item cards suggests the item development processes comply with Standard 3.2 very well. To minimize the potential for tests being affected by construct irrelevant characteristics, items are carefully reviewed and edited, and reading passages and item stimuli are carefully selected and developed. For example, the *Item Review Checklist* includes check points that focus on the linguistic (e.g., do the stem and options match grammatically?), communicative (e.g., are supporting graphics necessary, appropriate, and clear?), cognitive (e.g., does the item address important knowledge and skills?), and other important characteristics of the items. In the item flagging criteria file (*024_Criteria to Flag Items for Editing_E*), 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. In the bias sensitivity criteria document (*026_Bias Sensitivity Criteria_E*), the item bias, fairness, and sensitivity review process and criteria are documented. These review processes are helpful to minimize construct irrelevant variance for the items.

In the EOCEP and SC READY guidelines for item analysis and form construction files (*EOCEP Forms Construction Guidelines_101817.pdf; SC READY Forms Construction Guidelines_101817.pdf*), DRC provided test blueprints that clearly describe the construct(s) to be measured, the desired attributes of the assessment, and the distribution of items and score points for each measured construct. In the DRC *Item Development Manual*, test developers describe how items are developed and reviewed to ensure they measure the intended construct.[10] For example, the first task of the item development process is to develop and/or review the test/item specifications and blueprints. Before writing items, the item writers are trained to focus on the content standards for a given program or project to gain a full understanding of the fundamental principles underlying what is to be taught and assessed. These procedures help to ensure items are designed to measure what they are intended to measure.

The item cards for the sampled items included information such as the content area, standard(s) the item addresses, depth of knowledge (DOK), and estimated item difficulty. This information provides 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 coverage gaps between the items and the standards. Alignment study results presented in Chapter 2 provide additional information about how well the items measure the intended standards/indicators and represent the DOK levels of the standards.

As mentioned in our first report (Dickenson et al., 2017), we found that item review guidelines and checklists vary in their comprehensiveness across documents. For instance, document 023 only provides a brief item review checklist. Appendix A of the *Item Development Manual* document and *the Item Review Content and Fairness Checklists* file provide a very detailed content review checklist. It may be helpful to include references to other detailed guidelines and checkpoints in all documents so item writers or reviewers can use all available information to review items and check the quality of items.

The item development processes are generally the same across different subjects of the EOCEP assessments. Furthermore, the item development processes of the EOCEP and the SC READY assessments follow the procedures described in the *DRC Item Development Manual* and also do not differ substantially.

In the additional documents we received since our first report, DRC listed the specifications and described how items were developed and reviewed to ensure they measure the intended construct. Given the additional information, the rating for Algebra 1 and the SC READY assessments for Standard 3.2 in our first report (Dickinson et al., 2017) should be upgraded from a score of 4 to a score of 5.

[10] Test developers may include DRC staff and SCDE staff, depending on the assessment.

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 indicates that key aspects of Standard 3.3 are addressed by South Carolina’s test development process. As described in the bias sensitivity criteria file (*026_Bias Sensitivity Criteria_E*), DRC conducts item review meetings to review all items for bias, fairness, and sensitivity issues. The review committee is comprised of 12 -15 South Carolina educators and several DRC facilitators. The content and sensitivity review meetings are typically five days in length and held in Columbia, South Carolina. Following each meeting, DRC staff documents all changes and concerns raised during the meeting and provides all documentation to SCDE staff to make final decisions regarding item edits.

In the bias sensitivity criteria document (*026_Bias Sensitivity Criteria_E*) and in one of the item writer training materials *(August 2016 Fairness in Testing Manual, file 012F*), test developers provide definitions of bias and sensitivity, discuss different types of bias, and describe topics to avoid, topics of concern, and special circumstances. Sample items with bias, fairness, and sensitivity concerns are provided to support training for the bias and sensitivity review. In one of the item writer training materials (*Item Review Content and Fairness Checklists, file 012F*), the test developers provide a detailed fairness item review checklist. This checklist helps ensure test items are accessible to a diverse student population with respect to gender, race, ethnicity, geographic region, socioeconomic status, language, disability, and other factors. All these practices suggest the item development processes are, for the most part, consistent with Standard 3.3. 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 for the Biology 1 and English 1 assessments. Customized formats (e.g., braille, large-print, sign language) are available for students with documented disabilities. Previous studies on accommodations for students with disabilities were reviewed and the universal design process was followed to improve examinees’ participation in the assessment. The three item cards provided in the 013F folder show specific edits for the items by SCDE accommodation experts to assist students with disabilities and English Language Learners (ELLs).

The EOCEP and SC READY guidelines for item analysis and form construction documents indicate items with a differential item function (DIF) flag of "C" should be avoided. They also indicate items with a DIF flag of "B" should be considered carefully and, when included, balanced among favored gender and ethnicity groups. This suggests the psychometric property of an item is used to evaluate whether the item is appropriate for all relevant subgroups. In many of these documents, the test developers did not explicitly describe the relevant subgroups under consideration (e.g., race and ethnicity groups) and how test validity, reliability, and precision are considered for specific subgroups.

Besides some accommodation studies, the authors did not mention any other studies related to examinee subgroups that are considered when constructing the test. It is unclear how analyses are carried out using pilot and field test data to detect aspects of test design, content, and format that might distort test score interpretation for the intended uses of the test scores for subgroups and individuals. Thus, we believe some improvements could be made to the item development processes to further strengthen adherence to Standard 3.3. Our final SC READY and Algebra 1 ratings for this standard remain at the level 4 rating provided in our first report.

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 EOCEP and SC READY documents provides evidence that key aspects of Standard 4.0 are being met. The test developers clearly described the purposes and uses of the tests in the EOCEP and the SC READY technical reports (i.e., *SC READY 2017 Technical Report_100917.pdf, 2016–17 EOCEP Technical Report for HumRRO.pdf*). The documented item development processes provide evidence of test fairness, reliability, and validity to support the intended uses of the test scores for individuals in the intended examinee population. For example, Chapters 7 and 8 of the EOCEP technical report specifically describe how the items and test are to be designed to ensure reliability and validity.

Test developers documented steps taken during the design and development process to provide evidence of fairness, reliability, and validity. For example, the overall item development process is documented (*021_Flowchart Item Dev Process_E)* as well as the item review guidelines, checklists, item writer training materials, and quality assurance procedures. However, some of these documents could benefit from additional detail to more thoroughly describe certain steps. For instance, some steps presented in the item development flowchart are not well documented (e.g., committee review process for the field test item data, process of reviewing RFP requirements, state curriculum, style guide, scope and criteria of the test). In addition, the test developers briefly referred to quality assurance procedures associated with the item development processes, but provided no detailed descriptions of the quality control procedures for each step in the item development process. Without detailed information, it is difficult to evaluate how well the various steps contribute to high-quality items.

The technical reports we received since our first report provided more details about the purposes and uses of the tests and how the test is designed in a way that supports the validity of interpretations of the test scores for their intended uses. Thus, we increased our final rating of this standard for SC READY and Algebra 1 from a level 3 rating to a level 4 rating.

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 suggests key aspects of Standard 4.8 are met; however, some improvements could be made in either the test review process or the documentation to better address this standard. Empirical results and expert judgments are used to review items and scoring criteria. For example, experts use results from empirical analyses and a set of psychometric guidelines (e.g., recommended ranges for *p*-values, item-total correlations, and differential item functioning—DIF values) review and select items. The scoring guide file (*025_English TDA Scoring Guides Anchor Papers Practice Sets_E.pdf*) describes how the scoring criteria were developed based on live student work by experts' judgments that include DRC test developers, in consultation with scoring experts and the SCDE. Expert judges are used in the bias, fairness, and sensitivity review, and the item review for test accommodations. However, the documents we received did not provide enough information for us to judge the extent to which expert judgments are appropriately used.

Test developers documented the recruitment process as well as item writers' qualifications and relevant experiences. Test developers also documented the types of activities and materials used to train item writers and item bias, fairness, and sensitivity reviewers. However, no information was provided about how item review committee members (e.g., reviewers for bias, fairness and sensitivity; accommodation experts) are selected or the extent to which SCDE and DRC documents experts' qualifications, relevant experiences, and demographic characteristics. Similarly, in the quality assurance (QA) file (*022_Quality Assurance Procedures for Item Development_E.pdf*), the experience and qualifications of staff who perform QA procedures at different levels are only briefly described. The existing documentation should be expanded to include additional information and details regarding the background and characteristics of expert judges and QA staff. Our final ratings of this standard for SC READY and Algebra 1 remain at a score of 4.

Task 1: Discussion

We evaluated DRC's item development processes for the Biology 1 and English 1 EOCEP assessments. In addition, we updated our evaluation results for Algebra 1 EOCEP and SC READY assessments to reflect additional information received since our first report. Our evaluation is based on available documentation on item development and review processes collected from DRC and SCDE. Results from other chapters of this report provide additional information such as the item-standard alignment results and the empirical item-level statistics to further inform the quality of test items. We found the processes used to develop items for the reviewed assessments adhere to industry best practices to a great extent. On a 5-point rating scale, three of the *Test Standards* received a rating of 4 and one *Test Standard* received a rating of 5.

We found the test developers clearly described the purposes and uses of the tests. Item writers are carefully selected and trained. Item development processes follow well-established industry procedures. Items undergo multiple rounds of reviews from various perspectives, such as content, bias, fairness and sensitivity, and accommodations. Readability and grade level appropriateness are considered during the item development processes. Quality assurance procedures are in place to oversee the entire process and identify potential issues. Our evaluation of the sample items revealed the items adhere to item quality guidelines and feedback from each round of the review was incorporated to improve item quality.

We requested 13 different sources of information/documentation pertaining to South Carolina's SC READY and EOCEP item development processes. Information/documentation was provided that addressed each of our requests, suggesting that DRC generally documents steps taken during the item development process. However, we noted some of these documents could be improved by including additional information or details about certain aspects of the item development process. For example, the test developers may consider documenting more detailed quality control and quality assurance procedures associated with each item development step. The test developers may also consider implementing guidelines related to reviewing and revising items based on empirical results. Currently, there are only guidelines about *selecting* items using empirical results.

In addition, we recommend DRC and SCDE document how (a) item review committee members (e.g., reviewers for bias, fairness and sensitivity, accommodation experts) are selected and (b) their qualifications, relevant experiences, and demographic characteristics are recorded. The item review guidelines and checklists vary in their comprehensiveness across documents. It may be helpful to include references to detailed guidelines and checkpoints in all documents so item writers or reviewers can use all available information.

While our evaluation was quite positive, there are additional ways item development processes may be improved. For example, studies on pilot and field test data can be conducted to detect aspects of item design, content, and format that might introduce construct irrelevant issues for specific subgroups and individuals. Usability studies can be conducted to examine students' interactions with the items. Cognitive lab studies can be conducted to collect information about students' thinking and reasoning processes. Evidence-centered design (ECD) principles and models can be employed in the item development processes. Results from additional research such as usability studies and cognitive labs could further inform the item development processes and strengthen the reliability, validity, and fairness of items for all examinees.

Chapter 2: Review Standards Alignment and Item Quality (Task 2)

Richard Deatz & Tanya Longabach

Task 2: Introduction

The South Carolina College-and-Career Ready Assessments (SC READY) and End-of-Course Examination Program (EOCEP) assessments were developed based on South Carolina's academic standards and test blueprints. Alignment studies address a vital question related to the validity of test scores: Does the test content adequately reflect the content that students are expected to learn as outlined in the state standards?

HumRRO conducted a workshop in which content experts reviewed alignment of the SC READY test items and the South Carolina College-and-Career Ready Standards (SCCCRS). Content experts also reviewed the alignment of the EOCEP English 1 and Biology 1 test items and the South Carolina standards for English 1 and Biology 1.[11] The purpose of the alignment reviews was to evaluate the extent to which students' test scores reflected content knowledge and skills at the breadth and depth outlined in the content domain (as specified in the South Carolina Standards). This chapter describes the alignment method, results, and discussion of the overall alignment of the SC READY and the EOCEP assessments (English 1 and Biology 1) to the respective South Carolina Standards.

Task 2: Method

Several methods of alignment are in use (e.g., Forte, 2017; Porter, 2002; Webb, 1997, 1999, 2005). These methods all involve panelists evaluating several aspects of the content standards and test items, and statistically analyzing their ratings to determine the extent to which the content standards and test items are aligned. For this study, HumRRO used a method that combined elements of Norman Webb's (Webb, 1997; 1999; 2005) and HumRRO's (e.g., Nemeth, Purl, & Smith, 2016) alignment methods to evaluate alignment of the SC READY and EOCEP assessments to the South Carolina Standards. We recruited highly qualified educators to provide ratings of alignment and item quality. To maintain the independent and external nature of the study, neither DRC nor South Carolina Department of Education (SCDE) staff participated in the alignment workshop.

Alignment Method

To address concerns regarding traditional alignment methods (including Webb's method), such as not considering a state's test blueprints or the impact on the degree of alignment when there is a large number of content standards (or indicators), we used a hybrid approach that included some aspects of both Webb's and HumRRO's alignment methods. Our approach and the six criteria we used to investigate alignment and item quality are presented next.

Items Represent Intended Content

This criterion is a check of alignment between content standards and test items. Simply stated, this involves a check of the content standard or indicator (i.e., the most detailed level of the standards) assigned to each item during the item writing process, by a group of independent panelists who did not develop the items. For this task, panelists rated items as not aligned,

[11] A prior report provided the findings of the alignment review for Algebra 1 EOCEP (Dickinson, Chen, & Swain, 2017).

partially aligned, or fully aligned to the designated standard.[12] For panelists to have rated an item as partially aligned to the designated standard, there must have been some content in the item that was not covered by the linked standard.[13] For not aligned or partially aligned ratings, panelists provided an explanation for why the item was not covered by the linked standard and identified another content standard to which the item was better aligned, if applicable. We will share with DRC and/or SCDE a password-protected document with the item ids for which most panelists rated the items as "not aligned" to the linked standard. This password-protected document will include panelists' comments/explanations of their ratings.

Items Represent Intended Categories

This criterion is a check of alignment between the test blueprint and test items. For this criterion, we compared the number of items specified for each standard/indicator in the test blueprint to the actual number of items linked to each standard/indicator by the panelists (this is similar to Webb's categorical concurrence criterion). The test blueprints include ranges for the number of items for each category (e.g., domain, strand, standard). This criterion was met when the actual number of items linked to each category were within the target ranges specified on the test blueprints.

Evaluation of Test Blueprint

The Request for Proposal (RFP) required "an evaluation of the test blueprint," which will ". . . include analyses and recommendations as to the test blueprint needed to provide valid and reliable results for the intended purposes of the assessments" (see RFP page 16). In addition to analyzing whether the number of test items linked to standards/indicators coincided with the target number of items specified for each category in the test blueprints, panelists also provided qualitative feedback on whether the test blueprint adequately covered what students should know and be able to do (based on the standards). We framed this discussion by explaining that it is not feasible to test every standard/indicator on a single assessment—for example, the ELA grade 5 SC READY assessment has 153 indicators. Because a decision was needed about what content to address in the test, panelists were asked to discuss until they reached consensus on the following questions:

- Does the test blueprint adequately cover what students should know and be able to do (based on the standards)?
- Is there anything under-emphasized or missing from the test blueprint?
- Is there anything on the test blueprint that is over-emphasized?

The facilitators captured the panelists' responses to these questions as well as their recommendations for improving the test blueprint.

[12] This differs from Webb's method in that panelists verified the quality of the item -to-standard link assigned by the item writers rather than creating their own independent item-to-standard linkages and then comparing those linkages to the linkages assigned by the item writers.
[13] If the content in the item was fully addressed by the standard, then items were rated as "fully aligned." Items need not cover the entire standard to be "fully aligned."

Depth of Knowledge (DOK) Consistency

Depth of knowledge (DOK) refers to the complexity of cognitive processing required of students. The DOK consistency criterion indicates whether there is consistency between the complexity of knowledge required by the standards/indicators and the complexity of knowledge required to correctly answer the items linked to those standards/indicators. Complexity and difficulty can be, and often are, correlated; however, it is important to note that complexity and difficulty are not the same. Test items can be difficult (i.e., many students answer the items incorrectly indicated by low *p*-values), *and* require a low level of cognitive processing. For example, consider the science test item, “Recall the atomic weight of chlorine.” This test item requires a low level of cognitive processing, but it is a difficult item to correctly answer. The converse is also true—that is, test items can require a high level of cognitive complexity (e.g., evaluating multiple sources of information), and still be items that many or most students answer correctly.

Because the South Carolina test blueprints do not include DOK levels for the domains/strands/standards and because the test maps (i.e., item summary information provided by DRC) do not include DOK levels for the items (both of which are used in HumRRO’s DOK consistency criterion), we adopted Webb’s DOK consistency criterion. The panelists provided DOK ratings on items and standards/indicators using the following scale:

- Level 1 Recall and Reproduction – Recall of information (i.e., facts, terms, simple procedures); student either knows the answer or does not; the answer does not need to be “figured out” or “solved.”
- Level 2 Skills/Concepts – Includes the engagement of some mental processing beyond recalling, reproducing, or writing a response; it requires both comprehension and subsequent processing of information.
- Level 3 Strategic Thinking – Requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels.
- Level 4 Extended Thinking – Cognitive demand is high and complex; requires evaluation of multiple sources or independent pieces of evidence; may require extended time to apply significant conceptual understanding and higher-order thinking.

The panelists’ item DOK ratings were compared to the DOK ratings on the standards/indicators linked to those items. Per Webb’s guidance for this criterion to be met, at least 50% of the items linked to the standard needed to be at or above the DOK level for that standard.

Evaluation of Item Quality

Because the RFP required an evaluation of test item quality (see RFP page 15), panelists independently rated each item on several aspects of item quality: (a) clarity of presentation, (b) accuracy of content, (c) grade-level appropriateness, (d) supports research-based instruction,[14] and (e) unbiased content or presentation. Panelists entered “yes” if the item met the quality indicator and “no” if the item did not meet the quality indicator. If “no” was selected, panelists explained their reasoning for why the item did not meet the item quality indicator. Panelists were informed that items had undergone extensive review and field testing, and to flag only items for

[14] If a student could figure out the correct answer without knowing the content (e.g., item cueing, implausible distractors), then the item was deemed as *not* supporting research-based instruction. This is how this quality indicator was operationalized for the purposes of this workshop. Evaluation of items for supporting research-based instruction was a requirement specified in the RFP (see Section III, part ‘a,’ pg. 15).

which they identified a substantive issue. All flagged items required an explanation. We will share with DRC and/or SCDE a password-protected document with items for which most of the panelists indicated the item had a quality issue. This password-protected document will include panelists' comments/explanations of their ratings.

Overall Holistic Evaluation

At the end of the workshop, panelists completed a final holistic evaluation form, which asked the panelists to provide overall, holistic evaluations of the (a) alignment between items and standards, (b) consistency between the DOK of standards and the DOK of items linked to those standards, and (c) quality of items for allowing students to demonstrate their learning. The evaluation forms included space for panelists to enter qualitative feedback.

Alignment Workshop

HumRRO collected the alignment data during a 2–3-day workshop (depending on grade level) in Louisville, Kentucky on October 5–7, 2017. The following information regarding the subject/grade level panel groups, panelists, training, materials, and workshop is provided to describe how the alignment method was operationalized.

Subject and Grade Panel Groups

We reviewed alignment of items and standards for the following assessments administered during the 2016–17 academic year:

- SC READY ELA grades 3 - 8
- SC READY Math grades 3 - 8
- EOCEP English 1 fall/winter
- EOCEP English 1 spring
- EOCEP Biology 1 fall/winter
- EOCEP Biology 1 spring

The alignment workshop involved six panel groups: (a) five educators for the SC READY ELA grades 3–5 panel, (b) six educators for the SC READY ELA grades 6–8 panel, (c) six educators for the SC READY math grades 3–5 panel, (d) six educators for the SC READY math grades 6–8 panel, (e) five educators for the English 1 (fall/winter and spring) panel, and (f) five educators for the Biology 1 (fall/winter and spring) panel.

Panelists

As suggested by the Education Oversight Committee (EOC), to maintain external independence we recruited Kentucky educators who were experienced at implementing rigorous content standards, including Common Core-based content standards. The Kentucky Academic Standards are similar to the South Carolina Standards in both organization and content. We created crosswalks between the Kentucky Academic Standards and the South Carolina Standards to demonstrate this similarity. (See the report Addendum for crosswalks between the two sets of standards for ELA grades 3–8, math grades 3–8, English 1, and Biology 1, respectively.)

Educators were selected based on their prior experience teaching the content areas and grade levels. Each panel included at least one nationally recognized content expert and most panels included multiple teachers who were National Board Certified Teachers (NBCTs). Moreover, several of the recruited teachers had participated as content experts in prior alignment studies for other states and/or for national testing programs. Table 2.1 presents professional and demographic characteristics of the panelists.

Table 2.1. Professional and Demographic Characteristics of Panelists

No. of Panelists	Experience		Education			Gender		School		
	Average Teaching Years (SD)	No. of NBCTs	No. with Bachelor Degree	No. with Masters Degree	No. with EdS	No. of Females	No. of Males	No. Urban	No. Sub-urban	No. Rural
33	14.00 (7.21)	15[a]	6	24	3	29	4	8	14	11

Note. NBCTs stands for National Board Certified Teachers and EdS stands for Education Specialist.
[a] 12 teachers completed NBCT certification and three were in the process of being certified.

Facilitator Training

Prior to the workshop, facilitators (i.e., leaders of each panel group) attended a 3-hour training session that included an overview of the South Carolina assessment system, alignment process steps, and examples of the rating forms. The alignment steps for facilitators were summarized in a Facilitator Instructions document with specific procedural and annotated guidance to ensure the facilitators provided consistent facilitation across panels. Facilitators participated in a detailed review of the Facilitator Instructions document in combination with the corresponding panelist rating forms.

Panelist Training

Panelist training was conducted in two ways: (a) alignment familiarization training on Day 1 of the workshop as a full group and (b) targeted procedural training in specific panel groups prior to starting each alignment task. The full group training focused on the South Carolina assessment system and included information regarding the roles of the Executive Oversight Committee, SCDE, DRC, HumRRO, and panelists; definition of alignment; why alignment is important; alignment process; and cognitive complexity. The panel-specific training focused on specific task processes, rating definitions, navigation and use of rating forms, and calibration activities to reinforce panelists' shared understanding. All panelists signed non-disclosure agreements.

Materials

During the workshop, panelists evaluated the alignment between the standards and test items by reviewing paper copies of test items (screen shots from the online test system) and entering their ratings into Excel® rating forms. The item presentation and rating forms are discussed in more detail below.

Test Items and Forms. Panelists evaluated operational test items for the 2016–17 SC READY ELA and math assessments for grades 3–8, and the EOCEP fall/winter 2016–17 and spring 2017 English 1 and Biology 1 assessments. The assessments are administered online and via paper administration; however, aside from the technology-enhanced items on the online assessments, the items are essentially identical. For the technology-enhanced items on the online assessments, HumRRO printed screen shots of the technology enhancements (e.g., dropdown menus) so panelists would understand what students experienced when taking the online assessment. Table 2.2 lists the number of items from each grade/form.[15] Although there were some duplicate items on the fall/winter and spring forms for English 1 and Biology 1, items were only reviewed once. Because the test items are secure, this report does not include any examples of items or references to specific item content.

Instructions and Rating Forms. Panelists were given instructions describing the rating tasks, codes to be entered into the Excel rating forms, supporting materials, and laptop computers loaded with the excel rating forms (see Appendix A for example instructions). Panelists completed three rating forms. The first was completed as a group (via consensus) to provide depth of knowledge (DOK) ratings for the content standards and indicators (see Appendix B). The second form was completed by consensus to compare the test blueprints to the full content standards (see Appendix C). The third form, an item rating form, captured individual ratings on the item linkage to standard/indicator, item DOK, and item quality (see Appendix D).

Table 2.2. Number of SC READY and EOCEP Items Reviewed by Each Panel

Subject	Grade or Form	Items	Total Items
SC READY ELA 3–5	3	69	207
	4	69	
	5	69	
SC READY ELA 6–8	6	81	243
	7	81	
	8	81	
SC READY Math 3–5	3	50	162
	4	56	
	5	56	
SC READY Math 6–8	6	60	182
	7	60	
	8	62	
English 1	Winter 2016-17	55	110
	Spring 2017	55	
Biology 1	Winter 2016-17	60	120
	Spring 2017	60	

[15] The documentation requested from DRC for the alignment task included a request for "test maps with item meta data (e.g., item ID, assigned standard link, assigned DOK, test form number, item sequence, item difficulty, item type, items status such as operational or field test)." For this request, DRC provided files labeled, "Test Maps and Forms." Item information was also requested for Task 6 (Review of Item Parameters). The documentation requested for Task 6 included "CTT statistics (*p*-values, point-biserials)." For this request, DRC provided files labeled "Item Analysis." There were some minor differences in the number of items included in these files for some grades and content areas; thus, there are some minor differences between the number of items reported in Table 2.2 and the number of items reported in the tables in Chapter 6 (Task 6).

Workshop Activities

After the group-wide training, panelists split into their respective panels and received panel-specific training. HumRRO facilitators led panels through the workshop activities. HumRRO facilitators provided general suggestions and comments when appropriate; however, they emphasized they would not provide explicit direction on how to rate standards or items because panelists were valued as the content experts. Each panelist used Excel rating forms already loaded onto their assigned laptop and HumRRO facilitators provided support as needed for working with the electronic rating forms.

Activity 1. Panelists first provided DOK ratings for the South Carolina standards and indicators. Panelists independently assigned a DOK level to one standard or indicator, for the first few standards/indicators, and then discussed their individual ratings until the group reached consensus. When all panelists felt comfortable with the task they followed a similar process by providing independent ratings for several standards at a time, and then discussing until they reached consensus for each standard/indicator. If a panel was unable to reach a majority consensus rating, the highest DOK level discussed was entered as the final DOK rating for the standard/indicator.

Activity 2. Next, the panelists reached a consensus decision on whether the test blueprint adequately covers the essential knowledge and skills included in the South Carolina Standards. To make this decision, panelists reviewed the full set of South Carolina Standards for their assigned content domain(s) (e.g., ELA grade 3) and considered if any critical standards were omitted from the test blueprint or if the test blueprint overemphasized a content domain. They engaged in discussion until they arrived at a consensus decision. The facilitator recorded the panel’s feedback and comments in an Excel form.

Activity 3. Next, panelists received specific instructions to rate the test items. As a calibration activity, panelists rated the first few items individually and then discussed the ratings as a panel. Once panelists were comfortable making ratings and calibrated in their ratings, they continued the item rating activity independently. A recalibration activity was conducted at the beginning of the second and third days of the workshop to ensure panelists maintained a common rating approach.

Panelists rated the individual items on (a) depth of knowledge required to correctly respond to the item, and (b) the degree of alignment (i.e., how well the item linked to the identified standard/indicator). If the panelists felt the item did not fully match the standard/indicator to which it was linked, they entered their explanation for why the content in the item was not fully covered by the linked standard/indicator. If appropriate, panelists identified a secondary standard/indicator they believed was more closely linked to the item.

Panelists also rated each item on several aspects of item quality: (a) clarity of presentation, (b) accuracy of content, (c) grade-level appropriateness, (d) supports research-based instruction, and (e) unbiased content or presentation. Panelists entered “yes” if the item met the quality indicator and “no” if the item did not meet the quality indicator. If “no” was selected, panelists explained their reasoning for why the item did not meet the item quality indicator.

At the end of each day and before the end of the final day, the facilitator reviewed the individual panelists’ ratings for substantive discrepancies (e.g., one panelist rated an item DOK as a level “1” and all other panelists rated it a level “3”). When widely discrepant ratings were discovered, the facilitator engaged the panelists in a discussion to ensure the discrepancy was not due to a misunderstanding or mistake.

Activity 4. The final activity was for panelists to provide overall, holistic evaluations of the (a) alignment between items and standards, (b) consistency between the DOK of standards and the DOK of items linked to those standards, and (c) quality of items for allowing students to demonstrate their learning. The evaluation forms included space for panelists to enter qualitative feedback. Panelists also provided feedback on the quality of the training, rating processes, and workshop materials (see Appendix E for the results of the panelists' feedback).

Task 2: Results

The following section summarizes results from the analyses of panelists' alignment and item quality ratings.

Interrater Reliability

Table 2.3 presents the interrater reliability coefficients for panelists' independent ratings of item DOK. We used the intra-class correlation coefficient (ICC; Shrout & Fleiss, 1979) as a measure of consistency in the panelists' ratings. An ICC of .70 is generally considered sufficient for research purposes, although ICCs of .80 and above are preferred when ratings are used to make important or high-stakes decisions (e.g., promotion) (Graham, Milanowski, & Miller, 2012). As shown in Table 2.3, panelists demonstrated strong levels of consistency for the majority of independent item DOK ratings—that is, .70 and above, with the exception of English 1 fall/winter (i.e., ICC = .618), which was just slightly below the benchmark. Panelists' were very consistent on their independent ratings on quality of item link, which resulted in very low variance among raters; thus, we do not report ICCs on the quality of link ratings, as reporting the ICC values based on this low variance would be misleading.

Table 2.3. Interrater Consistency Coefficients on Item DOK Ratings

Content Area/Grade	ICC
ELA 3	0.965
ELA 4	0.975
ELA 5	0.954
ELA 6	0.854
ELA 7	0.885
ELA 8	0.851
MATH 3	0.847
MATH 4	0.831
MATH 5	0.867
MATH 6	0.754
MATH 7	0.833
MAT 8	0.811
ENG 1 Fall/ Winter	0.618
ENG 1 Spring	0.776
BIO 1 Fall/ Winter	0.813
BIO 1 Spring	0.919

SC READY ELA Alignment Results

Items Represent Intended Content

The percentage of items at each level of alignment—fully aligned, partially aligned, and not aligned—was calculated for each panelist and averaged across panelists. The quality of the link was calculated only using the primary linked standard. That is, when/if panelists entered a secondary standard/indicator, the quality of link for the primary standard was used for this analysis. Table 2.4 provides the average percentage of items at each level of alignment. As can be seen, the percentage of items that were cumulatively rated partially or fully aligned is over 93% across all ELA grades.

Table 2.4. Percentage of SC READY ELA Items at Alignment Levels, by Grade

Content Area/Grade	% Items Not Aligned	% Items Partially Aligned	% Items Fully Aligned	% Items Partially or Fully Aligned
ELA 3	0.00	0.29	99.71	100.00
ELA 4	2.62	0.29	97.09	97.38
ELA 5	1.45	0.00	98.55	98.55
ELA 6	4.33	5.77	89.9	95.67
ELA 7	6.26	4.8	88.94	93.74
ELA 8	1.23	3.09	95.68	98.77

Items Represent Intended Categories

As a check of alignment between the test blueprint and test items, this criterion was met when the actual numbers of items linked to each category were within the target ranges specified on the test blueprints. For this criterion, we compared the number of items specified for each blueprint category to the actual number of items that panelists (within a panel group) linked to each category. To calculate the number of items linked to each category, the number of items each panelist rated as aligned or partially aligned with that category was first calculated and then averaged across all panelists (within each panel group). If a panelist rated an item not aligned with the identified standard/indicator, they entered a secondary standard/indicator. In this case, analyses were conducted with the secondary standard rather than the primary identified standard. The items that were rated not aligned and for which no secondary standard was entered were excluded from the analysis.

As shown in Table 2.5, the mean number of items linked to each domain, when rounded, was within the target number of items specified in the test blueprint.

Additional detail on the mean number of items linked to each reporting category (i.e., a finer-grain category than Domain) is provided in Table 2.6. As shown in this table, the mean number of items linked to each reporting category was within the targeted number of items specified in the test blueprint.

Table 2.5. Summary of SC READY ELA Blueprint Content Coverage Results, by Domain within Grade

Content Area/ Grade	Domain	Mean Number of Linked Items	SD	Target Number of Items
ELA 3	Reading - Literary Text	19.20	0.45	19
ELA 3	Reading - Informational Text	18.80	0.45	19
ELA 3	Writing	21.00	0.00	30[a]
ELA 3	Inquiry	10.00	0.00	
ELA 4	Reading - Literary Text	19.00	0.00	19
ELA 4	Reading - Informational Text	18.80	0.45	19
ELA 4	Writing	25.00	0.00	30
ELA 4	Inquiry	6.00	0.00	
ELA 5	Reading - Literary Text	19.00	0.00	19
ELA 5	Reading - Informational Text	19.00	0.00	19
ELA 5	Writing	24.00	0.00	30
ELA 5	Inquiry	7.00	0.00	
ELA 6	Reading - Literary Text	21.83	0.41	21
ELA 6	Reading - Informational Text	29.00	0.00	29
ELA 6	Writing	22.00	0.00	30
ELA 6	Inquiry	8.00	0.00	
ELA 7	Reading - Literary Text	20.83	0.41	21
ELA 7	Reading - Informational Text	28.83	0.75	29
ELA 7	Writing	23.50	0.55	30
ELA 7	Inquiry	6.67	0.82	
ELA 8	Reading - Literary Text	21.00	0.00	21
ELA 8	Reading - Informational Text	29.00	0.00	29
ELA 8	Writing	24.00	0.00	30
ELA 8	Inquiry	7.00	0.00	

[a]According to the blueprint, Writing/Inquiry has 46 possible points. However, these standards include a text-dependent analysis item, that has 16 possible points, which was not included in the alignment review. Therefore, the value of this item was subtracted from the total points possible for Writing/Inquiry.

Table 2.6. Summary of SC READY ELA Blueprint Content Coverage Results, by Reporting Category within Grade and Domain

Content Area/ Grade	Domain	Reporting Category	Mean Number of Linked Items	SD	Target Number of Items
ELA 3	Reading - Literary Text	Meaning and Context	11.00	0.00	9-11
ELA 3	Reading - Literary Text	Language, Craft, and Structure	8.20	0.45	8-10
ELA 3	Reading - Informational Text	Meaning and Context	10.00	0.00	9-11
ELA 3	Reading - Informational Text	Language, Craft, and Structure	8.80	0.45	8-10
ELA 3	Writing/ Inquiry	Meaning, Context, and Craft	14.00	0.00	10-17
ELA 3		Language	7.00	0.00	7-14
ELA 3		Inquiry	10.00	0.00	6-10
ELA 4	Reading - Literary Text	Meaning and Context	9.00	0.00	9-11
ELA 4	Reading - Literary Text	Language, Craft, and Structure	10.00	0.00	8-10
ELA 4	Reading - Informational Text	Meaning and Context	9.00	0.00	9-11
ELA 4	Reading - Informational Text	Language, Craft, and Structure	9.80	0.45	8-10
ELA 4	Writing/ Inquiry	Meaning, Context, and Craft	13.00	0.00	10-17
ELA 4		Language	12.00	0.00	7-14
ELA 4		Inquiry	6.00	0.00	6-10
ELA 5	Reading - Literary Text	Meaning and Context	11.00	0.00	9-11
ELA 5	Reading - Literary Text	Language, Craft, and Structure	8.00	0.00	8-10
ELA 5	Reading - Informational Text	Meaning and Context	10.00	0.00	9-11
ELA 5	Reading - Informational Text	Language, Craft, and Structure	9.00	0.00	8-10
ELA 5	Writing/ Inquiry	Meaning, Context, and Craft	17.00	0.00	10-17
ELA 5		Language	7.00	0.00	7-14
ELA 5		Inquiry	7.00	0.00	6-10
ELA 6	Reading - Literary Text	Meaning and Context	11.00	0.00	11-13
ELA 6	Reading - Literary Text	Language, Craft, and Structure	10.83	0.41	8-10
ELA 6	Reading - Informational Text	Meaning and Context	16.00	0.00	15-17
ELA 6	Reading - Informational Text	Language, Craft, and Structure	13.00	0.00	12-14
ELA 6	Writing/ Inquiry	Meaning, Context, and Craft	13.00	0.00	10-17
ELA 6		Language	9.00	0.00	7-14
ELA 6		Inquiry	8.00	0.00	6-10
ELA 7	Reading - Literary Text	Meaning and Context	11.83	0.41	11-13
ELA 7	Reading - Literary Text	Language, Craft, and Structure	9.00	0.00	8-10
ELA 7	Reading - Informational Text	Meaning and Context	13.50	0.55	15-17
ELA 7	Reading - Informational Text	Language, Craft, and Structure	15.33	1.03	12-14
ELA 7	Writing/ Inquiry	Meaning, Context, and Craft	12.50	0.55	10-17
ELA 7		Language	11.00	0.00	7-14
ELA 7		Inquiry	6.67	0.82	6-10
ELA 8	Reading - Literary Text	Meaning and Context	13.00	0.00	11-13
ELA 8	Reading - Literary Text	Language, Craft, and Structure	8.00	0.00	8-10
ELA 8	Reading - Informational Text	Meaning and Context	16.00	0.00	15-17
ELA 8	Reading - Informational Text	Language, Craft, and Structure	13.00	0.00	12-14
ELA 8	Writing/ Inquiry	Meaning, Context, and Craft	16.00	0.00	10-17
ELA 8		Language	8.00	0.00	7-14
ELA 8		Inquiry	7.00	0.00	6-10

Evaluation of Test Blueprint

Panelists discussed whether the test blueprint adequately covers what students should know and be able to do, as described in the standards. For all ELA grades, the panelists felt the blueprints adequately cover what the students should know and be able to do; however, panelists provided several suggestions on how the blueprint could be improved.

For ELA grades 3–5, the panelists expressed that the Inquiry standard was difficult to assess via the format of the assessment (i.e., primarily multiple-choice items). Consequently, the panelists suggested removing the Inquiry standard from the assessment and distributing those questions to cover word analysis and grade-level phonics in the Principles of Reading strand.

For ELA grades 6–8, the panelists similarly expressed that the Inquiry standard was difficult to assess via the format of the assessment. These panelists suggested replacing the Inquiry standard with the Communication standard. The panelists also felt there was little difference between the grades 6 and 7 standards, although they felt the grade 7 assessment was considerably more difficult than grade 6 assessment. In addition, the panelists expressed some concern that the standards had the same weights across grades 6, 7, and 8; they suggested the standards should be weighted differently across these grades to reflect the expected increase in skills.

DOK Consistency

Webb's DOK consistency criterion examines the 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. Per Webb's guidance, at least 50% of the items linked to the standard/indicator must be at or above the DOK level for that standard/indicator. Table 2.7 provides a summary, by grade, of the consistency between the DOK of the standards and the DOK of the items linked to those standards. In grades 4 and 6, the DOK level of over 50% of items was at or above the DOK level of the standards; for the other grades, the DOK level of the majority of items was below the DOK level of the standards.

Table 2.7. DOK Consistency Results for SC READY ELA, by Grade

Content Area/Grade	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
	Mean	SD	Mean	SD	Mean	SD	Mean
ELA 3	55.10	4.10	38.80	4.20	6.10	0.60	44.90
ELA 4	46.50	2.80	50.60	2.90	2.90	0.00	53.50
ELA 5	73.80	3.30	25.20	3.40	0.90	0.80	26.20
ELA 6	47.70	6.30	41.90	5.20	10.40	3.20	52.30
ELA 7	68.40	3.40	29.70	5.20	1.90	2.10	31.60
ELA 8	69.10	8.50	27.80	6.80	3.10	2.70	30.90

Taking a finer-grain look at DOK consistency (i.e., by domain), we see in Table 2.8 that the Inquiry domain, in particular, tended to have items with lower DOKs than the standards to which they were linked. Additional detail on DOK consistency by ELA reporting category (finest-grain level) is provided in Appendix F.

Table 2.8. DOK Consistency Results for SC READY ELA, by Domain within Grade

Content Area/ Grade	Domain	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
		Mean	SD	Mean	SD	Mean	SD	Mean
ELA 3	Reading - Literary Text	40.60	5.20	47.90	6.00	11.50	2.40	59.40
ELA 3	Reading - Informational Text	49.00	3.70	40.40	3.90	10.60	0.30	51.00
ELA 3	Writing	64.80	5.40	35.20	5.40	0.00	0.00	35.20
ELA 3	Inquiry	74.00	11.40	26.00	11.40	0.00	0.00	26.00
ELA 4	Reading - Literary Text	24.20	4.70	70.50	4.70	5.30	0.00	75.80
ELA 4	Reading - Informational Text	54.30	3.80	45.70	3.80	0.00	0.00	45.70
ELA 4	Writing	44.80	1.80	51.20	1.80	4.00	0.00	55.20
ELA 4	Inquiry	100.00	0.00	0.00	0.00	0.00	0.00	0.00
ELA 5	Reading - Literary Text	75.80	6.00	21.10	6.40	3.20	2.90	24.30
ELA 5	Reading - Informational Text	93.70	4.40	6.30	4.40	0.00	0.00	6.30
ELA 5	Writing	53.30	3.50	46.70	3.50	0.00	0.00	46.70
ELA 5	Inquiry	100.00	0.00	0.00	0.00	0.00	0.00	0.00
ELA 6	Reading - Literary Text	69.40	8.50	30.60	8.50	0.00	0.00	30.60
ELA 6	Reading - Informational Text	44.30	4.00	36.80	3.60	19.00	3.60	55.80
ELA 6	Writing	26.30	12.20	61.70	10.10	11.90	9.80	73.60
ELA 6	Inquiry	56.30	23.40	39.60	18.40	4.20	6.50	43.80
ELA 7	Reading - Literary Text	66.40	10.30	33.60	10.30	0.00	0.00	33.60
ELA 7	Reading - Informational Text	83.90	6.30	16.10	6.30	0.00	0.00	16.10
ELA 7	Writing	46.00	5.60	47.60	7.00	6.40	7.10	54.00
ELA 7	Inquiry	88.90	20.20	11.10	20.20	0.00	0.00	11.10
ELA 8	Reading - Literary Text	85.70	8.00	14.30	8.00	0.00	0.00	14.30
ELA 8	Reading - Informational Text	62.10	8.70	37.90	8.70	0.00	0.00	37.90
ELA 8	Writing	61.10	10.10	28.50	3.10	10.40	9.00	38.90
ELA 8	Inquiry	76.20	14.80	23.80	14.80	0.00	0.00	23.80

Evaluation of Item Quality

Panelists independently rated each item on several aspects of item quality: (a) clarity of presentation, (b) accuracy of content, (c) grade-level appropriateness, (d) supports research-based instruction, and (e) unbiased content or presentation. When averaged across panelists, over 97% of the items were considered clear, accurate, grade appropriate, supporting research-based instruction, and free of bias across grades 3–8 (see Table 2.9).

Table 2.9. Percentage of SC READY ELA Items with Positive Ratings on Each Item Quality Indicator, by Grade

Content Area/Grade	Clarity	Accuracy	Grade-level appropriate	Supports Research-based Instruction	Free from Bias
ELA 3	98.55	98.55	99.42	100.00	100.00
ELA 4	98.55	98.55	100.00	100.00	100.00
ELA 5	100.00	100.00	100.00	100.00	100.00
ELA 6	97.78	99.03	99.68	99.68	100.00
ELA 7	97.69	99.37	98.95	100.00	98.95
ELA 8	99.51	99.75	100.00	100.00	100.00

Overall Holistic Evaluation

At the end of the workshop, each panelist completed a final, overall evaluation form in which the panelist was asked to provide a final holistic rating of the alignment between items and standards. This rating was made on a 5-point scale, where 5 = perfectly aligned and 1 = not aligned. Panelists were also asked to share qualitative feedback on the alignment. All five grade 3–5 panelists and four of the six grade 6–8 panelists believed the overall alignment of items and standards was good. Two grade 6–8 panelists believed the overall alignment needed some improvement; however, their comments indicated this rating applied to the grades 6 and 7 assessments, while the overall alignment for the grade 8 assessment was good. In addition, one grade 6–8 panelist commented that there was minimal coverage of Argumentative standards and two panelists commented that the Informational Texts were over-represented. Both grade 3–5 and 6–8 ELA panels suggested the SC READY assessments could be improved by eliminating test questions for the Inquiry standard. Panelists felt those standards would be better assessed in other ways, such as performance-based testing.

Regarding perceptions of the overall consistency between the DOK of standards and the DOK of items, four of the five grade 3–5 panelists indicated that the item DOK levels were lower than the DOK levels of the standards to which they were linked. For the grade 6–8 panel, half the panelists stated the items were generally written below the DOK of the standard, particularly at grades 6 and 7.

Regarding perceptions of overall item quality, one grade 6–8 panelist commented that there was a bias towards female protagonists in most literary passages. Also, two grade 3–5 panelists stated that test items assessed the content in the same way when they covered the same standard/indicator.

SC READY Math Alignment Results

Items Represent Intended Content

The percentage of items at each level of alignment—fully aligned, partially aligned, and not aligned—was calculated for each panelist and averaged across panelists. The quality of the link was calculated only for the primary linked standard. That is, when panelists entered a secondary standard/indicator, the quality of link for the primary standard was used for this analysis. Table 2.10 provides the average percentage of items at each level of alignment. As can be seen, the percentage of items that were cumulatively rated partially or fully aligned is over 96% across all math grades.

Table 2.10. Percentage of SC READY Math Items at Alignment Levels, by Grade

Content Area/ Grade	% Items Not Aligned	% Items Partially Aligned	% Items Fully Aligned	% Items Partially or Fully Aligned
Math 3	4.00	2.33	93.67	96.00
Math 4	0.00	0.00	100.00	100.00
Math 5	0.00	0.60	99.40	100.00
Math 6	0.00	6.98	93.02	100.00
Math 7	1.67	8.61	89.72	98.33
Math 8	3.49	4.03	92.47	96.50

Items Represent Intended Categories

For this criterion, we compared the number of items specified for each blueprint category to the actual number of items panelists (within a panel group) linked to each category. To calculate the number of items linked to each category, the number of items each panelist rated as aligned or partially aligned with that category was first calculated and then averaged across all panelists (within each panel group). If a panelist rated an item not aligned with the identified standard/indicator, they entered a secondary standard/indicator. In this case, analyses were conducted with the secondary standard rather than the primary identified standard. The items that were rated not aligned and for which no secondary standard was entered were excluded from the analysis.

As shown in Table 2.11, the mean number of items linked to each standard, when rounded, was within the target number of items specified in the test blueprint.

Evaluation of Test Blueprint

Panelists discussed whether the test blueprint adequately covers what students should know and be able to do. Overall, the panelists felt that the grade 4 math test blueprint adequately covers what students should know and be able to do per the standards. For the other SC READY math grades, the panelists felt the coverage of the standards by the test blueprint could be improved.

For grade 3, the panelists felt the “Number Sense and Base Ten” and “Number Sense and Operations – Fractions” categories should have more emphasis given they are the “foundation of future math understanding.” They also felt there was not enough variety of graphing data items and there was an overuse of interpreting bar graphs.

For grade 5, the panelists felt that there was an over-emphasis of standard 5.G.2 (Geometry, about coordinates) and standard 5.G.1 (Geometry, define a coordinate system), and that the items that addressed those standards required low-level thinking. These panelists suggested increasing the allocation of points for “Number Sense and Base Ten,” “Number Sense and Operations – Fractions,” and “Algebraic Thinking and Operations” to 11–14 points to reflect the number of standards and collective complexity of standards within those categories. They also suggested reducing the number of points allocated to Geometry and Measurement and Data Analysis to 10–12 points to reflect the lower number of standards within those categories.

Table 2.11. Summary of SC READY Math Blueprint Content Coverage Results, by Standard within Grade

Grade	Standard	Mean Number of Linked Items	*SD*	Target number of items
3	Number Sense and Base Ten	7.00	0.00	7-9
	Number Sense – Fractions	8.00	0.00	7-9
	Algebraic Thinking and Operations	13.00	0.00	13-16
	Geometry	9.00	0.00	7-9
	Measurement and Data Analysis	13.00	0.00	13-16
4	Number Sense and Base Ten	12.00	0.00	10-12
	Number Sense – Fractions	12.00	0.00	11-14
	Algebraic Thinking and Operations	12.00	0.00	11-14
	Geometry	9.00	0.00	8-10
	Measurement and Data Analysis	11.00	0.00	11-14
5	Number Sense and Base Ten	10.00	0.00	10-13
	Number Sense – Fractions	12.00	0.00	10-12
	Algebraic Thinking and Operations	13.00	0.00	10-13
	Geometry	10.00	0.00	10-12
	Measurement and Data Analysis	11.00	0.00	11-14
6	The Number System	14.00	0.00	12-15
	Ratios and Proportional Relationships	10.00	0.00	8-10
	Expressions, Equations, and Inequalities	14.83	0.41	12-15
	Geometry and Measurement	9.00	0.00	8-10
	Data Analysis and Statistics	11.67	0.52	11-13
7	The Number System	13.00	0.00	13-15
	Ratios and Proportional Relationships	10.00	0.00	8-10
	Expressions, Equations, and Inequalities	12.00	0.00	12-14
	Geometry and Measurement	12.00	0.00	11-13
	Data Analysis, Statistics, and Probability	13.00	0.00	13-15
8	The Number System	9.00	0.00	9-11
	Functions	13.83	0.41	11-14
	Expressions, Equations, and Inequalities	16.17	0.41	12-16
	Geometry and Measurement	14.00	0.00	12-16
	Data Analysis, Statistics, and Probability	9.00	0.00	9-11

For grade 6, panelists commented that the Number System and Expressions, Equations, and Inequalities categories were appropriately weighted on the test blueprint. They felt the weight for Ratios and Proportional Relationships should be increased because they felt that category was more important than Geometry and Measurement. They also felt Data Analysis and Statistics should be given less weight. These panelists suggested the Number System, Expressions, Equations, and Inequalities, and Ratios and Proportional Relationships categories each should be weighted 25%, while the Geometry and Measurement and Data Analysis and Statistics categories each be weighted 12.5%.

For grade 7, panelists felt the proportional weightings should replicate their grade 6 recommendations.

For grade 8, the panelists felt the blueprint more accurately reflected what students should know and do than did the blueprints for the grades 6 and 7 assessments; however, they still suggested some improvements. Specifically, they suggested the Number System and Data Analysis, Statistics, and Probability categories should have less weight, and the weight for Functions, Geometry and Measurement, and Expressions, Equations, and Inequalities, should be increased.

DOK Consistency

Webb's DOK consistency criterion determines 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. Per Webb's guidance, at least 50% of the items linked to the standard/indicator must be at or above the DOK level for that standard/indicator. Table 2.12 provides a summary, by grade, of the consistency between the DOK of the standards and the DOK of the items linked to those standards. On average, the DOK level of over 50% of the math items at all grades was at or above the DOK level of the standards.

Table 2.12. DOK Consistency Results for SC READY Math, by Grade

Content Area/ Grade	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
	Mean	SD	Mean	SD	Mean	SD	Mean
Math 3	44.30	8.10	44.30	7.80	11.30	3.00	55.60
Math 4	27.70	6.80	53.30	5.50	19.00	5.50	72.30
Math 5	25.30	2.60	58.00	4.00	16.70	3.50	74.70
Math 6	26.60	17.40	66.40	20.80	7.00	4.40	73.40
Math 7	25.00	15.80	67.20	17.40	7.80	3.30	75.00
Math 8	26.20	16.90	71.30	16.40	2.40	1.70	73.70

Table 2.13 provides a summary of DOK consistency, by standard and within grade. As can be seen, the DOK level of over 50% of the items was at or above the DOK level of the standards for the majority of the standards. The exceptions were grade 3 Geometry and grade 3 Measurement and Data Analysis. For these two standards, the majority of items were rated below the DOK of their linked standards.

Table 2.13. DOK Consistency Results for SC READY Math, by Standard within Grade

Content Area/ Grade	Standard	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above
		Mean	SD	Mean	SD	Mean	SD	SD
Math 3	Number Sense and Base Ten	26.20	14.00	73.80	14.00	0.00	0.00	73.80
Math 3	Number Sense – Fractions	18.80	10.50	56.30	10.50	25.00	7.90	81.30
Math 3	Algebraic Thinking and Operations	46.20	16.90	38.50	15.40	15.40	4.90	53.90
Math 3	Geometry	61.10	6.10	38.90	6.10	0.00	0.00	38.90
Math 3	Measurement and Data Analysis	56.40	4.00	30.80	4.90	12.80	4.00	43.60
Math 4	Number Sense and Base Ten	31.90	18.60	58.30	21.10	9.70	6.30	68.00
Math 4	Number Sense – Fractions	45.80	10.20	45.80	10.20	8.30	0.00	54.10
Math 4	Algebraic Thinking and Operations	25.00	10.50	68.10	12.30	6.90	13.40	75.00
Math 4	Geometry	7.40	5.70	59.30	13.50	33.30	14.10	92.60
Math 4	Measurement and Data Analysis	22.70	5.00	34.80	10.60	42.40	7.40	77.20
Math 5	Number Sense and Base Ten	23.30	13.70	55.00	13.80	21.70	7.50	76.70
Math 5	Number Sense – Fractions	31.90	3.40	62.50	4.60	5.60	6.80	68.10
Math 5	Algebraic Thinking and Operations	14.10	7.60	84.60	8.40	1.30	3.10	85.90
Math 5	Geometry	28.30	9.80	46.70	10.30	25.00	5.50	71.70
Math 5	Measurement and Data Analysis	30.30	4.70	34.80	3.70	34.80	3.70	69.60
Math 6	The Number System	13.10	8.40	77.40	15.30	9.50	8.70	86.90
Math 6	Ratios and Proportional Relationships	36.70	22.50	63.30	22.50	0.00	0.00	63.30
Math 6	Expressions, Equations, and Inequalities	41.10	38.00	58.90	38.00	0.00	0.00	58.90
Math 6	Geometry and Measurement	16.70	15.30	70.40	23.00	13.00	8.40	83.40
Math 6	Data Analysis and Statistics	22.90	10.60	62.90	17.70	14.30	8.60	77.20
Math 7	The Number System	24.40	16.40	73.10	18.70	2.60	4.00	75.70
Math 7	Ratios and Proportional Relationships	26.70	31.40	73.30	31.40	0.00	0.00	73.30
Math 7	Expressions, Equations, and Inequalities	23.60	29.50	76.40	29.50	0.00	0.00	76.40
Math 7	Geometry and Measurement	13.90	10.10	56.90	13.40	29.20	4.60	86.10
Math 7	Data Analysis, Statistics, and Probability	35.90	6.30	57.70	11.70	6.40	7.60	64.10
Math 8	The Number System	40.70	19.50	59.30	19.50	0.00	0.00	59.30
Math 8	Functions	22.40	18.70	75.10	21.40	2.50	3.80	77.60
Math 8	Expressions, Equations, and Inequalities	7.10	11.40	85.60	12.10	7.30	7.30	92.90
Math 8	Geometry and Measurement	42.90	20.20	57.10	20.20	0.00	0.00	57.10
Math 8	Data Analysis, Statistics, and Probability	25.90	28.70	74.10	28.70	0.00	0.00	74.10

Evaluation of Item Quality

When averaged across panelists, over 95% of the grades 3–5 items were rated clear, accurate, grade appropriate, supporting research-based instruction, and free of bias (see Table 2.14).

Table 2.14. Percentage of SC READY Math Items with Positive Ratings on Each Item Quality Indicator, by Grade

Content Area/ Grade	Clarity	Accuracy	Grade Appropriate	Supports Research- based Instruction	Free of Bias
Math 3	97.33	99.00	98.33	99.00	100.00
Math 4	98.81	100.00	100.00	100.00	99.70
Math 5	99.70	100.00	100.00	100.00	100.00
Math 6	96.93	99.44	98.04	95.25	98.60
Math 7	95.28	99.17	98.61	98.89	100.00
Math 8	98.66	98.66	99.46	97.58	100.00

Overall Holistic Evaluation

At the end of the workshop, each panelist provided a final holistic rating (i.e., 5 = perfectly aligned, 1 = not aligned) of the alignment between items and standards. Panelists also shared qualitative feedback regarding the item-standard alignment. All grades 3–5 panelists and five (of six) grades 6–8 panelists rated the alignment as good. Their qualitative comments regarding overall alignment were positive; however, one panelist in each group indicated there were too many questions that assessed patterns and coordinate graphing.

Regarding perceptions of the overall consistency between the DOK of standards and the DOK of items, panelists in both grade span groups indicated that there was reasonable consistency.

Regarding perceptions of overall item quality, the majority of panelists reported that the items were age appropriate, straight-forward, and fair. A few panelists commented that alternative items types (e.g., performance-based, open response) would allow for greater demonstration of student learning.

English 1 Alignment Results

Items Represent Intended Content

The percentage of English 1 items at each level of alignment—fully aligned, partially aligned, and not aligned—was calculated for each panelist and averaged across panelists. The quality of the link was calculated only for the primary linked standard. Table 2.15 provides the average percentage of items at each level of alignment. As can be seen, the percentage of items cumulatively rated as partially and fully aligned was just over 86% for the fall/winter form and nearly 100% for the spring form.

Table 2.15. Percentage of English 1 Items at Alignment Levels, by Form

Form	% Items Not Aligned	% Items Partially Aligned	% Items Fully Aligned	% Items Partially or Fully Aligned
Fall/Winter	13.50	6.57	79.93	86.50
Spring	0.36	1.82	97.82	99.64

Items Represent Intended Categories

This criterion was met when the actual number of items linked to each blueprint category was within the target ranges specified on the test blueprints. For this criterion, we calculated the number of items each panelist rated as aligned or partially aligned for each blueprint category and then averaged across all panelists.

As shown in Table 2.16, the mean number of items linked to each standard was within the target number of items specified in the test blueprint for all standards (strands), with one exception. For Writing, the mean number of items linked to this standard was slightly below the target number of items for both the fall/winter and spring forms.

Table 2.16. Summary of English 1 Blueprint Content Coverage Results, by Standard within Form

Form	Standard	Mean Number of Linked Items	SD	Target Number of Items
Fall/ Winter	Inquiry	4.60	1.52	4-8
	Reading Literary Text	19.40	0.55	18-25
	Reading Informational Text	22.40	1.14	16-25
	Communication	3.80	0.45	2-6
	Writing	4.80	0.45	6-12
Spring	Inquiry	4.00	0.00	2-6
	Reading Literary Text	19.00	0.00	4-8
	Reading Informational Text	25.00	0.00	18-25
	Communication	2.00	0.00	16-25
	Writing	5.00	0.00	6-12

Evaluation of Test Blueprint

Overall, based on the standards, panelists felt the blueprint appropriately reflected what students should know and be able to do. However, some panelists noted the Inquiry standard was “not very realistic” to assess on a standardized test.

DOK Consistency

Table 2.17 provides a summary, by form, of the consistency between the DOK of the standards and the DOK of the linked English 1 items. For the fall/winter form, slightly less than 50% of the items received DOK ratings at or above the DOK ratings of their linked standards, while the DOK ratings of just over 50% of the items for the spring form were at or above the standard DOK ratings.

Table 2.17. DOK Consistency Results for English 1, by Form

Form	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
	Mean	SD	Mean	SD	Mean	SD	Mean
Fall/ Winter	56.40	5.30	38.90	4.40	4.70	2.80	43.60
Spring	46.90	7.90	50.90	7.40	2.20	0.80	53.10

Table 2.18 provides a summary, by standard, of DOK consistency for each English 1 form. For the fall/winter form, all the standards had fewer than 50% of their linked items rated at or above the DOK of the standard, with one exception. All the items linked to the Writing standard were at or above that standard's DOK level. In contrast, most of the items linked to the Writing standard for the spring form were rated below that standard's DOK level. Additionally, the majority of items linked to the Inquiry standard for the spring form were rated below that standard's DOK level. Reading Literary Text, Reading Informational Text, and Communication had the majority of their items rated at or above those standards' DOK levels.

Table 2.18. DOK Consistency Results for English 1, by Standard within Form

Form	Standard	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
		Mean	SD	Mean	SD	Mean	SD	
Fall/ Winter	Inquiry	56.80	7.10	36.40	13.30	6.90	9.60	43.30
	Reading Literary Text	64.90	11.40	35.10	11.40	0.00	0.00	35.10
	Reading Informational Text	60.00	10.30	35.60	9.80	4.40	3.10	40.00
	Communication	61.70	26.10	38.30	26.10	0.00	0.00	38.30
	Writing	0.00	0.00	76.00	26.10	24.00	26.10	100.00
Spring	Inquiry	85.00	13.70	15.00	13.70	0.00	0.00	15.00
	Reading Literary Text	47.40	17.80	52.60	17.80	0.00	0.00	52.60
	Reading Informational Text	36.80	8.20	59.20	7.70	4.00	2.80	63.20
	Communication	0.00	0.00	90.00	22.40	10.00	22.40	100.00
	Writing	84.00	16.70	16.00	16.70	0.00	0.00	16.00

Evaluation of Item Quality

When ratings were averaged across panelists, virtually all the items were considered clear, accurate, grade appropriate, supporting research-based instruction, and free of bias across both forms (see Table 2.19).

Table 2.19. Percentage of English 1 Items with Positive Ratings on Each Item Quality Indicator, by Form

Form	Clarity	Accuracy	Grade Appropriate	Supports Research- based Instruction	Free of Bias
Fall/ Winter	99.27	100.00	98.91	100.00	99.27
Spring	94.91	99.27	98.18	99.27	98.90

Overall Holistic Evaluation

Four of the five panelists rated the overall alignment of items and standards as good while one panelist indicated the overall alignment needs major improvement. It should be noted the comments provided from that panelist were not specific and suggested general disapproval of multiple-choice tests. Other panelists' comments indicated the fall/winter test form was not as well-aligned as the spring test form.

Panelists also commented about the consistency between the standards and the DOK of items linked to those standards. Three of the five panelists felt the item DOK levels were what they expected for the linked standards, while two felt the DOK levels were too low, particularly for the fall/winter test form.

Regarding the quality of items, all panelists reported that the fall/winter assessment was not as strong as the spring assessment with regard to standard representation, strength of link to standards, and representation of DOK levels.

Biology 1 Alignment Results

Items Represent Intended Content

The percentage of items at each level of alignment—fully aligned, partially aligned, and not aligned—was calculated for each panelist and then averaged. Table 2.20, which provides the average percentage of items at each level of alignment, shows that the percentage of items that were cumulatively rated partially or fully aligned was 95% or more for both forms.

Table 2.20. Percentage of Biology 1 Items at Alignment Levels, by Form

Form	% Items Not Aligned	% Items Partially Aligned	% Items Fully Aligned	% Items Partially or Fully Aligned
Fall/ Winter	5.02	3.34	91.64	94.98
Spring	2.68	5.69	91.64	97.33

Items Represent Intended Categories

As shown in Table 2.21, the mean number of items linked to each standard was within the target number of items specified in the test blueprint, with one exception. Standard HB.3 (The student will demonstrate the understanding that all essential processes within organisms require energy which in most ecosystems is ultimately derived from the Sun and transferred into chemical energy by the photosynthetic organisms of that ecosystem) on the spring form was one item short of meeting the target number of items.

Table 2.21. Summary of Biology 1 Blueprint Content Coverage Results, by Standard within Form

Form	Standard	Mean Number of Linked Items	SD	Target Number of Items
Fall/ Winter	H.B.1 The student will use the science and engineering practices, including the processes and skills of scientific inquiry, to develop understandings of science content	8.00	0.00	8-10
	H.B.2 The student will demonstrate the understanding that the essential functions of life take place within cells or systems of cells.	14.80	0.45	12-18
	H.B.3 The student will demonstrate the understanding that all essential processes within organisms require energy which in most ecosystems is ultimately derived from the Sun and transferred into chemical energy by the photosynthetic organisms of that ecosystem.	9.00	0.00	8-10
	H.B.4 The student will demonstrate an understanding of the specific mechanisms by which characteristics or traits are transferred from one generation to the next via genes.	11.00	0.00	8-12
	B.5 The student will demonstrate an understanding of biological evolution and the diversity of life.	8.00	0.00	8-12
	H.B.6 The student will demonstrate an understanding that ecosystems are complex, interactive systems that include both biological communities and physical components of the environment.	9.00	0.00	8-10
Spring	H.B.1 The student will use the science and engineering practices, including the processes and skills of scientific inquiry, to develop understandings of science content	8.00	0.00	8-10
	H.B.2 The student will demonstrate the understanding that the essential functions of life take place within cells or systems of cells.	16.20	0.45	12-18
	H.B.3 The student will demonstrate the understanding that all essential processes within organisms require energy which in most ecosystems is ultimately derived from the Sun and transferred into chemical energy by the photosynthetic organisms of that ecosystem.	7.00	0.00	8-10
	H.B.4 The student will demonstrate an understanding of the specific mechanisms by which characteristics or traits are transferred from one generation to the next via genes.	9.80	0.45	8-12
	B.5 The student will demonstrate an understanding of biological evolution and the diversity of life.	9.80	0.45	8-12
	H.B.6 The student will demonstrate an understanding that ecosystems are complex, interactive systems that include both biological communities and physical components of the environment.	9.00	0.00	8-10

Evaluation of Test Blueprint

Panelists discussed the extent to which the test blueprint adequately covered what students should know and be able to do. The panelists felt the number of items reflected the number of indicators within each standard, resulting in a balanced test blueprint.

DOK Consistency

As can be seen in Table 2.22, the item DOKs of over 70% of the items for the fall/winter (72.10%) and spring (70.60%) forms were lower than the DOKs of the standards to which they were linked.

Table 2.22. DOK Consistency Results for Biology 1, by Form

Form	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
	Mean	SD	Mean	SD	Mean	SD	Mean
Fall/Winter	72.10	4.90	24.80	4.40	3.00	2.50	27.80
Spring	70.60	2.60	26.70	2.50	2.70	1.50	29.40

Table 2.23 provides a summary of DOK consistency, for items by standard, for each Biology 1 form. The items aligned to standard HB.5 (The student will demonstrate an understanding of biological evolution and the diversity of life) were at the same DOK level or higher level as the standard while most items aligned to the other standards were rated below the DOK levels of the standards.

Table 2.23. DOK Consistency Results for Biology 1, by Standard within Form

Form	Standard	% Below Standard Level		% At Standard Level		% Above Standard Level		% At and Above Standard Level
		Mean	SD	Mean	SD	Mean	SD	
Fall/Winter	HB1	100.00	0.00	0.00	0.00	0.00	0.00	0.00
	HB2	67.00	15.60	33.00	15.60	0.00	0.00	33.00
	HB3	100.00	0.00	0.00	0.00	0.00	0.00	0.00
	HB4	83.60	10.00	16.40	10.00	0.00	0.00	16.40
	B5	0.00	0.00	77.50	18.50	22.50	18.50	100.00
	HB6	77.80	0.00	22.20	0.00	0.00	0.00	22.20
Spring	HB1	75.00	0.00	25.00	0.00	0.00	0.00	25.00
	HB2	91.40	3.20	8.60	3.20	0.00	0.00	8.60
	HB3	100.00	0.00	0.00	0.00	0.00	0.00	0.00
	HB4	67.60	12.60	32.40	12.60	0.00	0.00	32.40
	B5	0.00	0.00	83.80	8.80	16.20	8.80	100.00
	HB6	86.70	5.00	13.30	5.00	0.00	0.00	13.30

Evaluation of Item Quality

When averaged across panelists, over 97% of the items on both forms were rated as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias (see Table 2.24).

Table 2.24. Percentage of Biology 1 Items with Positive Ratings on Each Item Quality Indicator, by Form

Form	Clarity	Accuracy	Grade Appropriate	Supports Research-based Instruction	Free of Bias
Fall/ Winter	99.17	100.00	98.33	97.50	100.00
Spring	98.33	100.00	97.66	97.99	100.00

Overall Holistic Evaluation

Four of the five panelists rated the overall alignment as good while one indicated the overall alignment needs major improvement. It should be noted the comments provided from this panelist were focused primarily on the low DOK level of items and indicated adequate representation of the content with only “some exceptions.”

Four of the five panelists also indicated the item DOK levels were lower than they expected for the linked standards, stating there were a high number of DOK level 1 items.

Other panelists’ comments suggested that improvements could be made by reducing the number of fact-based questions, increasing the number of questions on evolution, and ensuring questions and answer choices do not provide cues to the correct answer.

Task 2: Discussion

SC READY ELA

Overall, results from the alignment workshop indicate there is good alignment between the items on the SC READY ELA assessments and the South Carolina College-and–Career Ready Standards (SCCCRS) for ELA. In addition, the numbers of items linked to each domain and reporting category were within the targets specified on the test blueprint, thereby indicating that the items on the test address the intended categories specified on the blueprint. Based on the standards, panelists believed the blueprint covered what students should know and be able to do. However, most ELA panelists (in both the elementary and the middle school grades) felt the Inquiry domain could be more effectively assessed via other formats (e.g., performance-based testing).

Aside from grades 4 and 6, the panelists felt the DOK levels of the standards tended to be higher overall than the DOK levels of the items linked to those standards. This was especially true for the Inquiry domain. The SCDE should consider including target DOK levels in its test blueprints to improve consistency between the DOK levels of the standards and those of the items linked to those standards.

Finally, the panelists provided an external check on several aspects of item quality. The panelists rated the vast majority of items as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias. This confirms that the previously completed Content and Bias/Sensitivity Reviews were effective in ensuring item quality.

SC READY Math

The overall results from the alignment workshop indicate there is good alignment between items on the SC READY math assessments and the standards they were designed to assess. Furthermore, the number of items linked to each standard is within the target number of items specified on the test blueprint, indicating the items on the test address the intended categories specified on the blueprint.

Overall, the panelists felt that the grade 4 SC READY math blueprint adequately covered what students should know and be able to do. However, panelists felt the coverage of the standards specified on the blueprints for the other SC READY math grades could be improved.

For grade 3, panelists suggested that the weights for the Number Sense and Base Ten and Number Sense and Operations – Fractions categories should be increased because they represent the foundation of future math understanding. They also felt there was not enough variety among the graphing data items and there were too many items that assessed interpreting bar graphs.

For grade 5, the panelists felt there was an over-emphasis of standards 5.G.2 (Geometry, about coordinates) and 5.G.1 (Geometry; “define a coordinate system”), and the items that addressed those standards required low-level thinking. They suggested increasing the allocation of points for Number Sense and Base Ten, Number Sense and Operations – Fractions, and Algebraic Thinking and Operations to 11–14 points to reflect the number of standards and collective complexity of the standards in those categories. They also suggested reducing the number of points allocated to Geometry and Measurement and Data Analysis to 10–12 points to reflect the lower number of standards in those categories.

For grade 6, the panelists indicated the weight for Ratios and Proportional Relationships should be increased because they believed that standard is more important than the Geometry and Measurement standard. They also felt the weight for Data Analysis and Statistics should be decreased. In sum, they felt Number System, Expressions, Equations, and Inequalities, and Ratios and Proportional Relationships should each be weighted 25% and Geometry and Measurement and Data Analysis and Statistics should each be weighted 12.5%. They made this same recommendation for grade 7.

Finally, for grade 8, the panelists felt that the blueprint better reflected what should be tested than did the blueprints for grades 6 and 7, but they still felt that improvements could be made. Specifically, they suggested weights for Number System and Data Analysis, Statistics, and Probability should be decreased while weights for Functions, Geometry and Measurement, Expressions, Equations, and Inequalities, and Functions should be increased. Given the panelists’ concerns about the representation of the standards on the SC READY Math test blueprints, we recommend the SCDE convene another group of South Carolina content experts to review the test blueprints to ensure they appropriately represent the math SCCCRS.

In contrast to SC READY ELA, the panelists’ ratings for SC READY math indicated that, overall, there was reasonable consistency between the DOK levels of the standards and the DOK levels of the items linked to those standards.

Finally, the panelists rated the vast majority of the SC READY math items as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias. This confirms that the previously completed Content and Bias/Sensitivity Reviews were effective in ensuring item quality.

English 1

Results from the alignment workshop indicated there is good alignment between items on the English 1 EOCEP assessment and the South Carolina standards, although the alignment ratings tended to be stronger for the spring than for the fall/winter form.

The number of items linked to each standard (strand) was within the target number of items specified on the test blueprint, with one exception. The number of Writing items was slightly below the target number specified on the test blueprint for both the fall/winter and spring forms. This suggests the SCDE should consider adding one or two more Writing items to the English 1 EOCEP. Based on the standards, panelists indicated the English 1 EOCEP test blueprint appropriately reflected what students should know and be able to do.

The fall/winter and spring forms were close to meeting the DOK consistency criterion (i.e., at least 50% of the items linked to the standard at or above the DOK level for that standard); slightly fewer than 50% of the items on the fall/winter form received DOK ratings that were at or above the DOK ratings of the standards. Interestingly, all Writing items on the fall/winter form were at or above the DOK level of the Writing standard; however, this pattern was reversed for the spring form such that most Writing items on the spring form were rated below the DOK level of the Writing standard. Given the differences found between the fall/winter and spring forms, the SCDE should consider having South Carolina English 1 content experts review the fall/winter and spring forms for consistency.

Finally, the English 1 panelists provided an external check on several aspects of item quality. The panelists rated the vast majority of items as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias, confirming previous reviews were effective in ensuring item quality.

Biology 1

Results from the alignment workshop indicated there is good alignment between the items on the Biology 1 EOCEP assessments and the South Carolina standards. The number of items linked to each standard was within the target number of items specified on the test blueprint for all standards, with one exception—standard HB.3 (The student will demonstrate the understanding that all essential processes within organisms require energy which in most ecosystems is ultimately derived from the Sun and transferred into chemical energy by the photosynthetic organisms of that ecosystem) on the spring form was one item short of meeting the target number of items. In addition, based on the standards, panelists felt the test blueprint adequately represented what students need to know and be able to do.

Panelists found the item DOK levels of over 70% of the Biology 1 items on the fall/winter and spring forms were at lower DOK levels than those of the standards to which they were linked. The items aligned to standard B.5 (The student will demonstrate an understanding of biological evolution and the diversity of life.) were at the same DOK level or higher as the standard while most items aligned to the other standards were rated below the DOK levels of the standards. The SCDE should consider including target DOK levels in its test blueprints to improve consistency between the DOK levels of the standards and items linked to those standards.

The panelists rated the vast majority of items as clear, accurate, grade appropriate, supporting research-based instruction, and free of bias, confirming that the previously completed Content and Bias/Sensitivity Reviews were effective in ensuring item quality.

Chapter 3: Review Test Construction Processes (Task 3)

Matt Swain, Amanda Koch, & Adam Beatty

Task 3: Introduction

Forms construction refers to the assembly of test items into forms that meet certain specifications for content, statistical properties, and construct representation. We evaluated the test form construction processes based on eight best practices described in the *Test Standards*.

The current chapter generally follows the same organization as the chapter in the first report (Dickinson, Chen, & Swain, 2017) where we reviewed the test construction processes for the SC READY assessments and the Algebra 1 EOCEP assessment. In this chapter, we evaluate the test construction processes for the English 1 and Biology 1 EOCEPs. In addition, based on receipt of new documents since delivering the first report, we also provide final ratings of fidelity to the forms construction standards for Algebra 1 and SC READY. Finally, we update the SC READY ELA and math ratings to include findings from a site visit we conducted to observe forms assembly.

Task 3: Method

Documents and Datasets Reviewed

We worked in cooperation with the South Carolina Education Oversight Committee (EOC), the South Carolina Department of Education (SCDE), and the Data Recognition Corporation (DRC), with primary support provided by DRC, to obtain documentation related to South Carolina's test construction processes. We also searched the SCDE website to identify additional relevant information. We also received additional documents from DRC that were relevant to Algebra 1 and SC READY, which we used to update our earlier evaluation of Algebra 1 and SC READY forms construction (Dickinson et al., 2017). Table 3.1 summarizes the forms construction documents and datasets we reviewed.

Table 3.1. Forms Construction Documents and Datasets Reviewed

Report Section	Document Filename
English 1	
Fidelity to Forms Construction Standards	030_Forms Construction Guidelines_E.pdf [a]
	032_Guidelines for Ordering Items_English1_E.pdf [a]
	033_Guidelines for making changes within a test form_E.pdf [a]
	034_Test Form Construction Process_E.pdf [a]
	037_Guidelines for Forms Creation_E.pdf [a]
	038_Quality Assurance Procedures for Test Construction_E.pdf [a]
	2016-17 EOCEP Technical Report for HumRRO.pdf [a]
	5.2 EOCEP Examination Relationships with Other Benchmark Tests.pdf [a]
	3.11 EOCEP Bio Eng Principal Component Analysis.pdf [a]
	2016 EOCEP ENG1_ALG1 Standard Setting Report_091316.pdf [ac]
	2016 EOCEP ENG1_ALG1 Standard Setting Report_SCDE Addendum.pdf [ac]
Item Bank Metadata	036_ENG_Item metadata_Eligible items_E.xlsx Metadata_2016_2017.xlsx

Table 3.1. (Continued)

Report Section	Document Filename
Biology 1	
Fidelity to Forms Construction Standards	EOCEP Forms Construction Guidelines_101817.pdf [a]
	DRC Item Development Tech Manual_101817.pdf [a]
	030_Forms Construction Guidelines_E.pdf [a]
	031_Guidelines for Ordering Items_Biology 1_E.pdf [a]
	033_Guidelines for making changes within a test form_E.pdf [a]
	034_Test Form Construction Process_E.pdf [a]
	037_Guidelines for Forms Creation_E.pdf [a]
	038_Quality Assurance Procedures for Test Construction_E.pdf [a]
	2016-17 EOCEP Technical Report for HumRRO.pdf [a]
	3.11 EOCEP Bio Eng Principal Component Analysis.pdf [a]
	5.2 EOCEP Examination Relationships with Other Benchmark Tests.pdf [a]
	2017 EOCEP BIO1 Standard Setting Report_091817.pdf [ac]
	2017 EOCEP BIO1 Standard Setting Report_SCDE Addendum.pdf [ac]
Item Bank Metadata	035_BIO_Item metadata_Eligible items_E.xlsx [a]
Algebra 1	
Fidelity to Forms Construction Standards	030_Forms Construction Guidelines_E.pdf [a]
	033_Guidelines for making changes within a test form_E.pdf [a]
	034_Test Form Construction Process_E.pdf [a]
	037_Guidelines for Forms Creation_E.pdf [a]
	038_Quality Assurance Procedures for Test Construction_E.pdf [a]
	2016-17 EOCEP Technical Report for HumRRO.pdf [a]
	5.2 EOCEP Examination Relationships with Other Benchmark Tests.pdf [a]
	2016 EOCEP ENG1_ALG1 Standard Setting Report_091316.pdf[ac]
	2016 EOCEP ENG1_ALG1 Standard Setting Report_SCDE Addendum.pdf [ac]
SC READY	
Forms Construction Site Visit	016_Guidelines for Item Analysis and Form Construction_R.pdf [b]
Fidelity to Forms Construction Standards	SC READY Forms Construction Guidelines_101817.pdf [a]
	DRC Item Development Tech Manual_101817.pdf [a]
	027_SC READY and SCPASS Spring 2016 Test Mode Comparability Study.pdf [a]
	SC READY 2017 Technical Report_100917.pdf [a]
	5.2 SC READY Multi-State Common Calibrations.docx [a]
	2016 SCREADY Standard Setting Report.pdf [a]
	2016 SC READY Standard Setting Report_SCDE Addendum.docx [a]
	2016 SCREADY Vertical Moderation Report.pdf [ac]

[a] Document received between delivery of the first report and the current report.
[b] Document received prior to the first report.
[c] Document reviewed but not cited in this report.

Procedures for Reviewing Documents and Datasets[16]

Two HumRRO staff independently rated each relevant *Test Standard* after reviewing the documents related to each assessment. These staff then met and participated in a discussion until they reached a consensus rating. Table 3.2 describes the rating scale that staff applied. The goal was to quantify the fidelity of the practices as described in the forms construction documents to the *Test Standards*. In addition to the numeric rating, we provided comments regarding specific aspects of the *Test Standard* that were missing from the documentation. The first part of the Results section is organized by *Test Standard* and includes the text of the standard, our assigned rating, and an explanation of what was not found in the documentation provided by the testing contractor.

Table 3.2. Rating Scale for Evaluating Strength of Evidence for Test Standards

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

[a] Materials include all documents and data provided, emails or phone calls with SCDE/DRC staff, as well as information available online.

Procedures for Reviewing Item Bank Metadata

We used procedures to review the English 1 and Biology 1 metadata similar to those used to review the Algebra 1 and SC READY metadata for the first report (Dickinson et al., 2017). First, we imported the Excel spreadsheets provided by DRC into SAS 9.4. Then we focused on reviewing the descriptive statistics of the item bank to determine how well they related to the target form statistics. This review allowed us to determine how readily the item bank would allow for building test forms.

Procedures for Reviewing SC READY Forms Construction Meeting

During the week of August 14, 2017, DRC and South Carolina Department of Education (SCDE) staff assembled SC READY forms for spring 2018 operational testing of grades 3–8 in ELA and math. Two HumRRO staff observed the first three days of forms assembly (i.e., August 14–16). One observer focused on the test forms constructed for ELA while the second observer focused on the test forms constructed for math.

[16] The process for reviewing materials for adherence to relevant *Test Standards* is the same process as used in Task 1 (Review of Item Development Processes), Task 4 (Review of Test Administration Processes), and Task 5 (Review of Scaling, Equating and Scoring Processes).

To aid in their review of the SC READY forms construction, HumRRO staff developed checklists based on the procedures in the *016_Guidelines for Item Analysis and Form Construction_R.pdf* document. The observers used the checklist to guide their observations and reviews (see ELA and Math columns in Appendix G).

Table 3.3 contains the rubric for the site visit checklist ratings. The rubric for the checklists enabled the observers to provide a quantitative rating of fidelity between the documented procedure and the observed procedure. The rating scale ranged from 1–5, with higher ratings indicating greater fidelity between the documented and observed step. After observing forms being constructed, both HumRRO staff met and determined consensus ratings that merged their observations from both groups. These consensus ratings are found in the Consensus column of Appendix G.

Table 3.3. Rating Codes for Site Visit to Forms Construction Meeting

Rating Level	Description
1	Documented procedure was not followed; actual procedure did not resemble documented procedure.
2	Documented procedure was rarely followed, or was followed incompletely or mostly incorrectly.
3	Documented procedure was followed some of the time, but not all the time. Aspects/steps of the procedure may have been missing or may not have been documented.
4	Documented procedure was mostly followed most of the time. Extraneous aspects/steps were rarely included.
5	Documented procedure was followed; there were no additional aspects/steps taken than what was planned.

Task 3: Results

English I

Fidelity to Forms Construction Test Standards

Table 3.4 presents HumRRO's evaluation ratings for adherence of English 1 forms construction to the relevant *Test Standards*. Two HumRRO staff independently reviewed the materials and assigned ratings. They then met and discussed their ratings until they reached consensus. The rationale for each rating follows this table.

Table 3.4. English 1 Evaluation Results Based on the Test Standards

Test 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.	4
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

Rationale for English 1 Test Standards Evaluation Ratings

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).

There was no test specifications document; however, there was a test blueprint and the purpose of the EOCEP program (including the English 1 assessment) was clearly described on the SCDE Website.[17] Both the Website and the *030_Forms Construction Guidelines_E.pdf* document mentioned the assessment contributes 20% towards a student's final grade for English 1, which is a required course. The definition of the construct was described by the test blueprint located in Appendix B of the *030_Forms Construction Guidelines_E.pdf* document. Additionally, the *2016–17 EOCEP Technical Report for HumRRO.pdf* document showed that English 1 test forms for fall/winter, spring, and summer met those blueprints. It was clear the intended examinee population includes students finishing their English 1 coursework as the intended use is to contribute to the students' final grades. These confluent descriptions form a basis for test specifications but they are not compiled in a unified document or Website. Such a document is recommended. This document could contain the rationale supporting test uses as well.

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.

As noted, there was no single test specifications document; however, the information required by this standard was found across several documents and Websites. The content, length, item formats comprising the exam, and psychometric properties were described in *030_Forms Construction Guidelines_E.pdf*. Most remaining requirements of the standard were found online. The procedures for test administration were found online and were very detailed.[18] The *030_Forms Construction Guidelines_E.pdf* document indicated a Rasch measurement model is used to calibrate items; however, there was no information about how examinees are scored or how a range of eventual scale scores are derived. The *2016–17 EOCEP Technical Report for HumRRO.pdf* mentioned examinees are not timed but they are allowed a full day to complete the assessment unless their IEP/504 Plan indicates otherwise.

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 specifications. Our understanding is that all PBT items are ported from the CBT with a few substitutions. That is, some item types can only be administered on computer (i.e., technology enhanced [TE] items) and these are swapped for items in the same content standard on the PBT

[17] https://ed.sc.gov/tests/high/eocep/

[18] https://ed.sc.gov/tests/tests-files/eocep-files/eocep-spring-2017-test-administration-manual-tam/

version of an exam. As mentioned in a phone interview on March 1, 2017 with SCDE and DRC staff, item-level mode differential item functioning (DIF) is explored using the Educational Testing Service's (ETS's) Delta method. Items with category "C" DIF are sent to item developers for review, although SCDE staff indicated items rarely reach that level of DIF for mode comparisons. The vast majority of South Carolina students (98%) complete the exam online. The 2% who complete the PBT version could be matched with 2% of online students (i.e., propensity score matching) to conduct mode comparability analyses to verify equivalent forms and comparable scores.

***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**.*

This standard addresses variations in testing accommodations for students with disabilities. As it applies to forms construction, paper forms are clearly necessary for students whose IEP/504 Plans require them. There are some tools available on the online testing platform for students with disabilities (i.e., visual impairments); however, these features were not described in the documents provided. There also are some descriptions of accommodations online for EOCEP assessments.[19] Because we evaluated this standard from a forms construction perspective, we found the documentation sufficiently addressed the need and procedures to adapt online forms to paper 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 *DRC Item Development Tech Manual_101817.pdf* is an exhaustive description of the life cycle of an item. We applaud DRC for this detailed description that goes above the standard. The procedure for selecting field test (FT) items is well-documented in terms of number of items, their placement, and statistics. We agree with the practice of SCDE reviewing a composed form (operational and field test items), as a review of the pool of operational items would not provide a complete picture from an examinee's perspective.

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 items are field tested using a sample of students who come from the same population who are administered 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. According to an email received from DRC on October 24, 2017, the EOCEP forms are assembled using item parameters that are based on only South Carolina students, which complies with this standard. There are no concerns with using other state's FT data to place FT items on South Carolina forms. However, these FT parameters should be updated with only South Carolina student data before use as an operational item in a pre-equated form.

[19] http://ed.sc.gov/tests/assessment-information/testing-swd/accommodations-and-customized-forms/

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 documentation clearly refers to use of a Rasch model to calibrate new item parameters and equate them to a common scale. These parameters are used to generate form-level difficulty estimates and make comparisons across forms. Our review revealed a disconnect between the use of a Rasch model to calibrate and equate items and the use of classical test theory (CTT) parameters to assemble forms. We are unclear as to how forms can be pre-equated when CTT parameters are used to assemble forms rather than the equated Rasch difficulties. The *038_Quality Assurance Procedures for Test Construction_E.pdf* document states:

> [t]he use of standardized-test construction software enables the construction of forms with similar test characteristic functions and standard errors of measurement curves, and DRC's calibration and equating designs ensure that scaled scores are comparable across different forms of each test. (page 1)

However, it does not appear this information applies to the EOCEP assessments as CTT parameters are used to create forms, not item response theory (IRT)-based aspects like test characteristic curves. This should be clarified in the documentation.

We could not find information about how the sample is used for estimating item properties nor procedures to ensure the sample is of adequate size and diversity for the procedure. The new *EOCEP Forms Construction Guidelines_101817.pdf* document indicates pre-equating will be based on the first field test administration of the item. According to an email received on October 24, 2017 from DRC, these item parameters are computed using South Carolina student data, which complies with this standard. We are aware a post-equating check is performed and agree that a check for item drift aligns with best practice.

The primary documentation mentions CTT item parameter targets—mean *p*-value and median point-biserial range—are provided to guide form-level evaluation. This documentation mentions items are screened for DIF using the ETS Delta method; items with DIF flags of "C" are not considered but those with "B" may be considered. The documents do not specify 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 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 that investigated if test scores of these two modes are comparable for English 1, 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. As mentioned in Standard 4.4, we recognize few examinees complete the paper-and-pencil version of the EOCEP assessments so there is likely little concern for a mode effect to sway the item parameters. However, a small study using propensity score matching could be conducted to elucidate if mode differences exist. 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 construct-irrelevant variance.

English 1 Item Bank Metadata

DRC provided an eligible English 1 item bank that contained 365 items, and included content codes and item statistics. We did not include Tier 2 items in the eligible pool as these items were designated for use as a last choice. All items were included in these analyses, including those with a status of “OPReady” and “FTReady.” Table 3.5 presents classical item statistics for the eligible item bank (*k* = 365).

Table 3.5. English 1 Item Bank Descriptive Statistics

	k	Min	Max	Median	Mean	*SD*
***p*-values**	365	0.31	0.86	0.58	0.58	0.13
Point-Biserial Correlations	365	0.20	0.77	0.45	0.45	0.12

Although the mean *p*-value is 0.58, the target mean for form assembly is 0.65, which is slightly easier than what the bank provides overall. The guidelines state *p*-values should range between 0.30 and 0.85. With 99.7% of items in the bank falling within that range, it is highly unlikely the target range will be violated. Figure 3.1 depicts the distribution of *p*-values for the eligible bank. The target mean *p*-value of 0.65 also shows how many items are less than the target mean difficulty.

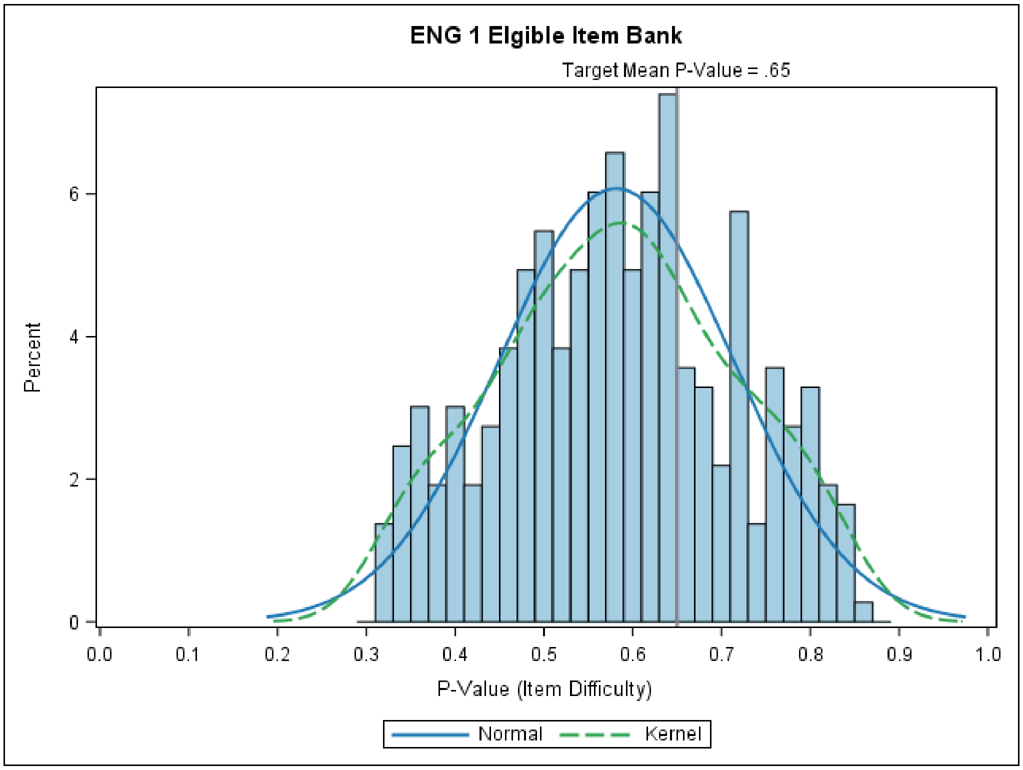


Figure 3.1. Distribution of p-values from the eligible English 1 item bank.

Figure 3.2 depicts the distribution of point-biserial correlations in the eligible English 1 item bank. The median and mean point-biserial are both at 0.45, which is the upper end of the target median range of 0.35 to 0.45. This is a desirable characteristic of the item bank because items with higher item discrimination (point-biserial correlations) have a stronger relationship with the construct being assessed. That is, items with high point-biserial correlations do well at delineating between low- and high-performing examinees. Although a range of item discrimination parameters is desired, higher point biserial correlations are better than lower ones. Finally, it is noteworthy that there are no items with point-biserial correlations below .20, the lower limit according to the guidelines.

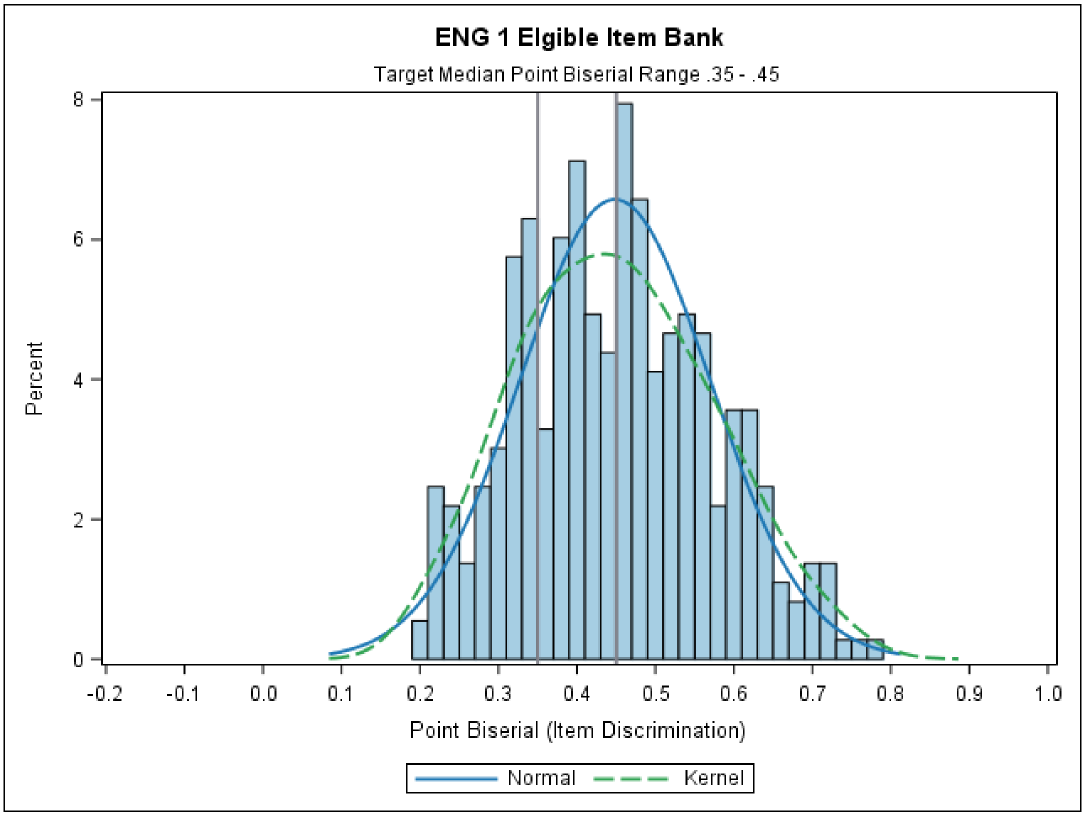


Figure 3.2. Distribution of point-biserial correlations from the English 1 eligible item bank.

Biology 1

The findings for Biology 1 mirror those presented for English 1, largely because the documentation provided by DRC is the same for both assessments.

Fidelity to Forms Construction Standards

Table 3.6 presents the final, consensus rating assigned to each *Test Standard* under review relative to the documentation available for the Biology 1 assessment. Here, we explain each rating, highlighting areas where the documentation exceeded or perhaps did not meet requirements outlined for that *Test Standard*.

Table 3.6. Biology 1 Evaluation Results Based on the Test Standards

Test 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.	4
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

Rationale for Biology 1 Test Standards Evaluation Ratings

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).

There was no "test specifications" document; however, there was a test blueprint and the purpose of the EOCEP program (including the Biology 1 assessment) was clearly described on the SCDE Website.[20] Both the Website and the *030_Forms Construction Guidelines_E.pdf* document indicated that the exam is weighted 20% in a student's final grade for Biology 1, a required course. The definition of the construct was described by the test blueprint located in Appendix B of the *030_Forms Construction Guidelines_E.pdf* document. And the *2016-17 EOCEP Technical Report for HumRRO.pdf* document shows that Biology 1 forms for fall/winter, spring, and summer met those blueprints. It was clear the intended examinee population includes students finishing their Biology 1 coursework as the intended use is to contribute to the students' final grades. These confluent descriptions form a basis for test specifications but they are not in a unified document or Website. Such a document is recommended. This document could contain the rationale supporting test uses as well.

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.

As noted, there was no single "test specifications" document; however, the information required by this standard was found across several documents and websites. The content, length, item formats comprising the exam, and psychometric properties were described in *030_Forms Construction Guidelines_E.pdf*. Most of the remaining requirements of the standard were found online. The procedures for test administration were found online and were very detailed.[21] The *030_Forms Construction Guidelines_E.pdf* document mentioned that a Rasch measurement model is used to calibrate items; however, there was no information about how examinees are scored or how a range of eventual scale scores are derived. The *2016-17 EOCEP Technical Report for HumRRO.pdf* indicates that examinees are not timed but they are allowed a full day to complete the exam unless their IEP/504 Plan indicates otherwise.

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 specifications. Our understanding is that all PBT items are ported from the CBT with a few substitutions. That is, some item types can only be administered on computer (i.e.,

[20] https://ed.sc.gov/tests/high/eocep/
[21] https://ed.sc.gov/tests/tests-files/eocep-files/eocep-spring-2017-test-administration-manual-tam/

technology enhanced [TE] items) and these are swapped for items in the same content standard on the PBT version of an exam. As mentioned in a phone interview on March 1, 2017 with SCDE and DRC staff, 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. The vast majority of South Carolina students (98%) complete the exam online. The 2% who do complete the PBT version could be matched with 2% of online students (i.e., propensity score matching) to conduct mode comparability analyses to verify equivalent forms and comparable scores.

***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**.*

This standard addresses variations in testing accommodations for students with disabilities. As it applies to forms construction, paper forms are necessary for some students as required by their IEP/504 Plans. 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 for online EOCEP assessments.[22] However, we are evaluating this standard from a forms construction perspective and find that the documentation addresses sufficiently the need and procedures to adapt online forms to paper.

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 *DRC Item Development Tech Manual_101817.pdf* is an exhaustive description of the life cycle of an item. We applaud DRC for this detailed description that goes above the standard. The procedure for selecting field test (FT) items is well-documented in terms of number of items, their placement, and statistics. We agree with the practice of SCDE reviewing a composed form (operational and field test items). A review of just operational items would not be a complete picture from an examinee's perspective.

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 who are administered 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. According to an email received from DRC on October 24, 2017, the EOCEP forms are assembled using item parameters that are based on only South Carolina students, which complies this standard. There are no concerns with using other state's FT data to place FT

[22] http://ed.sc.gov/tests/assessment-information/testing-swd/accommodations-and-customized-forms/

items on South Carolina test forms. However, these FT parameters should be updated with only South Carolina student data before use as an operational item in a pre-equated form.

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 documentation clearly refers to use of a Rasch model to calibrate new item parameters and equate them to a common scale. These parameters are used to generate form-level difficulty estimates and make comparisons across forms. Our review revealed a disconnect between the use of a Rasch model to calibrate and equate items and the use of classical test theory (CTT) parameters to assemble forms. We are unclear as to how forms can be pre-equated when CTT parameters are used to assemble forms rather than the equated Rasch difficulties. The *038_Quality Assurance Procedures for Test Construction_E.pdf* document states:

> "[t]he use of standardized-test construction software enables the construction of forms with similar test characteristic functions and standard errors of measurement curves, and DRC's calibration and equating designs ensure that scaled scores are comparable across different forms of each test." (page 1)

However, it does not appear that this information applies to the EOCEP assessments as CTT parameters are used to create forms, not item-response theory (IRT) based aspects like test characteristic curves. This should be clarified in the documentation.

We could not find documentation to address this part of the standard: "the sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure." The new *EOCEP Forms Construction Guidelines_101817.pdf* document indicates that "pre-equating will be based on the first field test administration of the item." According to an email received on October 24, 2017 from DRC, these item parameters are computed using South Carolina student data, which complies with this standard. We are aware that a post-equating check is performed and agree that a check for item drift aligns with best practice.

The primary documentation mentions CTT item parameter targets. That is, a mean *p*-value and median point-biserial range is provided to guide form-level evaluation. It is mentioned that items are screened for differential item functioning (DIF) 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 for Biology 1, or if item parameters are similar. The new *EOCEP Forms Construction Guidelines_101817.pdf* document states, "separate conversion tables are used for the online and print forms." We interpret that as two scoring tables are produced for paper and online test separately. However, if this is not true and if data from paper- and computer-based forms are combined to estimate (calibrate) item parameters used to assemble forms, then there could be a situation where the "true" item parameters (that is, with mode effects removed) do not meet the psychometric guidelines. As mentioned in Standard 4.4, we acknowledge that few examinees complete the paper-and-pencil version of the EOCEP assessments and there is likely little concern for a mode effect to sway the item parameters. However, a small study using propensity score match could elucidate if mode differences existed and satisfy Standards 4.4 and 4.13. 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 construct-irrelevant variance.

Biology 1 Item Bank Metadata

DRC provided an eligible Biology 1 item bank that contained 330 items, and included content codes and item statistics. We did not include Tier 2 items in the eligible pool as these items were designated for use as a last choice. All items were included in these analyses, including those with standard set designations of "BIO04" and "BIO15" status of "OPReady" and "FTReady." Table 3.7 presents classical item statistics for the eligible item bank (k = 330).

Table 3.7 Biology 1 Item Bank Descriptive Statistics

	k	Min	Max	Median	Mean	SD
***p*-values**	330	0.30	0.89	0.49	0.51	0.12
Point-Biserial Correlations	330	-0.20	0.60	0.36	0.36	0.08

Although the mean *p*-value is 0.51, the target for form assembly is 0.65, which is moderately easier than what the bank provides overall. The guidelines state *p*-values should range between 0.30 and 0.85. With 97.8% of items in the bank falling within that range, it is unlikely the target range will be violated.[23] Figure 3.3 depicts the distribution of *p*-values for the eligible bank. The target mean *p*-value of 0.65 also shows how many items are less than the target mean difficulty.

[23] Task 6 (see chapter 6) provides *p*-values for the operational fall/winter and spring forms. Table 5.14 in chapter 6 shows that no items on the operational forms were below the lower end of the target range (i.e., 0.30). Moreover, the mean *p*-value of the operational forms was 0.591, which is somewhat higher than the mean *p*-value of the item bank, although still slightly lower than the target mean *p*-value of 0.65.

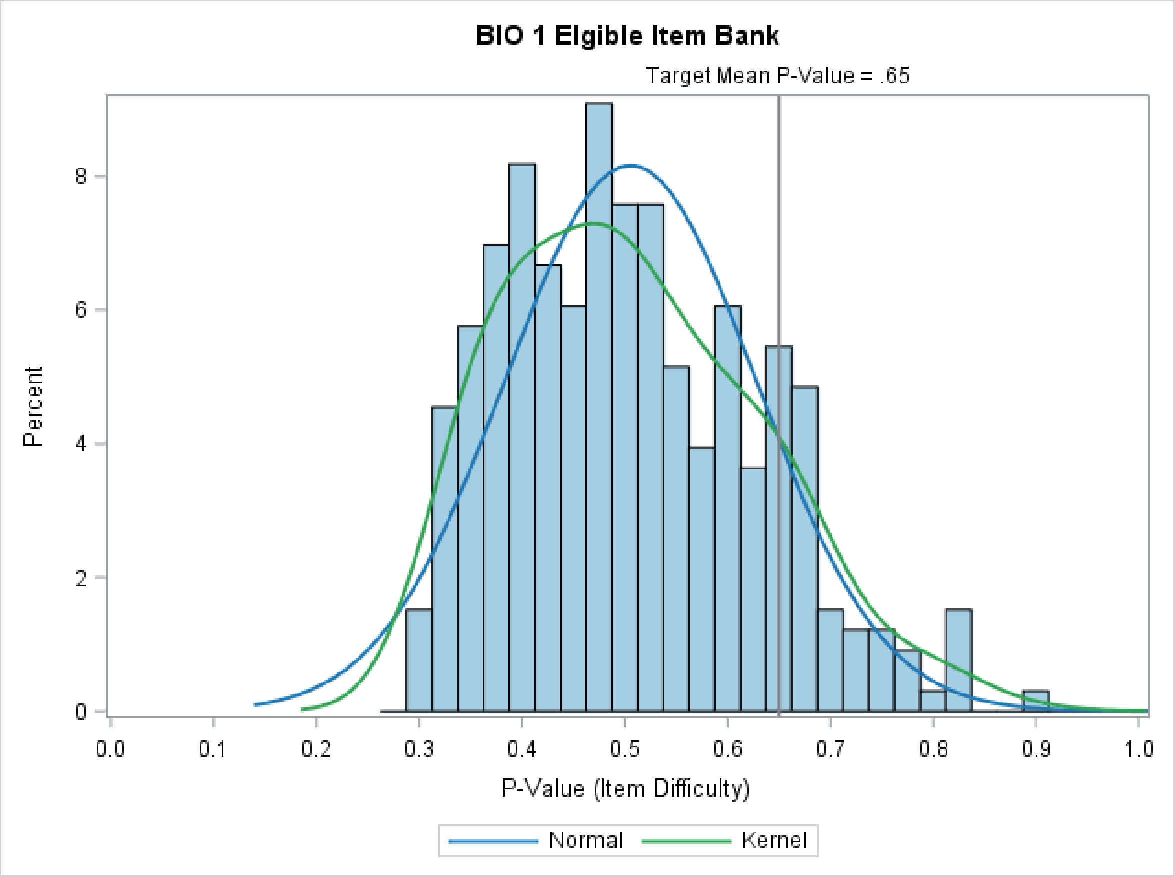


Figure 3.3. Distribution of p-values from the Biology 1 eligible item bank.

Point-biserial correlations for items in the eligible bank are a bit closer to the form targets. The median and mean point-biserial are both at 0.36, which is within the target median range of 0.35 to 0.45. Figure 3.4 depicts the number of items that are close to this range. Also notable is that no items have point-biserial correlations below 0.20, the lower limit according to the guidelines.

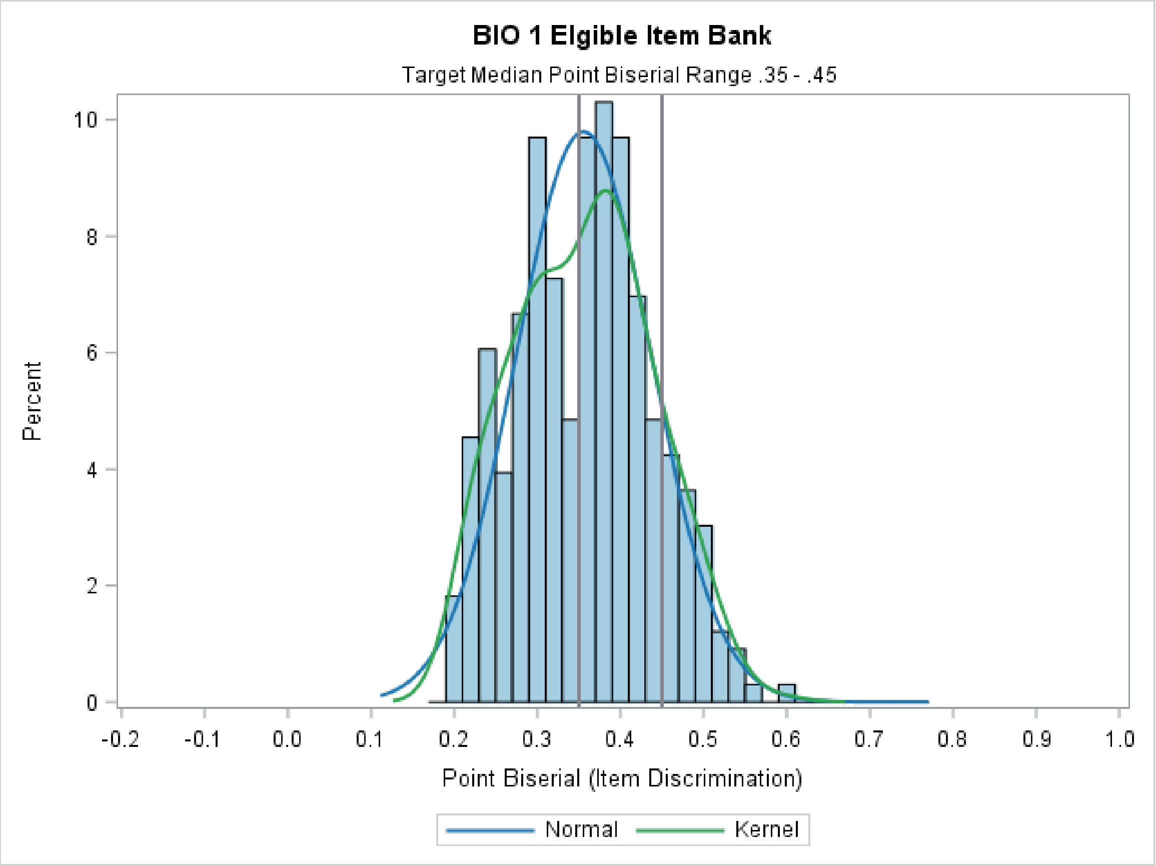


Figure 3.4. Distribution of point-biserial correlations from the Biology 1 eligible item bank.

Algebra 1

Fidelity to Forms Construction Standards

Based on recommendations offered in our first report (Dickinson et al., 2017), several new documents were provided, some of which were relevant to the Algebra 1 EOCEP exam (see Table 3.1). Table 3.8 contains updated ratings for Algebra 1 based on our review of the new documents. One rating increased from the first report (Standard 4.2) based on the newly provided information. Because documentation was identical, the rationale for the Algebra 1 ratings mirrors that of the Biology 1 and English 1 and is not repeated here.

Table 3.8. Final Algebra 1 Evaluation Results Based on the Test Standards

Test 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.	4[a]
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

[a]This rating was updated from the first report after new documentation was received.

SC READY

In the first report (Dickinson et al., 2017), SC READY forms construction documents were reviewed to gauge their fidelity with the same *Test Standards* identified for the EOCEP assessments. Since that first report was submitted, we were provided with more documentation (see Table 3.1 above) *and* we conducted an on-site observation of forms construction for SC READY. Thus, we incorporated this new information for SC READY into the current report.

On-site Observation of SC READY Forms Construction

The HumRRO observers' consensus ratings on adherence to the steps in the Forms Construction Checklist (Appendix G) were collapsed across steps and mean ratings computed. Based on a 5-point rating scale, a mean rating of 4.50 was obtained for ELA (8 steps observed) and a mean rating of 4.11 was obtained for math (9 steps observed) on the 5-point rating scale.[24] Supporting notes were provided from both observers (see Appendix G).

Based on observations of the SC READY assembly of test forms, we provide the following recommendations:

- Standard 4.9 states that when trying out items, "the sample(s) should be as representative as possible of the population(s) for which the test is intended." If items on the SC READY assessment include items from DRC's college- and career-readiness (CCR) item bank for which item statistics are based on students in other states (i.e., not South Carolina students), then this standard could be compromised.[25] If SC READY assessments include items for which the item statistics come from students in other states, then additional detail should be provided on that population of students to ensure that it is representative of the South Carolina population of students.
- The psychometrician appeared to use an Excel macro to compute form statistics. At one point, the formulas did not encompass all rows in the spreadsheet, and therefore form statistics did not represent all items on the form. However, this was discovered and corrected by the psychometrician. Given the high-stakes nature of the decisions based on form statistics, we recommend quality checks be conducted of the Excel macro to ensure the formulas are accurate. Additionally, the process could be modified to rely less on manual modification of Excel spreadsheets (e.g., copying and pasting of item information from different Excel spreadsheets) as input to the macro. For example, a column with identified item IDs could be prepared. Then, a macro could be created that merges the item IDs, selecting only those identified, and dynamically create a new spreadsheet that automatically accounts for the number of items.
- When participants reject items for inclusion on a form, the participants' reasons for rejection did not appear to be documented. We recommend including item rejection explanations within the item bank. This information would be useful for editors to correct information or allow staff to immediately exclude these items during future forms assembly.
- Approximately 25% of items are refreshed each year. However, there does not appear to be a mechanism to track how long an item has been on a form. We recommend the item bank include the year and the form(s) on which the item was last used. Given there are only two

[24] The HumRRO observers attended three of the five days of the Forms Construction Meeting; consequently, some steps were not observed.

[25] 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.

years of data in the existing item bank, this is not a pressing need, but the recommendation should be implemented soon.

- The SCDE may want to consider requesting that DRC create a statistical program that assembles forms to satisfy content and psychometric requirements simultaneously. These forms would then be reviewed by content specialists to identify concerns and be revised as needed. Enacting such a process would be more efficient by removing some of the manual steps involved in the current forms construction process, while still leveraging the expertise of the content experts in the areas in which they uniquely contribute.
- During the forms construction meeting, when the content specialists had difficulty finding items to satisfy certain content standards, they appeared to pull items from other states' item banks. However, it was necessary to align these items to the SCCCRS before they could be used on a form. We recommend this alignment work be completed in a more thoughtful manner rather than on-the-fly. Alignment work can take time and include deliberation with other content experts.
- Not all meeting participants were actively engaged in aspects of forms construction during the forms construction meeting. Some participants had considerable periods of time in which they waited for others to finish a step so they could begin their step. Specifically, the SCDE staff's time was not used consistently during the meeting. Consideration should be given to restructuring the way SCDE content experts participate in the forms construction meeting. One suggestion may be for DRC content specialists to develop drafts of the forms, DRC psychometricians review them, and DRC content specialists revise them, all prior to the in-person forms construction meeting (SCDE could virtually attend this portion of the meeting if desired, which would save travel expenses). The in-person meeting could then begin with SCDE content expert reviews of the forms that DRC created.

Overall, it is important to note that the overall mean ratings from the observation checklist were quite high, thereby indicating fidelity between the actual forms construction steps and the documented forms construction steps. Additionally, the HumRRO's observers noted the forms construction meeting was well organized.

Final Forms Construction Evaluation Results for SC READY

Since the first report was submitted (Dickinson et al., 2017), we were provided with more documentation (see Table 3.1 above) *and* we conducted the aforementioned site visit of forms construction. Thus, we incorporated the new information for SC READY into Table 3.9.

Table 3.9. Final SC READY Evaluation Results Based on the Test Standards

Test 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.	4[a]
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.	4[a]
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.	5[a]

[a]These ratings increased from the first report after new documentation and information became available.

Rationale for SC READY Test Standards Evaluation Ratings

The rationale for the ratings in Table 3.9 are presented next. Observations based on new information that was not available for our first report are incorporated into our original evaluation below.

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).

Although not stated in the documents provided, the purpose of the SC READY exams does appear in the *SC READY Test Administration Manual* (TAM).[26] A primary use of test scores is to meet the annual accountability requirements defined by South Carolina 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 SC READY 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 online TAM.

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 is inferred they are the same. *The SC READY Forms Construction Guidelines_101817.pdf* document states that companion items presented on the paper forms have item characteristics similar to their computer-based form counterparts. That assumes there are no item-level mode effects or differences in performances based on mode of delivery. Based on the study presented in *027_SC READY and SCPASS Spring 2016 Test Mode Comparability Study.pdf*, mode DIF is a rare occurrence particularly for the math assessments.

[26] http://ed.sc.gov/tests/tests-files/sc-ready-files/2016-sc-ready-test-administration-manual-tam/

The study did not indicate if mode DIF affected "validity of score interpretations;" for example, no correlations of scores with external variables (i.e., concurrent validity) were reported.

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 assessments 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 only paper-based testing 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 items on an ELA and math form are refreshed each year with field test items. However, during the form assembly site visit, HumRRO staff did not observe a mechanism to track how long an item has been on a form. We recommend the item bank be revised to indicate the year(s) and form(s) the item was last used. Given there are only two years of data in the item bank, this is not a pressing need, but should be implemented soon. Moreover, the *030_Forms Construction Guidelines_E.pdf* document states "items chosen for operational use should not have appeared on the most recent two administrations." We recommend that items be retired by age rather than random chance.

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.

We noticed during the on-site forms assembly meeting that some items may have come from the DRC CCR item bank that had not yet been field tested in South Carolina. We are concerned the item statistics are based on students from states other than South Carolina, as *5.2 SC READY Multi-State Common Calibrations.docx* describes. This practice would not satisfy the portion of this standard specifying the sample is as "representative as possible of the population(s) for which the test is intended." There could be differences in the ability distributions of South Carolina students and the students who provided data for the item. With item parameters such as *p*-values and point-biserials used to assemble forms, this concern is even greater. However, we have no concern with using other states' items as FT items, except perhaps for grade 3 ELA, given that the grade 3 form is pre-equated according to the *SC READY Form Construction Guidelines_101817.pdf* document. We assume a post-equating check is performed. This should be documented.

The *SC READY 2017 Technical Report_100917.pdf* described the pilot test conducted in 2014 to collect preliminary data for the item pool. Because pilot tests use a volunteer sample, resulting item parameters may be affected by the (a) motivation of the examinees and (b) representativeness of the convenience sample. As we have seen in other testing programs, these item parameters are not likely to be stable or correct and they should be used with caution.

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 assessments are identical in terms of their CTT targets. The guidelines for picking "good" items are also identical and satisfy that portion of the *Test Standards*. Our review of the documents indicates 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 appears to be used only for equating purposes and was not used for forms construction. During a phone interview with SCDE and DRC staff on March 1, 2017, DRC staff confirmed this assumption. The latest documentation (*SC READY Form Construction Guidelines_101817.pdf*) states that non-convergence is rare but does not address evidence of model fit.

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.

As described in our first report (Dickinson et al., 2017), according to a phone interview with SCDE and DRC staff on March 1, 2017, SCDE staff indicated all items are subjected to comparison between paper and computer-based data for mode differences. On this call, SCDE staff indicated no items have been categorized as an ETS "C" level since 2008. If any items were to reach that level, they would be sent for content review and not immediately made ineligible for future forms.

Since the first report, we received a document from DRC describing a mode comparison study on the spring 2016 SC READY assessments (*027_SC READY and SCPASS Spring 2016 Test Mode Comparability Study.pdf)*. Two separate methods of investigating mode of test administration differences are presented in this report. First, DRC utilized the Mantel-Haenszel (for MC items) or the Standardized Mean Difference (for TDA items) for detecting differential item functioning (DIF). If an item is flagged for DIF this indicates that one group outperformed the other group once the effects of differences in skill levels between the two groups have been removed. One of three severity classification categories was assigned to each item. The A category represents negligible DIF. The B category indicates moderate potential DIF, and the C category indicates that there is large potential DIF. For ELA there were only 2 of 449 items (across grades 3-8) that were identified as C-DIF. For math, there were no C-DIF items identified. This evidence supports the conclusion that the items on the paper-and-pencil tests and the items on the online tests are not functioning substantively different. DRC further investigated mode of test administration differences by calculating the difference in *p*-value (i.e., the proportion answering the item correctly) between the item in the paper-and-pencil and online modes. The *p*-value differences between the math paper-and-pencil and online tests across grades 3-8 showed that the percent of items with absolute value of the *p*-value difference greater than .06 was approximately 4%. For, ELA the percent of items across grades 3-8 with *p*-value differences greater than .06 was approximately 16%. In both cases, the direction of the difference favored students taking the paper-and-pencil test (i.e., greater proportion of students correctly answered items on the paper-and-pencil test). Moreover, for ELA even though the magnitude of the *p*-value differences tended to be small (i.e., less than .06 for most items), the vast majority of the ELA items had slightly higher *p*-values (i.e., easier) on the paper-and-pencil tests than on the online tests. The comparability study report presents these differences in terms of their impact on overall raw scores. For math, the differences in item *p*-values equates to less than one score point, on average, between paper-and-pencil and online tests across all grades. For ELA, the overall raw score differences range from about 1.4 score points to 3.3 score points, which is 4 to 8 times larger than the differences for math. The same analysis was conducted for the SCPASS science and social studies tests. The results for science and social studies were very similar to the results for math—that is, none of the items were flagged for C DIF and only about 4% to 5% were flagged for absolute *p*-value differences greater than .06.

Overall, the mode comparability study indicates that, for math, there is good comparability between paper-and-pencil and online administrations. For ELA, the C-DIF results indicate that the individual items are not functioning substantively different between paper-and-pencil and online administrations; however, the *p*-value analysis indicates that, overall, the items tend to be consistently slightly easier for students taking the paper-and-pencil tests, which equate to overall raw score differences that favor paper-and-pencil examinees. There are various possibilities that could account for this pattern of results. There may be a systematic bias against ELA items on the online test—for example, passages might be harder for examinees to scroll through and read (see Chapter 4, Task 4—Review of Test Administration—for additional discussion on this topic). We recommend that a study be conducted to determine whether construct irrelevant variance associated with the online items (e.g., difficulty scrolling to read passages, lack of familiarity with tools, etc.) may be contributing to the lower *p*-values for the online ELA items. Another possibility is that the population of students taking the online ELA tests might have lower ability than the population of students taking the online math tests. It is worth noting that the smallest number of online administrations occurred for ELA (as compared to math, science, and social studies). To further elucidate mode differences, a propensity score matching study could be conducted whereby those who complete the paper-and-pencil tests could be matched (via propensity score matching) to similar ability students taking the online tests to determine if mode differences exist among matched samples of test-takers.

Task 3: Discussion

This chapter presents an evaluation of DRC's test construction processes for English 1 and Biology 1 EOCEP assessments, as well as new information regarding the Algebra 1 EOCEP and SC READY assessments. Since the first report, we conducted a site visit of the SC READY forms assembly process, and we reviewed additional documentation on all the assessments. Therefore, our evaluation of test construction processes is now final for all the reviewed assessments. We evaluated several of the same documents as before for the three EOCEP assessments—thus, the results in the current report (particularly the fidelity to procedures section) are identical in many ways to that presented in the first report. For this report, we evaluated the item bank metadata for English 1 and Biology 1 separately, but had similar conclusions.

The evidence supporting the EOCEP assessments was found to have strong compliance with the applicable *Test Standards* for forms construction. The English 1 and Biology 1 assessments both had a mean rating of 4.40 (on a 5-point scale). The Algebra 1 EOCEP assessment mean rating increased from 3.87 to 4.00, given the addition of new information since our first report. The psychometric review of the item bank metadata for Biology 1 and English 1 was largely positive. Overall, the statistics for items in these banks suggest that form assembly should not be hindered. This should allow staff to focus on meeting content constraints, the principle goal of form assembly. The EOCEP item banks appear strong and should allow for forms to be assembled that meet psychometric guidelines.

Based on the additional information provided since our first report and information we obtained from observing the forms construction meeting, the SC READY mean rating rose from 4.40 to 4.60 (on a 5-point scale). The current report contains additional detailed feedback and suggestions, which mostly stem from our observation of test form construction. Some of our suggestions relate to the standards reviewed, but most of our recommendations address aspects we perceive as potential risks that could threaten the goal of a valid and reliable test form, such as conducting a quality control review of the manual steps involved in forms construction.

Chapter 4: Review Test Administration Procedures (Task 4)

Carrie Wiley & Jing Chen

Task 4: Introduction

The purpose of Task 4 was to document the extent to which the test administration processes of SC READY and EOCEP assessments follow best practices as described in the *Test Standards*.[27] In this chapter, we first introduce the methods we used to evaluate the test administration processes of the SC READY and EOCEP assessments. Then, we describe the results, organized by each standard, for the *Test Standards* that are relevant to test administration. Finally, we discuss our findings and provide recommendations for improving test administration procedures.

Task 4: Method

Documentation

We conducted a systematic document review to evaluate the test administration processes of SC READY ELA and math and EOCEP English 1, Biology 1, and Algebra 1. We worked in cooperation with the South Carolina Education Oversight Committee (EOC), the South Carolina Department of Education (SCDE), and the Data Recognition Corporation (DRC), with primary support provided by DRC, to obtain documentation of the South Carolina test administration processes for each assessment. We also searched the SCDE website to identify additional relevant information.

The documents we collected fall into several categories based on their foci, such as test administrator training materials, test accommodation guidelines, and test security procedures. Table 4.1 lists all the documents we collected and reviewed. These documents provided useful information about various steps and procedures related to South Carolina's test administration procedures.

Review Process[28]

Our evaluation of South Carolina's test administration processes was informed by the *Test Standards*. We identified 14 standards from the *Test Standards* that were directly relevant to test administration and rated the degree to which the documents we reviewed indicated compliance with each standard. The rating scale ranges from a score of 1 to 5, with higher scores indicating greater compliance with the standard. The relevant *Test Standards* can be found in the results section of this chapter.

[27] English 1, Biology 1, and Algebra 1 are the EOCEP assessments we evaluated for Task 4.
[28] The process for reviewing materials for adherence to relevant *Test Standards* is the same process as used in Task 1 (Review of Item Development Processes), Task 4 (Review of Test Administration Processes), and Task 5 (Review of Scaling, Equating and Scoring Processes).

Table 4.1. Test Administration Documents Reviewed

Document Focus	Document File Name	Relevant Assessment(s)	
		EOCEP	SC READY
Test Administration Manuals (TAMs) for computer-based tests (CBT) and paper-based tests (PBT)	Spring 2017 EOCEP TAM.pdf	X	
	Spring 2017 EOCEP MRRS.pdf	X	
	Spring 2017 SC READY_SCPASS MRRS.pdf		X
	Spring 2017 SC READY_SCPASS ADM.pdf		X
	Spring 2017 SC READY_SCPASS TAM.pdf		X
Technical Manuals	SC READY 2017 Technical Report_100917.pdf		X
	2016-17 EOCEP Technical Report for HumRRO.pdf	X	
Test administration systems	[a]DRC INSIGHT Technical Guide eDIRECT User Guide.pdf	X	X
Test administrator training materials (e.g., on-line tutorials, print tutorials) for CBTs and PBTs	2016-2017 Technical Training Presentation.pptx	X	X
	Spring 2017 EOCEP Pretest Workshop.pptx	X	
	Spring 2017 EOCEP STC TA Training Tool.pptx	X	
	Spring 2017 SC READY_SCPASS Pretest Workshop.pptx		X
	Spring 2017 SC READY_SCPASS STC TA Training Tool.pptx		X
Supplement materials from SC DOE website	https://ed.sc.gov/tests/high/eocep/ https://ed.sc.gov/tests/middle/south-carolina-college-and-career-ready-assessments-sc-ready/	X	X
Access to the test delivery systems, the online practice tests	Tutorials and Online Tools Training.docx	X	X
	Spring 2017 SC READY Brochure.pdf		X

[a]Indicates a folder that includes multiple files

The rating scale is presented in Table 4.2. For each of the relevant *Test Standards*, two HumRRO researchers independently assigned an overall rating based on the evidence collected and reached consensus through discussion of discrepant ratings.

Table 4.2. Rating Scale for Evaluating Strength of Evidence for Test Standards

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

[a] Materials include all documents and data provided, any emails or phone calls with SCDE/DRC staff, as well as information available online.

The information we collected indicates the test administration processes are generally the same for the SC READY and EOCEP assessments. Consequently, our results are presented across the SC READY and EOCEP assessments.

With any review of test administration procedures, fidelity of administration and adherence to protocols is vital to reduce the impact of construct-irrelevant variance on student achievement. The *Test Standards* for test administration can be classified into three categories: ensuring (a) documentation related to standardization and security is provided, (b) documentation is clear and usable, and (c) procedures outlined in the documentation are followed. Our review focuses largely on the first category and to a limited extent, the second category; that is, we discuss whether the provided documents clearly addressed each standard, but we cannot fully evaluate the extent to which they are clear and usable to test administrators and test users given the restricted scope of the Phase 2 evaluation. Our third and final report, to be delivered June 2018, will address adherence to protocols and the usability of the documentation and materials based on a small sample of observations of test administrations and interviews with test administrators.

Task 4: Results

Results are organized around the relevant *Test Standards* and include details from our documentation review of test administration procedures and processes to support judgments about the extent to which industry standards are met. Table 4.3 provides an overall rating (described above) for each relevant *Test Standard* after reviewing all available information related to each assessment.

Table 4.3. Evaluation Results for Test Administration Procedures Based on the Test Standards

Standard Number	Standard Content	Rating
Standard 3.10	When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.	4
Standard 4.5[a]	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 4.15	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.	5
Standard 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.	4
Standard 6.1	Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.	4

Table 4.3. (Continued)

Standard Number	Standard Content	Rating
Standard 6.2	When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.	4
Standard 6.3[b]	Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to test users.	3
Standard 6.4	The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.	5
Standard 6.5	Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.	4
Standard 6.6	Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.	5
Standard 6.7	Test users have the responsibility of protecting the security of test materials at all times.	5
Standard 7.7[b]	Test documents should specify user qualifications that are required to administer and score a test, as well as the user qualifications needed to interpret the test scores accurately.	5
Standard 7.8[b]	Test documents should include detailed instructions on how a test is to be administered and scored.	4
Standard 7.9[b]	If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.	5

[a]Indicates the Standard is applied to other aspects of the assessment and is also evaluated in Task 3 Review Test Construction Processes.
[b]Indicates the Standard references assessment-related processes other than test administration (e.g., scoring).

Rationale for Test Administration Test Standards Evaluation Ratings

This section is organized by the *Test Standards* that formed the basis of our evaluation. For each standard, we describe the rationales of our rating and explain to what extent the standard was met. We also provide suggestions for improvement to better align with the standard. We do not address elements of these standards that do not directly pertain to test administration (e.g., scoring or detecting cheating).

Standard 3.10 – When test accommodations are permitted, test developers and/or test users are responsible for documenting standard provisions for using the accommodation and for monitoring the appropriate implementation of the accommodation.

Evidence from the documents indicates that key aspects of Standard 3.10 are covered. Test developers documented standard provisions for using the accommodation. For example, detailed provisions of testing students with documented disabilities are documented in Appendix C of the TAMs for EOCEP and SC READY. In the User Guide that introduced the interface to the administrative functions of the DRC INSIGHT Online Learning System (*eDIRECT User Guide.pdf*), the test developers list all accommodations available for students testing online, and provide tips to conduct online oral administration and update and/or change accommodations for a single student and multiple students. Though there is training for test administrators to administer the tests with accommodations (e.g., *Spring 2017 EOCEP Pretest Workshop.pptx, Spring 2017 SC READY_SCPASS Pretest Workshop*), little evidence can be found in the documents that the implementation of the accommodations is carefully monitored to ensure that test administrators implement accommodations appropriately.

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.

In both the EOCEP and SC READY *Test Administration Manuals* (TAMs) and the *Administration Directions Manual* (ADM, [29]) the test developers clearly document the permissible variation in and rationale for test administration conditions, (e.g., different types of accommodations for students with disabilities). All three manuals have distinct sections detailing the procedures for online and paper-pencil testing. Appendix C in both TAMs provide definitions and administration procedures for specific accommodations. Additionally, the SC READY website has a useful FAQ document for district and school personnel regarding accommodation procedures.[30]

In addition, DRC, SCDE, and a team of South Carolina educators conducted a validity study to investigate the impact of oral/signed administration on the validity of SC READY ELA (sessions one and two) assessment.[31] They concluded that the use of oral/signed administration does not impact the validity of the assessment in grades 4–8. The study suggests that the test developers have collected evidence to investigate whether the target construct is altered by allowable variations.

Standard 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.

The TAMs and ADMs for online and paper/pencil testing of EOCEP and SC READY provided sufficient clarity and details. For example, the manuals included directions for both school test coordinators and test administrators, directions for administering both online and paper-and-pencil testing. In addition, the TAMs and ADMs included general test administration directions for all subjects and specific test administration directions for specific subjects for both online administration and paper/pencil administration. The verbal script in the ADMs provide enough details and clarity so that others can easily replicate the administration conditions and thereby support the reliability and validity of the assessments.

The TAMs also describe allowable variations in administration procedures. For instance, in the TAM for SC READY, the test developers list three acceptable alternatives for ensuring that students placed in Residential Treatment Facility (RTF) are appropriately assessed (see details on p.24). The process for reviewing requests for additional testing variations is also documented. For instance, in the TAM for SC READY, it is mentioned that “testing must be conducted during the published schedule for the specific test or District Test Coordinators (DTCs) must provide the SCDE with a written request for an alternative schedule” (p.25).

[29] The EOCEP ADM is embedded within the EOCEP TAM. The SC READY ADM is a stand-alone document.

[30] https://ed.sc.gov/scdoe/assets/File/tests/middle/scready/SC_READY_AccommodationsFAQ_FINAL2.pdf

[31] https://ed.sc.gov/tests/tests-files/sc-ready-files/memorandum-oral-administration-on-the-sc-ready-ela-12-12-16/

Although there is sufficient documentation to replicate administration conditions across various settings, the organization of the TAMs could be improved. The overall structure flows; however, the SCDE Policies section has information regarding all phases of the test administration process and may be confusing as a Site Test Coordinator (STC) or Test Administrator (TA) reads about processes that have not yet been discussed in the TAMs. For example, SC READY TAM (p. 36) details the timing and break procedures during administration; however, page 65 of the Test Administrator's Section only indicates that breaks should be scheduled as needed, with no reference to the details on page 36. Organizing all the necessary requirements in one section would minimize the need to reference multiple sections of the document, reducing the potential to miss policies and procedures pertinent to standardization, which is particularly concerning when sections do not prompt the STC or TA to review specific sections. The current SCDE Policies section could be included as an Appendix to highlight the specific Department of Education Policies in one document. Additionally, the TAMs indicate what TAs and Monitors are permitted to answer, but do not indicate in the ADM script a specific verbal response. Including scripted responses to frequently asked questions, particularly those that TAs and Monitors are not permitted to answer could improve standardization across administrations.

Standard 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.

Evidence from the documents indicates that key aspects of Standard 4.16 are covered. The Online Tools Training (OTT) and tutorials are available to students for both EOCEP and SC READY (see files *Spring 2017 SC READY Brochure.pdf* and *Tutorials and Online Tools Training.docx*). Sufficient details are provided to test takers so that they can respond to a task in the manner that the test developer intended. There are video tutorials that provide clear instructions about how to sign in and how to use basic and advance tools of the online testing system. Information such as item types, sample items for each item type, and scoring rubrics of the writing component is available to test takers before the test date. However, practice materials may not be available in formats that can be accessed by all test takers. We did not find practice materials in a form that can be accessed by students with disabilities. Practice materials may not be suitable for students with certain disabilities (e.g., deaf or hard of hearing and sign language accommodation), but practice materials with some types of accommodations (e.g., large-print) can be provided to make the materials more accessible to test takers.

Standard 6.1 – Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

Evidence from the documents indicates that key aspects of Standard 6.1 are covered. DRC provide appropriate training and documentation so that TAs understand the standardized procedures to follow. The TAMs include accepted standardized procedures for determining accommodations, minimum technology requirements, test time limits, test make-up policies, and other acceptable variations in test administration. There are training and pretest workshops for school test coordinators, TAs, and technology coordinators.[32]

[32] We did not observe actual live training sessions and our evaluation is based on the training materials only.

The training materials provide instructions for TAs for when they need to make adjustments if an accommodation is required. In the SC READY training materials, some exceptions for administering the assessments in the online format are specified. For example, students who cannot take online assessments due to their disabilities, as specified in their IEPs or 504 plans, may be tested in a paper-based format. In the Training tool slides (*Spring 2017 EOCEP STC TA Training Tool.pptx, Spring 2017 SC READY_SCPASS STC TA Training Tool.pptx*), the test developers provide case scenarios related to test security to train TAs to deal with different test security issues. Similar hands-on training or concrete examples for other phases of administration could be provided to TAs as well to improve the training to ensure that TAs carefully follow the standardized procedures. Additionally, we did not find documentation about usability studies or empirical research related to topics of test administration.

Standard 6.2 – When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

Test takers are informed about appropriate accommodations procedures. The TAMs state the requirements for notifying parents in advance of the testing schedule, testing format, and any special conditions that apply to the testing of their children. For both SC READY and EOCEP, there are student video tutorials about how to use accommodation features. Detailed information about accommodations is provided in the TAMs. For SC READY, there is a list of *Frequently Asked Questions* (FAQ) to address common questions from students and district and school personnel about accommodations and accessibility.[33] Clear lists are provided to students in advance of the testing date regarding online testing accommodations and paper-and-pencil testing accommodations for students with disabilities and English Language Learners. However, for EOCEP, information about accommodations is mainly provided in the TAM, which is less accessible for test takers. We recommend providing a list of online and paper-and-pencil testing accommodations for the EOCEP assessments that are designed specifically for students rather than TAs. The list could be similar to what is provided for the SC READY assessments (see the *SC READY Online and Paper/Pencil Tools and Supports* file).[34] Also, a FAQ list could be provided to students to address common questions about accommodations and accessibility.

Standard 6.3 – Changes or disruptions to standardized test administration procedures or scoring should be documented and reported to test users.

The TAMs generally specify procedures related to deviations from standard procedures, such as disruptions to testing environments (e.g., fire drills, bomb threats) and administering accommodations. However, the guidance does not clearly indicate how other changes or minor disruptions to standardized test administration procedures during operational testing should be documented and reported to ensure that other testing conditions do not systematically impact score interpretation (e.g., recording technology issues, loud noises, classroom management issues).

The TAMs clearly state the appropriate processes to report and document test security violations. For example, it is specified that the District Test Coordinator (DTC) and the School Test Coordinator (STC) are responsible for conducting a comprehensive investigation of each allegation. The DTC must prepare and submit to the SCDE all required documentation that serves as a summary of the information obtained from the investigation.

[33] https://ed.sc.gov/scdoe/assets/File/tests/middle/scready/SC_READY_AccommodationsFAQ_FINAL2.pdf

[34] https://ed.sc.gov/scdoe/assets/File/tests/middle/scready/SC%20Ready%20Accommodations%20Charts_12_31_15.pdf

Standard 6.4 – The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

The test administration processes follow this standard very well. In both the EOCEP and the SC READY TAMs, there is a section about the testing environment that specifies standards to be followed to provide a reasonably comfortable testing environment to test takers. The guidance specifies that schools must adhere to several standards to ensure that all students have an equal opportunity to perform their best on the test. Some examples of these standards include "tests should be administered in a familiar classroom or computer lab setting to reduce student test anxiety and simplify test security," "students should be tested in classrooms or computer labs that have good lighting and are well-ventilated with a reasonable temperature," and "classrooms and computer labs should be quiet and free from interruptions or distractions of any type." The technical guide documents (*DRC INSIGHT Technical Guide* and *eDirect User Guide*) provide technical instructions for using the online testing system. This helps to reduce distractions due to internet connectivity issues and technology failures and avoid construct-irrelevant variance.

Standard 6.5 – Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.

Instructions to test takers regarding how to respond and interact with the test delivery interface are clearly indicated in the TAMs, ADMs, Online Tools Training (OTT), and student tutorials. Guidance for how to interact with and navigate the delivery platform, use the available tools, and respond to items are provided. The *eDIRECT User Guide* and the TAMs state that STCs and TAs are responsible for (a) reviewing the OTT and Tutorial prior to testing, and (b) ensuring that students practice on the device they will be taking the operational test prior to testing.

While the OTT and the Tutorial adequately address the issue of test takers being provided appropriate instructions and practice prior to operational testing, the documents we reviewed do not detail the part of the standard that addresses monitoring those practice opportunities. The documents provide little information regarding providing guidance to the STCs and TAs to ensure that the practice opportunities lead to students acceptably interacting with the testing engine (e.g., navigating, marking responses).

One area of importance with online testing is that students understand how to scroll through passages commonly seen on ELA tests (and sometimes in other subjects). The EOCEP English 1 and Biology 1 passage navigation (as evidenced by our review of the OTT) has a seamless transparent blue bar with white font indicating if there is more text to scroll through at the bottom and top of the passage screen. The SC READY ELA test, however, uses a pagination navigation screen at the bottom of the passage. For example, if a passage has four pages to scroll through, the bottom left of the passage will say 'Page 1 of 4.' However, clicking to the next page is not immediately made clear—in order to do so, one must click the right side of the passage to advance forward or the left side to go backward. The script in the ADM does include specific instructions on how to navigate, but the OTT and Tutorial does not directly address this issue. We have some concerns that younger students, in particular, may have difficulty accessing the entire passage without appropriate practice, exposure, and guidance. The scrolling passage navigation as used in the EOCEP assessments might be easier for younger students; however, consideration to which passage navigation is most intuitive and easiest for younger students should be guided by usability studies or cognitive labs.

Additionally, there were some aspects of the Tutorial that might use language that is too advanced for younger students. For example, "The ELA test will be a two-day test. For ELA

Session 1, the extended response item will be a text dependent analysis or TDA item" could use simpler language or more teacher-guided direction for younger students.

Standard 6.6 – Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.

This standard includes providing (a) safeguards against fraudulent activities at the local school sites and during administration, and (b) measures to detect cheating during scoring processes. This standard was only reviewed in relation to documented procedures for ensuring the integrity of scores during test administration processes.[35] The EOCEP and the SC READY TAMs provide a separate section on test security including, state board regulations, reporting and documenting violations, and separate policies and procedures for administering online and paper tests. The TAMs also provide guidance for TAs to help reduce cheating by requiring seating charts, completion of security checklists, and providing helpful tips on how to separate students (e.g., privacy folders, space). The TA script in the ADMs includes a statement about the prohibition of electronic devices. Additionally, the training PowerPoint® files include several test security case scenario vignettes to help standardize TA understanding and implementation of test security policies and procedures.

One area that would benefit from additional specification relates to preventing breaches of accommodation policies. The TAMs identify the procedures to take should a violation occur, but there is little guidance on how to identify or minimize such breaches. It is possible, based on the criteria of who is eligible to serve as a TA, that the TA might not have sufficient knowledge of IEP/504 accommodations to be able to identify when a breach might occur.

Standard 6.7 – Test users have the responsibility of protecting the security of test materials at all times.

This standard largely means that all test users (at all levels) have the responsibility of protecting and securing test materials. Our review excludes documentation of procedures related to state agency actions (e.g., documents shown in court challenges) and focused on the responsibilities of those at the district and school level. The EOCEP and SC READY TAMs and training slides state the criteria for eligible DTCs, STCs, TAs, and Monitors and provides general requirements for ensuring test materials remain secure at all times. The TAMs include an overview of state laws regarding test security, completing required forms and checklists, and handling, storing, and returning materials.

Standard 7.7 – Test documents should specify user qualifications that are required to administer and score a test, as well as the user qualifications needed to interpret the test scores accurately.

The Task 4 review excludes documentation of procedures related to scoring and focuses only on test administration processes. (See Task 5 for a review of scoring.) The EOCEP and SC READY TAMs and training slides state the specific criteria for eligible DTCs, STCs, TAs, and Monitors and states that TAs must participate in a DTC- or STC-led training session. Although

[35] The scope of our Phase 2 evaluation reflects the documentation regarding test security processes, and not whether these policies and procedures are carried out with fidelity. For example, the TAM states, "the school should follow policies and procedures established by the district for investigating and documenting suspected cheating incidents (EOCEP p. 20)," but there is no specific guidance of what those district policies should include.

the TAMs state that certified and properly trained professionals administer the test for any administration (e.g., accommodations, residential treatment facilities), one area that would benefit from additional specification relates to the qualifications for administering IEP/504 accommodations. Since TAs need to be able to determine if students receive the appropriate accommodations, having separate requirements could help ensure that violations in the administration of accommodations do not go undetected.

Standard 7.8 – Test documents should include detailed instructions on how a test is to be administered and scored.

The Task 4 review excludes documentation of procedures related to scoring and focuses only on test administration. The EOCEP and SC READY TAMs*,* *eDirect User Guide*, and training slides clearly state the instructions on how the test should be administered. The TAMs provide both general overviews and detailed information of the test administration process including preparation (of TAs and students), test security, state policies, accommodations, and administration and monitoring of the online and paper tests. The *eDirect User Guide* provides information on how to prepare the required technology components, prepare test tickets, and complete student demographic information prior to the administration. As mentioned in other standards, our review focuses on the documentation provided. Some areas that could benefit from additional specification include (a) the inclusion of a timeline of responsibilities and actions for DTCs, STCs, and TAs, particularly surrounding the issue of entering student data, and (b) indication of how schools can receive technical help with the online test. We saw little indication of a Help Desk available for preparation and during actual administrations.

Standard 7.9 – If test security is critical to the interpretation of test scores, the documentation should explain the steps necessary to protect test materials and to prevent inappropriate exchange of information during the test administration session.

The Task 4 review excludes documentation of procedures related to scoring and focuses only on test administration processes. Overall, the EOCEP and SC READY TAMs and training slides explain what is required to protect test materials. They describe state law and policy regarding test security, and requirements for storing and handling materials. Additionally, the training slides include several test security case scenarios to ensure that TAs are trained on appropriate procedures regarding test material access and policies.

Task 4: Discussion

Our evaluation of EOCEP English 1, Biology 1, Algebra 1 and SC READY test administration processes focused on available documentation and materials. Furthermore, our review of Test Administration procedures focused on components of the *Test Standards* that focus specifically on test administration processes. The *Test Standards* that describe processes related to scoring or detecting cheating are not directly related to test administration per se, and thus, were not considered for the Task 4 review.

We generally found the test administration processes of EOCEP and SC READY reflected the 14 *Test Standards* with a mean rating of 4.40 on a scale of 1 to 5 (1 indicates no evidence to support the standard and 5 indicates evidence fully supports the standard). With the exception of one standard (Standard 6.3), we found the documented policies and procedures to mostly match the key aspects of industry standards. Among the key documents (TAMs, ADMs, OTTs, and Tutorial), policies and procedures were clearly stated, comprehensive, and would likely support standardized administrations across conditions.

Based on our review, we make the following recommendations to strengthen and improve the test administration processes.

- Ensure that test administrators (TAs) administer the assessment according to standard procedures:
    - More clearly identify (a) qualifications of TAs to administer accommodations, and (b) procedures to monitor the implementation of the accommodations.
    - More clearly describe procedures for systematically documenting and reporting changes and disruptions during the assessment.
    - More clearly organize the TAMs so that all requirements are readily highlighted and known to TAs.
    - Make available a technical help desk to assist with technical difficulties during the assessment.

- Reduce construct-irrelevant factors on score interpretation related to test preparation:
    - Include information from usability studies or empirical research related to test administration to ensure that the test materials are clear and usable for all grade levels and subjects, specifically the SC READY ELA Tutorial and passage interface.
    - Provide practice materials in formats that can be accessed by all test takers (e.g., provide practice materials with accommodations that can be accessed by students with disabilities).
    - For EOCEP, we recommend providing a list of online and paper-and-pencil testing accommodations designed specifically for students rather than TAs.
    - More clearly describe appropriate procedures for operationally preparing student test tickets and entering student data.

We plan to conduct site visits to observe processes related to test administration prior to our third and final report that will further explore fidelity of administration processes.[36]

[36] The current report serves as the final analysis of the SC READY assessments and the EOCEP assessments for Biology 1 and Algebra 1. The third report will include the final analysis of the English 1 assessment. Thus, the site visits to observe test administration will be limited to the English 1 test administration.

Chapter 5: Review Scaling, Equating, and Scoring Processes (Task 5)

Hillary Michaels & Jing Chen

Task 5: Introduction

HumRRO conducted a document review to evaluate the scaling, equating, and scoring processes for the South Carolina College-and-Career Ready (SC READY) Assessments in English language arts (ELA) and math, and for the End-of-Course Examination Program (EOCEP) English I, Biology I, and Algebra I assessments as well as. The purpose of this task is to document the extent to which the equating, scaling, and scoring processes of SC READY and EOCEP assessments follow best practices described in *Test Standards.*

In this chapter, we first introduce the methods we used to evaluate the equating, scaling, and scoring processes of SC READY and EOCEP. Then, we describe the results organized by each standard identified in the *Test Standards* that are relevant to the equating, scaling, and scoring processes. Finally, we discuss our findings and provide recommendations for improvement.

Task 5: Methods[37]

We conducted a systematic document review based on industry standards to evaluate the equating, scaling, and scoring processes of SC READY ELA and math and the EOCEP English 1, Biology 1, and Algebra 1 assessments. We worked in cooperation with the South Carolina Education Oversight Committee (EOC), the South Carolina Department of Education (SCDE), and the Data Recognition Corporation (DRC), with primary support provided by DRC, to obtain documentation of the equating, scaling, and scoring processes for each assessment. We also searched the SCDE website to identify additional relevant information.

The documents we collected fall into several categories based on their foci, such as technical specifications for item calibration, equating, scoring, documentation of item scoring procedures, and rater training materials. Table 5.1 lists the 37 documents we collected and reviewed. These documents provided useful information about various steps and procedures associated with the equating, scaling, and scoring processes.

[37] The process for reviewing materials for adherence to relevant *Test Standards* is the same process as used in Task 1 (Review of Item Development Processes), Task 4 (Review of Test Administration Processes), and Task 4 (Review of Test Administration).

Table 5.1. Documents Reviewed for Task 5 – Equating, Scaling, and Scoring

Document Focus	Document/Folder File Name	Relevant Assessment(s)	
		SC READY (ELA, math)	EOCEP (Algebra 1, English 1, Biology 1)
Technical specifications for item calibration, equating, and scoring. Technical reports and special studies.	[a]024F_EOCEP Reports_Technical_Standard Setting_Special Studies		X
	[a]025F_SC READY Reports_Technical_Standard Setting_Special Studies	X	
	[a]029F_Reading PLDs	X	
	SC-MAP-Linking-Study	X	
Documentation of item scoring procedures; Quality assurance processes for automated scoring	[a]028F_Phase I_Item Development _Forms Construction Document	X	X
	043_Item Scoring and Quality Control	X	X
Scorer training materials (TDA only).	[a]015F_SC READY Scorer Training Materials	X	
Criteria for scorer qualification (TDA only)	039_SC READY Scorer Qualification	X	
Processes for monitoring scorer accuracy and consistency (TDA only)	040_SC READY Scorer Accuracy and Consistency	X	
Documentation related to creation of vertical scales (SC READY only)	[a]027_2017 SC READY Vertical Equating	X	
	042_SC READY Creation of Vertical Scales	X	
	047_SC READY Horizontal_Vertical Linking Process	X	
Sample 2016-2017 student and school score reports	041_EOCEP Score Report Users Guide		X
	045_Spring 2017 SC READY Score Report Users Guide	X	

[a] Indicates a folder including multiple files.

This evaluation of the equating, scaling, and scoring processes was informed by industry best practices as outlined in the *Test Standards*. We identified 16 standards directly relevant to our work. To evaluate the quality of the available information against the standards, a rating scale was developed. The rating scale is presented in Table 5.2. The identified standards are listed in Table 5.3. For each identified standard, at least two HumRRO researchers independently assigned a rating based on evidence reviewed. The researchers compared and discussed their initial ratings and rationales to reach a final consensus rating for each standard for both the SC READY and EOCEP assessments.

Table 5.2. Rating Scale for Evaluating Strength of Evidence for Test Standards

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

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

Task 5: Results

Results are organized around the relevant *Test Standards* and include details from our documentation review of test equating, scaling, and scoring processes to support judgments about the extent to which industry standards are met. Table 5.3 provides ratings by each relevant standard for each assessment program (SC READY and EOCEP). The materials for the SC READY assessments (ELA grades 3-8 and math grades 3-8) were the same or nearly the same; thus, the findings for SC READY apply across all the SC READY assessments. Similarly, the materials for the EOCEP assessments (English 1, Biology 1, Algebra 1) were the same or nearly the same; thus, the findings for EOCEP apply across all three assessments. The results show that the SC READY and EOCEP assessments have documents providing evidence that most, if not all, of the relevant *Test Standards* are well covered. Because the EOCEP assessments are structured differently than the SC READY assessments (e.g., there is no vertical scale for the EOCEPs), some of the standards are not applicable (NA) to the EOCEP assessments.

Table 5.3. Evaluation Results Based on the Test Standards

Standard Number	Standard Content	SC READY Rating	EOCEP Rating
Standard 5.1	Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.	4	4
Standard 5.2	The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.	4	4
Standard 5.5	When raw scores or scale scores are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretation should be explained clearly.	5	5
Standard 5.6	Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which the scores are reported.	4	4
Standard 5.8	Norms, if used, should refer to clearly described populations. These populations should include individuals or groups with whom test users will ordinarily wish to compare their own examinees.	4	NA
Standard 5.12	A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used inter-changeably.	4	4

Table 5.3. (Continued)

Standard Number	Standard Content	SC READY Rating	EOCEP Rating
Standard 5.13	When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.	4	4
Standard 5.15	In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specification and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented.	4	NA
Standard 5.17	When scores on tests that cannot be equated are linked, direct evidence of score comparability should be provided, and the examinee population for which score comparability applies should be specified clearly. The specific rationale and the evidence required will depend in part on the intended uses for which score comparability is claimed.	4	NA
Standard 5.18	When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.	4	NA
Standard 5.21	When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.	5	5
Standard 5.22	When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.	5	5
Standard 5.23	When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.	4	5
Standard 6.8	Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgement should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.	4	NA
Standard 6.9	Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.	4	NA
Standard 6.10	When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how score are intended to be used.	4	4

The following section is organized by the *Test Standards* used for the Task 5 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 where appropriate.

Standard 5.1 - *Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations.*

SC READY: The technical documentation, such as the *Technical Report* and *Score Report Users' Guide* clearly outline the purpose of the test. The *Score Report Users' Guide* includes information on the score levels, types of items, and the set of generated reports with descriptions of how reported data should be interpreted and used. The SC READY individual student reports include scale scores and information about score precision and related performance levels and performance level descriptors (PLDs). The reports include an ELA Reading subscale performance level reflecting the State's interest in reading. Subscale Reading PLDs are computed, but are not provided on the student report or referenced as a link.

The SC READY assessments are vertically scaled based on a linking study. The method used to create the vertical scale is psychometrically sound and has utility for performing several types of analyses at the system, district, and school levels; for example, examining mean scores of students across grades from year to year to look for changes in growth patterns could help the state or district to determine if large scale programmatic changes have their desired impact. However, the vertical scale could be potentially misleading to some stakeholders, including teachers, parents, and students due to the large overlap in the scale from one grade to the next and given that the same reporting scale is used across grades. As a result, stakeholders may erroneously conclude, for example, that a student in grade 3 scoring near the maximum and who has a score that is equivalent to that of typical eighth grader has mastered grade 8 content. Even though the scores are the same, this would be an erroneous conclusion given that the grade 8 content on which the grade 8 assessment is based is very different than the grade 3 content on which the grade 3 assessment is based. To help guard against such potential confusions, the *Score Report Users' Guide* should more clearly explain interpretations of the vertical scale and their limitations.

EOCEP: The technical documentation, such as the *Technical Report* and *Score Report Users' Guide* clearly outline the purpose of the test. Further, the *Score Report Users' Guide* includes information on score levels, types of items, and the set of generated reports with descriptions of how reported data should be interpreted and used at the summary and individual level. Beginning with the 2016-17 School Year, the EOCEP scale scores correspond to the Revised South Carolina Uniform Grading Scale (revised in 2016). The EOCEP individual student reports do not include error bands, but they include links to related references, such as performance level descriptors (PLDs), Uniform Grading Scale, and curriculum standards. The explanation of scale score limitations could be enhanced on the report by providing the standard error of measurement.

Standard 5.2 - ***The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.***

SC READY: The SC READY tests in grades 3-8 math and grades 4-8 ELA are post-equated. The grade 3 ELA test is pre-equated. This information is not readily available in the *Technical Report*, but it is included in other documentation such as the *Guidelines for Item Analysis and Form Construction*. Information on the scales and scale/score precision is provided in Chapters 7 and 8 of the *Technical Report*. Additional documentation on the horizontal and vertical linking is provided that adds details, such as which items are included in linking and the process for removing items from the linking set due to parameter drift. As mentioned in the rationale for Standard 5.1, additional detail and explanation should be provided to test users on interpretations and limitations of the recently developed vertical scale.

EOCEP: The EOCEP scales are described in a couple of documents including the *Technical Report* and *Guidelines for Item Analysis and Form Construction*. The purpose of the scales is described. Information about scale precision can be found in Chapter 7 of the *Technical Report* that includes the summary of the test reliability, overall and conditional standard errors of measurement, and score consistency results. Much of this information is also suggested by the steps included in the *Guidelines for Item Analysis and Form Construction.* However, the specific guidance provided by DRC psychometricians to test developers and for form selection and pre-equating is not outlined.

The EOCEP assessments are pre-equated. DRC took over the previous vendor's item pool and statistics. Original scale development is not referenced in these documents. Post-equating checks, when necessary, and general scaling and equating design, including linking the field test items to the item bank, are described in *Guidelines for Item Analysis and Form Construction*.

Standard 5.5 - ***When raw scores or scale scores are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretation should be explained clearly.***

SC READY: SC READY scale scores are criterion-referenced. Standard setting is conducted so that the scale scores and performance levels provide more descriptive information about what students scoring in a particular level know and can do. The *Standard Setting Report* provides cut score rationales and recommended score interpretation. The *Technical Report* provides classification consistency information for the population and for demographic categories of students such as gender, ethnicity, Students with Disabilities, and English Learners. These results provide evidence that examinees have been assigned to their appropriate category. As mentioned in the rationale for Standard 5.1, additional detail and explanation should be provided to test users on interpretations and limitations of the recently developed vertical scale.

EOCEP: EOCEP scores are criterion-referenced. Standard setting is conducted so that the scale scores and performance levels provide more descriptive information about what students scoring in a particular level know and can do. The *Standard Setting Report* and report addendum provide cut score rationales and recommended score interpretation. The *Technical Report* provides classification consistency information. These results provide evidence that examinees have been assigned to their appropriate category. In addition, the *SCDE Addendum to the Standard Setting Report* described policy-based adjustments based on (a) standard setting panelists' recommendations, (b) the confidence intervals for the panel recommended cut scores, (c) estimated percentages of students assigned to performance levels and Uniform Grading scale, and (d) approximate college ready percentages of the ACT WorkKeys and ACT.

Standard 5.6 - ***Testing programs that attempt to maintain a common scale over time should conduct periodic checks of the stability of the scale on which the scores are reported.***

SC READY: SC READY tests were placed onto the South Carolina grade level scale in the 2015-2016 school year. This is the base horizontal scale. In the 2016-17 school year, the vertical scale was developed after being discussed with South Carolina's Technical Advisory

Committee and conducting simulations. The vertical scale ranges from 100 to different score maximums at each grade. The performance levels set on the South Carolina grade level scale are directly comparable between the years. However, the within grade level scale scores are not directly comparable between the 2015-2016 and 2016-17 school years.

For the horizontal scale the *Forms Construction Guidelines* mention that item parameter drift is handled through equating and that linking constants are established for each administration. This way, all items in the bank do not have modified item difficulties. Moreover, the vertical scale documentation includes information stating that there are no current processes developed to check the stability of the vertical scale. However, vertical scale maintenance will be discussed at an upcoming Technical Advisory Committee meeting to put procedures in place.

EOCEP: The EOCEP seems to pull items from a bank to develop pre-equated test forms. The *Guidelines for Item Analysis and Form Construction* includes information indicating which items are being added to the pool. These items are field tested and linked to the operational test bank after test administration using an anchor set. The *Forms Construction Guidelines* document mentions that part of the post-equating check process checks for consistency with item parameter guidelines and test blueprints. The document also outlines conditions when the pre-equated results would need adjustment. The EOCEP test forms have not needed to be adjusted after post-equating checks.

Item response theory assumes that the items are not correlated and the underlying ability scale is unidimensional. To support the unidimensional assumption of the Biology 1 assessment, a principal component analysis was provided for review (*Principal Component Analysis*). The results suggested a small amount of multi-collinearity in the data.

Standard 5.8 – Norms, if used, should refer to clearly described populations. These populations should include individuals or groups with whom test users will ordinarily wish to compare their own examinees.

SC READY: The SC READY uses items in DRC's item bank. These items are used by other clients as mentioned in *SC READY Multi-State Common Calibrations* and in the *Score Report Users' Guide*. The student reports (for online and paper-pencil administrations) include normative information with the inclusion of percentile ranks based on the subset of items from the item bank. Percentile ranks are presented on the student report providing normative information for each student against other South Carolina students and other states with comparable standards. Next to the results, an explanation of percentile rank is presented. For the students who take SC READY by paper-pencil instead of online, Lexile and Quantile reports portraying the test taker's current reading or mathematical achievement are presented along with their estimated growth paths and college- and career- readiness ranges. According to the standard, norms should be clearly described. It is unclear whether the norms are user norms from the items in DRC's item bank, based on the DRC's 2011 TerraNova 3 national norming study, or from some other source.

EOCEP: No norms are reported on the EOCEP.

Standard 5.12 – A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used inter-changeably.

SC READY: The SC READY *Guidelines for Item Analysis and Form Construction* include general instructions to develop alternate forms including test blueprints. The document also includes handoffs, decisions, and reviews needed by SCDE and DRC, and between DRC content specialists and psychometricians, to develop new forms. However, currently, the SC READY assessments only include one online form and one paper/pencil form with over 90% identical items. Back-up forms would be desirable, in the event test security is compromised.

EOCEP: The EOCEP *Guidelines for Item Analysis and Form Construction* document provides general test blueprint information for constructing alternate forms. This document also includes handoffs, decisions, and reviews needed by SCDE and DRC, and between DRC content specialists and psychometricians, to develop new forms. The processes outlined in the documents indicate that the developed forms have similar statistical properties and content, and therefore, can be interchangeable. There is an on-line form for fall/winter, spring, and summer administrations, as well as a paper-and-pencil form for fall/winter, spring, and summer administrations. Back-up forms would be desirable, in the event test security is compromised.

Standard 5.13 – When claims of form-to-form score equivalence are based on equating procedures, detailed technical information should be provided on the method by which equating functions were established and on the accuracy of the equating functions.

SC READY: As mentioned for Standard 5.12, test blueprints and general forms development procedures indicate that the assessed content area constructs are consistent across administrations. The SC READY *Technical Report* includes a link to the test blueprints that specify the number of items and item types for each grade-level standard and tested subject. The *Guidelines for Item Analysis and Form Construction* include general procedures for developing the annual forms including information on the anchor sets.

The *Technical Report* describes equating procedures that allow forms to be used interchangeably. However, currently, the SC READY assessments only include one online form and one paper/pencil form with over 90% identical items. Furthermore, a couple of steps are vague in the equating procedures in the *Technical Report*. For example, no boundaries are provided to help the reader understand when results are unusual. One step states, “The distribution of students scoring in each achievement level should not vary unusually from year to year.” The boundaries for what “unusual” means may exist in other documents. In reviewing information in Chapter 7, clear flags for outlying statistics have been developed. After equating, item difficulties of the operational forms seem to match with those in the existing item pool. No mention of the processes for the grade 3 pre-equated ELA form and procedures for post-equating checks are highlighted in the *Technical Report*.

South Carolina reports reliability information on all tests for the entire population and subgroups of interest, such as different ethnicities. All reliability estimates are greater than 0.85, as recommended by the Technical Advisory Committee, except for Students with Disabilities on the Math grade 7 (α = 0.79) and grade 8 forms (α = 0.81). There is a comment that the estimates are calculated on “only Form A” (p. 42). No additional information on which form is considered Form A is provided. SC READY is usually administered by computer and there is a paper form and customized forms, such as large-print, as well. Standard errors and conditional standard errors of measurement at the cut scores are reported for the entire population since the reliability indicated that subgroups of students did not greatly differ from the population.

This standard requires that information on the size and relevant characteristics of examinee equating samples be described. The *Technical Report* includes a statement that all students who attempted the test are included (see Section 7.3) in the calibration sample. In contrast, the

SC READY Horizontal Linking Process includes a statement that the "SCDE requests a sample of at least 20,000 records" (p.1).

EOCEP: As mentioned for Standard 5.12, test blueprints and general forms development procedures indicate that assessed content area constructs are consistent across administrations. The *Technical Report* includes information about the item distribution by content domain for the fall/winter, spring, and summer administrations. Test reliability and conditional standards errors from the three administrations are also reported. The results indicate form equivalence; however, there is only one on-line form for fall/winter, spring, and summer and one paper-and-pencil form for fall/winter, spring, and summer. Information on test fairness is discussed in the *Technical Report*. It includes differential item functioning (DIF) results and a statement about the bias and sensitivity reviews.

The EOCEP *Technical Report* briefly introduced that the previous vendor conducted field tests with a sufficient number of items to create pre-calibrated item pools and to construct pre-equated operational-test forms for all tests. For all subjects, the Rasch-ability-score-to-scale-score conversion tables were produced prior to each test administration based on the item parameters in the pre-equated item pools. The equating process could be more thoroughly documented. The equating is conducted through pre-equating. We did not find detailed documentation of the item calibration process and evaluations of the adequacy of the equating functions following operational administration as required in this standard. No post-equating checks are presented in the *Technical Report*.

Standard 5.15 – In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specification and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented.

SC READY: SC READY has anchor items for vertical equating. In general, there are 15-18 anchors from the grade below. The characteristics of vertical anchor items are described in *SC READY Vertical Scale_Updated 101717.pdf*. However, some grades drop more vertical anchor items based on Robust Z (i.e., a statistic for detecting anomalous values). The general specifications and guidelines are included in the *Guidelines for Item Analysis and Form Construction* document. This document provides DIF information that the content and statistical characteristics of the anchor set reflect the test, but specific information was not provided for review.

EOCEP: The EOCEP does not use an anchor test design.

Standard 5.17 – When scores on tests that cannot be equated are linked, direct evidence of score comparability should be provided, and the examinee population for which score comparability applies should be specified clearly. The specific rationale and the evidence required will depend in part on the intended uses for which score comparability is claimed.

SC READY: A study was conducted to link the SC READY assessments to Northwest Evaluation Association's (NWEA) Measures Academic Progress (MAP) and concordance tables were provided. Results from the study provide evidence of score comparability. Students' MAP ELA scores can consistently classify students' proficiency (Level 3 or higher) status on the SC READY ELA tests 84-86% of the time, and MAP math scores can consistently classify students on the SC READY math tests 86-89% of the time. Data used in the linking study were collected in spring 2016 from matched students from 246 schools who completed both the SC READY and MAP.

The ELA sample included 78,320 students in grades 3-8, and the math sample included 78,063 students in grades 3-8. The NWEA mentioned that the results are only generalizable to test takers who do not differ significantly from the sample in the study. Data on the representativeness of the samples, content similarity of the tests and test reliabilities were not reported.

EOCEP: No linking with other assessments was conducted.

Standard 5.18 – When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.

SC READY: In the SC-MAP linking study document, the test developers briefly introduced the construction and intended interpretation of the assessments in the overview section. The test developers listed several limitations of the linking study such as the generalizability of the results. It would have been helpful if the report included more technical information on the linking methodology and the quality of linking. Details such as the constructs of each subject, the content similarity between assessments, the data collection design, and the reliability of the sets of scores being linked should also be included.

EOCEP: No linking with other assessments was conducted.

Standard 5.21 – When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

SC READY and EOCEP: The *Standard Setting Technical* Reports and Addenda are very thorough. Both used The Bookmark Method, a common item mapping method for setting defensible cut scores. The method is appropriate to the assessments and attends to how the results are used. The technical report clearly describes the discussion of test impact data with panelists after their second round of ratings, and the addenda clearly describes policy-based adjustments to the recommended cut scores.

The documentation includes description of the panelists, the Bookmark process, panelists training, and their results. Data from the three rounds are included such as the median, and minimum and maximum cut scores. Panelist variability is reported. The recommended cut scores are presented with impact data and confidence ranges.

Both the SC READY and EOCEP Technical Reports include classification consistency information, based on a beta-binomial model (Huynh, 1979; Huynh & Saunders, 1980) that support the final cut scores. Conditional standard errors of measurement at the cut scores are also reported.

Standard 5.22 – When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgements can bring their knowledge and experience to bear in a reasonable way.

SC READY and EOCEP: The *Standard Setting Technical* Reports and Addenda provide descriptions of panelist training. Panelists were introduced to and practiced the method, and reviewed the performance level descriptors and content standards. To better understand the student experience, panelists took the operational test. All rounds of rating were conducted individually, though panelists discussed their ratings after each round. In the post workshop survey, panelists generally indicated that training was clear and that they were at least partially confident in their bookmark placement. These processes indicate that panelists had a sound basis for making their judgements and were familiar with the skills and knowledge of students just transitioning into the higher achievement level based on the descriptions.

Standard 5.23 – When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

SC READY: Policymakers required the cut scores to (a) be internally consistent within content areas and grades, (b) be consistent across grades, and (c) indicate progress on the college- and career-readiness trajectory. The panelist-suggested cut scores did not adequately address the second requirement. The SCDE compared the results to the 2015 ACT Aspire and NAEP results (year unspecified). Information is provided in the *SC READY Standard Setting Report SCDE Addendum*. No data have been collected to empirically validate whether attaining the cut score (or above) on each grade level SC READY test predicts success at the next grade level.

EOCEP: In setting the final cut scores, South Carolina policymakers required that the cut scores (a) be based on college- and career-ready performance, (b) be linear with respect to the Uniform Grading Scale, and (c) produce reasonable distributions. DRC provided the confidence intervals of the panel-recommended cut scores. DRC describes how other student data (ACT, ACT subject tests, South Carolina's career-ready criterion, and ACT WorkKeys) were presented to the SCDE. The final EOCEP cut scores are consistent with the college- and career-ready impact data from these sources.

Standard 5.18 – When linking procedures are used to relate scores on tests or test forms that are not closely parallel, the construction, intended interpretation, and limitations of those linkings should be described clearly.

SC READY: In the SC-MAP linking study document, the test developers briefly introduce the construction and intended interpretation of the assessments in the overview section. They list several limitations of the linking study such as the generalizability of the results. It would have been helpful if the report included more technical information on the linking methodology and the quality of linking. Details such as the constructs of each subject, the content similarity between assessments, and the data collection design would have been informative.

EOCEP: The *Technical Report* includes a comment that the SCDE and DRC use a rapid scoring and reporting process for all test administrations. The *DRC Item Development Manual* includes information on the development of scoring keys, rubrics, and guidelines. Only the English I test includes items requiring handscoring. We will evaluate the scoring processes for the English 1 assessment in the third and final report for this project.

Standard 6.9 – Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.

SC READY: Training materials that include scorer training and anchor sets are provided in the folder 015F_SC READY Scorer Training Materials. As described in the *SC READY Scorer Accuracy and Consistency* document, rater accuracy is monitored by back reading, inter-rater reliability, and validity papers. We did not find information about systematic sources of scoring errors or required corrections. There is no rescoring policy if the inter-rater agreement levels are very low. In addition, we did not find information about trend scoring to maintain the consistency over time; however, we can infer that the consensus sets are used to maintain trend.

EOCEP: Not applicable for this reporting phase.

Standard 6.10 - When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how score are intended to be used.

SC READY: The *Score Report Users' Guide* includes multiple reports tailored to the needs and interests of different stakeholder groups—for example, students, teachers, and school administrators. The Guide includes interpretation material and is revised annually. The *Score Report Users' Guide* includes information on the score levels, types of items, and the set of generated reports with descriptions of how reported data should be interpreted and used. As mentioned in the rationale for Standard 5.1, additional detail and explanation should be included to test users on interpretations and limitations of the recently developed vertical scale.

The SC READY individual student reports include information about scale score precision and standard errors of measurement as well as related performance levels and performance level descriptors (PLDs). The reports include an ELA Reading subscale performance level reflecting the State's interest in reading. The subscale Reading PLDs are developed, but are not provided on the student report or referenced as a link. The student reports also include diagnostic information for reporting categories within ELA and math, designated by three ordinal categories: low, medium, and high. The *Technical Report* states that these ordinal categories for the diagnostic reporting categories within ELA and math do not directly correspond to the overall student performance levels (although the diagnostic category scores and overall scores are still correlated). This statement could also be included on the report or in the *Score Report Users' Guide*. Also reported is the text-dependent analysis score for the essay requiring text support and analysis from a reading passage. The score is followed by a description of the item. For all SC READY reports, including the Preliminary Grade 3 Reading Rosters, Student Rosters, Individual Student Reports, and Student Labels, interpretation information and Score Report Notes are included. Notes convey specific information about special circumstances, such as students requiring the Braille or Sign Language versions or students missing test results or taking only one day of the ELA test. Standard 6.10 suggests that research be conducted to verify that reports are correctly interpreted. No information about report revision is available at this time.

EOCEP: The *Score Report Users' Guide* includes multiple reports tailored to the needs and interests of different stakeholder groups—for example, students, teachers, and school administrators. The Guide includes interpretation material and is revised annually. The student report for EOCEP does not provide information about score precision. For example, there are no error bands that would indicate that the score is an estimate based on the test form. The purposes of the assessment are reiterated in the document. This is a way to remind users how to correctly use the data. Standard 6.10 suggests that research be conducted to verify that reports are correctly interpreted. No information about report revision is available at this time.

Task 5: Discussion

We evaluated the equating, scaling, and scoring processes for SC READY and EOCEP assessments. Our evaluation is based on available documentation collected from SCDE and DRC. The results from Chapter 7 of this report will provide additional information such as the accuracy of item calibration and equating procedures from independent replications. Our review focused on components of the *Test Standards* that specifically address scaling, equating, and scoring. We found the equating, scaling, and scoring processes of the SC READY and EOCEP assessments generally adhere to industry best practices. The *Technical Reports*, *Guidelines for Item Analysis and Form Construction,* and *Score Report Users' Guide* include a great deal of technical information presented in an understandable manner.

Based on our review, we make the following recommendations to strengthen and improve understanding of scaling, equating, and scoring of the SC READY and EOCEP assessments, and, thus, adherence to the relevant *Test Standards*.

For SC READY:

- Provide additional detail and explanation to test users on interpretation and limitations of the newly created vertical scale.
- Provide additional detail on the population on which the percentile ranks are based to ensure the population is representative of South Carolina students.
- Develop back-up forms.
- Include specific information regarding the grade 3 ELA test in the Technical Report.
- Resolve the discrepancy between the students included in the calibration sample. The *Technical Report* reports that all students who attempted the test are included (see Section 7.3) in the sample. The *SC READY Horizontal Linking Process* includes a statement that the "SCDE requests a sample of at least 20,000 records" (p.1).
- Provide more detailed information about how the content and statistical characteristics of the anchor set reflect the test.
- Conduct a study to empirically validate whether proficiency in one grade predicts proficiency in the next grade.
- Document rater qualifications for verification.
- Provide information or procedures for calibrating raters.
- Provide information or procedures on any rescoring policies.
- Include a statement in the *Score Report Users' Guide* about how the ordinal categories (low, medium, high) are not related to the overall student performance levels.
- Provide information or reference links to the subscale Reading PLDs on the student report.
- Conduct research to ensure that score reports are correctly interpreted.

For EOCEP:

- Provide more detailed information about the original scale development work done by the prior vendor.
- Develop back-up test forms for the fall/winter, spring, and summer administrations.
- Enhance technical documentation with detailed information about item calibration steps and evaluation of the adequacy of the equating functions following operational administration.
- Where appropriate, include post-equating verification information.
- Include a measure of score precision on individual student reports such as standard errors of measurement or error bands.
- Conduct research to ensure that score reports are correctly interpreted.

Chapter 6: Review of Psychometric Processing and Item Parameters (Task 6)

Emily Dickinson, Erin Banjanovic, & Justin Purl (HumRRO)

Task 6: Introduction

HumRRO conducted a review of psychometric processing for the SC READY grade 5 ELA assessment, along with a review of item parameters for (a) all grade levels of SC READY ELA, (b) all grade levels of SC READY Math, (c) EOCEP English 1, (d) EOCEP Biology 1, and (e) EOCEP Algebra 1. The purpose of this task was to satisfy the RFP's request for a specific evaluation of psychometric validity. The review of item parameters addresses the following elements of psychometric validity outlined in the RFP:

- Is the difficulty level of the item appropriate?
- Are the item discrimination statistics acceptable?
- Do the item characteristics support that the items were written in such a way as to reduce the likelihood that a student could get the item correct by guessing?

The review of psychometric processing represents an additional step intended to bolster the rigor of the psychometric validity evaluation by verifying that established psychometric processes and procedures are sound. Because psychometric processes and procedures were similar across all tests, we limited the review to a single grade and subject (SC READY grade 5 ELA). This assessment was identified by the Data Recognition Corporation (DRC) as having demonstrated the most notable changes from 2015-16 to 2016-17, and, thus, was of particular interest to the Education Oversight Committee (EOC) for this review.

Task 6: Method

The method and procedures used for Task 6 are discussed below. First, we discuss the method used to review psychometric processing. Then, we discuss the method used to review item parameters. The methods used were common across all assessments reviewed.

Review of Psychometric Processing

HumRRO first requested several documents and data sources to facilitate this review. These included technical specifications for item calibration and scaling, test maps, student data files, and item parameter files. We used WINSTEPS v.3.91.0 to independently estimate item parameters. We then compared our initially estimated item parameters with those provided by DRC.

After our initial parameter estimation, we followed up with requests for additional information to help us troubleshoot differences between our parameters and those provided by DRC. We also scheduled a telephone conference with key staff from the South Carolina Department of Education (SCDE) and DRC to clarify our understanding of processes and procedures and identify any additional documentation that would be helpful. During this discussion, we clarified our understanding of the data cleaning process and requested more documentation of the vertical and horizontal (e.g., year-to-year) equating processes and procedures. DRC staff agreed to assemble additional documentation for this purpose. DRC also agreed to share several interim WINSTEPS output files to help us fill gaps in the available technical documentation.

The steps in the parameter estimation process that we attempted to replicate, along with a rationale for independent replication of each step, are presented below.

1. **Import test map and arrange items in calibration order**. This replication step is key for ensuring that student response data are matched with the correct item-level characteristics (e.g., item type, possible score points). Small differences in parameter estimation have been observed when items are entered into WINSTEPS in a different order, so it is also important to ensure that independent calibrations enter items in the same order.

2. **Import student data, clean student data, and score items**. This replication step is key for ensuring that (a) all student records are properly read in, (b) certain records are removed from the calibration sample per previously established exclusion rules (e.g., accommodated student, duplicate records), and (c) items are correctly scored. Any differences in the above steps can result in parameter differences. Additionally, it is important to ensure that student data are sorted in the same way prior to entering WINSTEPS as different ordering can also lead to small differences in parameter estimation.

3. **Conduct free calibration of operational items**. This replication step estimates the initial item parameters, prior to any equating or linking. As with all other WINSTEPS calibrations, these estimates are impacted by WINSTEPS settings (e.g., correcting for statistical estimation bias or extreme scores), so it is important that independent calibrations use the same settings. The text-dependent item was not included in the initial calibration to minimize the influence of error associated with interscorer differences.

4. **Put grade 4 vertical linking items (administered to grade 5 students) on operational grade 5 scale**. The final, banked item parameters will reflect adjustments to the initial item parameters resulting from the linking process. This is the first step in computing those adjustments, and includes evaluating item drift to ensure that only appropriate anchor items are included in linking. Replicating this process is essential for ensuring the overall quality of the linking process.

5. **Compute vertical linking constant**. This replication step ensures that the calculated adjustments to the initial item parameters are correct and the appropriate item parameters have been used in the estimation.

6. **Put operational items on the 2017 vertical scale**- This replication step ensures that the computed vertical linking constant is correctly applied to all item parameters.

7. **Calibrate the text-dependent item on the vertical scale**. This replication step ensures that the estimation of the item parameters for the text-dependent item is correct.

8. **Compute horizontal linking constant**. This replication step ensures that adjustments made to the cut scores as a result of the horizontal linking process are correct.

Review of Item Parameters

HumRRO received item-level data files for each assessment reviewed. For EOCEP exams, operational items from the fall/winter 2016-17 and spring 2017 administrations were included in our evaluation. For SC READY, operational items from the 2016-17 assessment were included.

For each item, indexes of classical test theory (CTT)—item difficulty (p-values) and item discrimination (item-total correlation) were provided. For multiple-choice items, the percentage of students selecting each response option and point-biserial correlations were also provided. We first calculated the distribution of CTT item difficulty and discrimination statistics for each item type. Next, we flagged items with CTT item difficulty and discrimination statistics that failed to fall within an acceptable range of values (i.e., $p < .10$, $p > .95$, and item-total correlation < .10). Additionally, we flagged multiple-choice items with potentially problematic distractors. These flags identified items in which a distractor was chosen more often than the correct answer or a distractor had a point-biserial correlation higher than the correct response option. All of the above flags were based on work HumRRO has done previously for another assessment program and were selected because they reflect more stringent criteria than the key check criteria provided by DRC. While DRC has documented key check criteria, the DRC criteria were not employed in the current review as they are intended to identify items for potential mis-key issues, not items that may not belong in the item bank. Finally, the number of items flagged for differential item functioning (DIF) among gender, ethnicity, and test mode (SC READY only) subgroups was analyzed.

For the EOCEP spring 2017 assessments and the SC READY 2016-17 assessments, the availability of two Rasch fit statistics allowed us to conduct an additional examination. The standardized infit is an information-weighted fit statistic that is sensitive to unexpected item performance near a person's measurement level. The standardized outfit is an outlier-sensitive fit statistic that is sensitive to unexpected item performance far from a person's measurement level. Using a criterion that HumRRO has applied on behalf of another testing program, we flagged items with absolute standardized fit values greater than 3.6 for potential fit issues. This is a more stringent criterion than what is typically applied, thus flagging items with statistically significant misfit statistics at the <.001 level applied (see Linacre, 2014).

Task 6: Results

Review of Psychometric Processing

A robust replication of psychometric processing would yield independently estimated item parameters and independently calculated linking constants that match to the fourth decimal place, when provided with the (a) student data file, (b) various item- and test-level input files (e.g., test maps, anchor parameters), and (c) detailed documentation summarizing the process steps (including methods used, decision criteria applied, etc.). Through our review of available and requested documentation we could follow the logic of item calibration and scaling processes and procedures, but we were not able to perform a robust, independent replication.

There were several factors that influenced the level of independence in our replication and the degree to which we could replicate the final item parameters of record. Appendix H presents a detailed description of the circumstances that contributed to a review of the logic of item calibration and scaling processes, as opposed to a robust, independent replication down to the fourth decimal place.

Most notable was the amount of, and level of detail in, the available documentation describing processes and procedures. This was likely due, in part, to a lack of common understanding between HumRRO and DRC staff about what the replication task entailed and the files that were needed. HumRRO conducts psychometric replication for numerous testing programs and has replicated DRC's processes and procedures in the past for other state assessment systems. Because of this, we incorrectly assumed that the same type and specificity of documentation would be readily available for South Carolina. However, because DRC's contract with South Carolina does not include independent replication by a third party, this type of detailed documentation was not available. Consequently, to fill the gaps in our understanding we had to rely on DRC's WINSTEPS control files and interim output files rather than producing these wholly independently.

There was one unexpected change in the way that DRC generates student data files that also impacted our replication efforts. DRC has moved to a system that performs data cleaning at the time of extraction. Therefore, it was not possible for us to get an "uncleaned" data file with which to replicate data cleaning rules. Rather, we relied on DRC to provide us with the input data files used in their WINSTEPS runs. During the process of requesting the student data file that we would need to use to replicate their results, DRC discovered duplicate student records in the datafile used for their initial grade 5 ELA calibration for 2017. As a result, DRC re-estimated parameters using a corrected datafile and concluded that the estimation was invariant. To replicate their process and match the WINSTEPS output originally delivered, HumRRO proceeded by using the datafile with duplicate students.

Table 6.1 summarizes the results of HumRRO's replication. We were able to match some of the initial parameter estimates to the fourth decimal place, but did observe some differences at the third decimal place. The largest parameter difference (0.0233) was observed in the calibration of the text-dependent item onto the vertical scale. This was not a concurrent replication, and so resolving these differences is beyond the scope of this task. It is likely that differences in the initial calibration were due to differences in the WINSTEPS versions used by DRC and HumRRO. These small initial differences would then be compounded in subsequent steps.

Review of Item Parameters

The results from the review of item parameters are reported separately for each assessment. SC READY results are reported first, followed by EOCEP results.

Table 6.1 Summary of Replication Results for ELA Grade 5

Points of Comparison	Level of Match
Free calibration of 2017 operational (without TDA item)	Absolute parameter difference: • Average = 0.0003 • Max = 0.0028 • Min = 0.0000
Calibration of 18 grade 4 linking items on operational grade 5 scale	Absolute parameter difference: • Average = 0.0061 • Max = 0.0146 • Min = 0.0001
Vertical linking constant	• Documentation was specifically created for this task by DRC to replicate the decision criteria for item removal from the linking set via Robust Z and the linking constant estimation. • Matched the DRC constant to the fourth decimal place (i.e., 0.0001).
Calibration of operational items on vertical scale	Using DRC's vertical linking constant, absolute parameter difference: • Average = 0.0003 • Max = 0.0028 • Min = 0.0000
Calibration of TDA item parameter on the vertical scale	Absolute parameter difference across difficulty and step parameters for the TDA item: • Average = 0.0060 • Max = 0.0233 • Min = 0.0005
Raw to theta score table	Absolute theta differences: • Average = 0.0040 • Max = 0.0102 • Min = 0.0000
Horizontal linking constant	• Documentation was specifically created for this task by DRC to replicate the decision criteria for item removal from the linking set via Robust Z and the linking constant estimation. The additional documentation did not review the Wright and Bell linking method (an alternative explored), thus that could not be replicated. • Matched the unweighted link constant estimated from all the items and the Robust Z estimates (e.g., correlation between parameters, standard deviation ratio, Robust Z values). This was the final linking constant employed.

Note. TDA = text-dependent analysis.

SC READY ELA

Table 6.2 presents a summary of CTT item difficulty statistics for operational 2016-17 SC READY ELA items. Items with p-values greater than .95 indicates that these items were very easy for this group of examinees, while items with p-values less than .10 indicate that these items were very difficult for this group of examinees. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so ideally item p-values will fall between these values. As Table 6.2 shows, no item was flagged at any tested grade level for p-values falling outside of this acceptable range.

Table 6.2 also presents analysis of multiple-choice item distractors that further inform item difficulty. Items with one or more appealing distractors may have lower p-values. At each of the tested grade levels, fewer than 7% of multiple-choice items were flagged for having a distractor that was selected more often than the correct response.

Table 6.2 Item Difficulty Analysis: SC READY ELA

Grade	Item Type	Item p-values					Item Difficulty Flags % (N)		Distractor Flags % (N)
		N	Min	Max	Mean	SD	p-value above .95	p-value below .10	Distractor Selected More Often than Correct Response
3	MC	59	.276	.817	.552	.142	0	0	6.78 (4)
3	EB	4	.180	.428	.293	.102	0	0	NA
3	MS	4	.192	.481	.313	.126	0	0	
3	TE	3	.346	.757	.511	.217	0	0	
3	TDA	1	.295	.295	.295	--	0	0	
4	MC	64	.315	.872	.615	.143	0	0	3.13 (2)
4	EB	1	.475	.475	.475	--	0	0	NA
4	MS	3	.277	.443	.362	.083	0	0	
4	TDA	1	.258	.258	.258	--	0	0	
5	MC	63	.319	.858	.591	.132	0	0	3.17 (2)
5	EB	2	.453	.574	.514	.086	0	0	NA
5	MS	3	.241	.459	.371	.115	0	0	
5	TDA	1	.292	.292	.292	--	0	0	
6	MC	67	.187	.799	.575	.125	0	0	1.49 (1)
6	EB	8	.190	.631	.420	.144	0	0	NA
6	MS	5	.192	.444	.362	.104	0	0	
6	TDA	1	.249	.249	.249	--	0	0	
7	MC	69	.296	.805	.566	.119	0	0	2.90 (2)
7	EB	5	.344	.537	.459	.083	0	0	NA
7	MS	6	.220	.488	.359	.119	0	0	
7	TDA	1	.324	.324	.324	--	0	0	
8	MC	69	.323	.859	.608	.131	0	0	1.45 (1)
8	EB	7	.106	.644	.414	.182	0	0	NA
8	MS	4	.186	.535	.355	.165	0	0	
8	TDA	1	.438	.438	.438	--	0	0	

Note. MC= Multiple choice; EB= Evidence-based; MS= Multiple select; TE= Technology enhanced; TDA= Text-dependent analysis; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.3 presents a summary of CTT item discrimination statistics for operational 2016-17 SC READY ELA items. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in ELA. As Table 6.3 shows, one item in grade 6 was flagged for a low item-total correlation.

Table 6.3 also presents analysis of the correlation between multiple-choice item distractors and total test score. A distractor-total correlation that is higher than the key-total correlation would indicate that higher ability students are selecting the distractor more frequently than the correct response. At each of the tested grade levels, less than 4% of SC READY ELA multiple-choice items were flagged for having a distractor-total correlation higher than the key-total correlation.

Table 6.3 Item Discrimination Analysis: SC READY ELA

Grade	Item Type	Item-Total Correlations					Item Discrimination Flags % (N)	Distractor Flags % (N)
		N	Min	Max	Mean	SD	Item-total correlation below .10	MC distractor-total correlation higher than key-total correlation
3	MC	59	.143	.553	.376	.099	0	3.39 (2)
3	EB	4	.434	.585	.488	.068	0	NA
3	MS	4	.204	.592	.416	.175	0	
3	TE	3	.250	.326	.276	.043	0	
3	TDA	1	.664	.664	.664	--	0	
4	MC	64	.141	.576	.400	.102	0	1.56 (1)
4	EB	1	.213	.213	.213	--	0	NA
4	MS	3	.401	.611	.478	.116	0	
4	TDA	1	.523	.523	.523	--	0	
5	MC	63	.144	.580	.401	.090	0	1.59 (1)
5	EB	2	.614	.634	.624	.014	0	NA
5	MS	3	.366	.681	.507	.160	0	
5	TDA	1	.526	.526	.526	--	0	
6	MC	67	.000	.569	.410	.101	1.49 (1)	2.99 (2)
6	EB	8	.172	.671	.499	.153	0	NA
6	MS	5	.378	.563	.468	.067	0	
6	TDA	1	.546	.546	.546	--	0	
7	MC	69	.148	.564	.399	.091	0	2.90 (2)
7	EB	5	.447	.592	.527	.052	0	NA
7	MS	6	.120	.598	.413	.171	0	
7	TDA	1	.663	.663	.663	--	0	
8	MC	69	.134	.606	.406	.112	0	2.90 (2)
8	EB	7	.116	.680	.467	.201	0	NA
8	MS	4	.386	.685	.528	.134	0	
8	TDA	1	.687	.687	.687	--	0	

Note. MC= Multiple choice; EB= Evidence-based; MS= Multiple select; TE= Technology enhanced; TDA= Text-dependent analysis; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.4 presents a summary of results from differential item functioning (DIF) analyses for operational 2016-17 SC READY ELA items. DIF statistics provide an indication of whether items are functioning differently for different student groups, after taking into account underlying ability. DRC calculated the Mantel-Haenszel (MH) chi-square for dichotomous items and then identified items that demonstrated large DIF. As Table 6.4 demonstrates, no more than three items were flagged for DIF at any grade level. No items were flagged for gender DIF. Most flags were for items exhibiting DIF between black and white students. One grade 8 multiple-choice item was flagged for mode DIF. The presence of DIF is not sufficient for bias, but rather is a trigger for further scrutiny of an item. The small number of items flagged for DIF indicates that that were no systematic fairness issues with the operational SC READY ELA items.

Table 6.4 Differential Item Functioning (DIF) Analysis: SC READY ELA

Grade	Item Type	N	DIF Flags % (N)		
			Female/Male	Black/White	Online Mode/Paper Mode
3	MC	59	0	0	0
3	EB	4	0	0	0
3	MS	4	0	0	0
3	TE	3	0	0	0
3	TDA	1	0	0	0
4	MC	64	0	1.56 (1)	0
4	EB	1	0	0	0
4	MS	3	0	0	0
4	TDA	1	0	0	0
5	MC	63	0	1.59 (1)	0
5	EB	2	0	0	0
5	MS	3	0	0	0
5	TDA	1	0	0	0
6	MC	67	0	4.48 (3)	0
6	EB	8	0	0	0
6	MS	5	0	0	0
6	TDA	1	0	0	0
7	MC	69	0	0	0
7	EB	5	0	0	0
7	MS	6	0	0	0
7	TDA	1	0	0	0
8	MC	69	0	1.45 (1)	1.45 (1)
8	EB	7	0	14.29 (1)	0
8	MS	4	0	0	0
8	TDA	1	0	0	0

Note. MC= Multiple choice; EB= Evidence-based; MS= Multiple select; TE= Technology enhanced; TDA= Text-dependent analysis; N = number of items.

Table 6.5 summarizes Rasch item statistics from the 2016-17 administration. In grades 3 through 5, no items were flagged for high item difficulty, while between 1 and 7 items per grade were flagged for low item difficulty. Conversely, in grades 6-8, between 2 and 7 items per grade level were flagged for high item difficulty and none were flagged for low item difficulty. Multiple choice items were more frequently flagged than other item types.

The last column in Table 6.5 presents the number of items at each grade level flagged for not passing infit and outfit tests. This statistic indicates that these items demonstrated student response patterns that were not as expected given the item difficulty. Only two grade 6 and three grade 8 items were flagged. Overall, the available Rasch item statistics indicate that the 2016-17 operational SC READY items measured student achievement in ELA at appropriate levels of difficulty, and that items functioned as intended.

Table 6.5 Rasch Item Statistics: SC READY ELA

Grade	Item Type	Rasch Empirical Items Difficulty							Item Fit
		N	Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)	Not passing infit/outfit tests[a] % (N)
3	MC	59	-2.480	0.456	-0.982	0.736	0	10.17 (6)	0
3	EB	4	-0.362	1.104	0.400	0.600	0	0	0
3	MS	4	-0.632	1.059	0.302	0.728	0	0	0
3	TE	3	-2.083	0.029	-0.818	1.116	0	33.33 (1)	0
3	TDA	1	0.614	0.614	0.614	--	0	0	0
4	MC	64	-2.532	0.728	-0.814	0.788	0	6.25 (4)	0
4	EB	1	-0.099	-0.099	-0.099	--	0	0	0
4	MS	3	0.086	1.002	0.527	0.459	0	0	0
4	TDA	1	1.372	1.372	1.372	--	0	0	0
5	MC	63	-2.009	1.176	-0.270	0.721	0	1.59 (1)	0
5	EB	2	-0.134	0.425	0.146	0.396	0	0	0
5	MS	3	0.427	1.682	0.934	0.662	0	0	0
5	TDA	1	1.646	1.646	1.646	--	0	0	0
6	MC	67	-1.293	2.284	0.049	0.684	1.49 (1)	0	2.99 (2)
6	EB	8	-0.237	2.252	0.887	0.795	12.50 (1)	0	0
6	MS	5	0.736	2.248	1.204	0.624	20.00 (1)	0	0
6	TDA	1	2.199	2.199	2.199	--	100.00 (1)	0	0
7	MC	69	-0.862	1.900	0.471	0.633	0	0	0
7	EB	5	0.619	1.578	1.016	0.414	0	0	0
7	MS	6	0.888	2.372	1.590	0.656	33.33 (2)	0	0
7	TDA	1	1.898	1.898	1.898	--	0	0	0
8	MC	69	-1.006	2.111	0.582	0.720	4.35 (3)	0	2.90 (2)
8	EB	7	0.412	3.802	1.715	1.120	28.57 (2)	0	14.29 (1)
8	MS	4	0.980	2.989	1.982	0.936	50.00 (2)	0	0
8	TDA	1	1.562	1.562	1.562	--	0	0	0

Note. MC= Multiple choice; EB= Evidence-based; MS= Multiple select; TE= Technology enhanced; TDA= Text-dependent analysis; N = number of items; Min = Minimum; Max = Maximum.

[a] Infit/outfit tests based on criteria of >|3.6|.

SC READY Math

Table 6.6 presents a summary of CTT item difficulty statistics for the operational 2016-17 SC READY math items. Items with p-values greater than .95 were very easy for this group of examinees, while items with p-values less than .10 were very difficult. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so ideally item p-values will fall between these values. As Table 6.6 shows, no items were flagged for low difficulty, and one multiple-select item each at grades 5 and 8 were flagged for high difficulty.

Table 6.6 also presents analysis of multiple-choice item distractors that further inform item difficulty. Items with one or more appealing distractors may have lower p-values. At each of the tested grade levels, fewer than 13% of multiple-choice items were flagged for having a distractor that was selected more often than the correct response.

Table 6.6 Item Difficulty Analysis: SC READY Math

Grade	Item Type	Item p-values					Item Difficulty Flags-% (N)		Distractor Flags % (N)
		N	Min	Max	Mean	SD	P value above .95	P value below .10	Distractor Selected more often than Correct Response
3	MC	50	.222	.844	.609	.156	0	0	6.00 (3)
4	MC	56	.303	.880	.549	.132	0	0	3.57 (2)
5	MC	54	.275	.832	.539	.150	0	0	12.96 (7)
5	MS	2	.097	.354	.226	.182	0	50.00 (1)	NA
6	MC	54	.285	.894	.575	.134	0	0	3.70 (2)
6	SR	3	.529	.711	.640	.097	0	0	NA
6	MS	3	.404	.504	.455	.050	0	0	
6	TE	1	.353	.353	.353	--	0	0	
7	MC	54	.249	.839	.505	.142	0	0	11.11 (6)
7	SR	3	.360	.571	.488	.112	0	0	NA
7	MS	3	.153	.349	.277	.108	0	0	
7	TE	1	.145	.145	.145	.	0	0	
8	MC	55	.309	.817	.528	.135	0	0	5.45 (3)
8	SR	3	.154	.294	.227	.070	0	0	NA
8	MS	4	.070	.283	.192	.089	0	25.00 (1)	
8	TE	3	.324	.401	.366	.039	0	0	

Note. MC= Multiple choice; SR= Short Response; MS= Multiple select; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.7 presents a summary of CTT item discrimination statistics for operational 2016-17 SC READY math items. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in mathematics. As Table 6.7 shows, one grade 4 and one grade 5 item were flagged for a low item-total correlation.

Table 6.7 Item Discrimination Analysis: SC READY Math

Grade	Item Type	Item-Total Correlations					Item Discrimination Flags % (N)	Distractor Flags % (N)
		N	Min	Max	Mean	SD	Item-total correlation below .10	MC distractor-total correlation higher than key-total correlation
3	MC	50	.152	.607	.431	.107	0	2.00 (1)
4	MC	56	.094	.639	.413	.101	1.79 (1)	1.79 (1)
5	MC	54	.081	.603	.421	.097	1.85 (1)	1.85 (1)
5	MS	2	.358	.583	.471	.159	0	NA
6	MC	54	.193	.574	.412	.087	0	1.85 (1)
6	SR	3	.510	.543	.527	.017	0	NA
6	MS	3	.603	.620	.614	.009	0	
6	TE	1	.453	.453	.453	--	0	
7	MC	54	.133	.581	.385	.109	0	5.56 (3)
7	SR	3	.581	.606	.597	.014	0	NA
7	MS	3	.320	.519	.445	.109	0	
7	TE	1	.414	.414	.414	--	0	
8	MC	55	.173	.522	.390	.091	0	3.64 (2)
8	SR	3	.491	.589	.542	.049	0	NA
8	MS	4	.390	.462	.425	.037	0	
8	TE	3	.350	.570	.468	.111	0	

Note. MC= Multiple choice; SR= Short response; MS= Multiple select; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation.

Table 6.7 also presents analysis of the correlation between multiple-choice item distractors and total test score. A distractor-total correlation that is higher than the key-total correlation would indicate that higher ability students are selecting the distractor more frequently than the correct response. At each of the tested grade levels, less than 6% of SC READY math multiple-choice items were flagged for having a distractor-total correlation higher than the key-total correlation.

Table 6.8 presents a summary of results from DIF analyses for operational 2016-17 SC READY math items.

As Table 6.8 demonstrates, no more than four items were flagged for DIF at any grade level. No items were flagged for gender DIF. Most flags were for items exhibiting DIF between black and white students. One multiple-choice item each from grades 3, 6, and 7 was flagged for mode DIF. The presence of DIF is not sufficient for bias, but rather is a trigger for further scrutiny of an item. The small number of items flagged for DIF indicates that that were no systematic fairness issues with the operational SC READY math items.

Table 6.8 Differential Item Functioning (DIF) Analysis: SC READY Math

Grade	Item Type	N	DIF Flags % (N)		
			Female/Male	Black/White	Online Mode/Paper Mode
3	MC	50	0	2.00 (1)	2.00 (1)
4	MC	56	0	7.14 (4)	0
5	MC	54	0	3.70 (2)	0
5	MS	2	0	0	0
6	MC	54	0	1.85 (1)	1.85 (1)
6	SR	3	0	0	0
6	MS	3	0	0	0
6	TE	1	0	0	0
7	MC	54	0	3.70 (2)	1.85 (1)
7	SR	3	0	0	0
7	MS	3	0	33.33 (1)	0
7	TE	1	0	0	0
8	MC	55	0	1.82 (1)	0
8	SR	3	0	0	0
8	MS	4	0	0	0
8	TE	3	0	0	0

Note. MC= Multiple choice; SR= Short response; MS= Multiple select; TE= Technology enhanced; N = number of items.

Table 6.9 summarizes Rasch item statistics from the 2016-17 administration. In grades 3, 4, and 6 no items were flagged for high levels of difficulty. One grade 5 multiple-select item, one grade 7 multiple-select, one grade 7 technology-enhanced item, three grade 8 multiple-select, and four grade 8 short response items were flagged for high difficulty. The multiple-select and technology-enhanced items were disproportionately flagged for high levels of difficulty compared to multiple-choice and short response items. No items from grades 5, 7, and 8 were flagged for low levels of difficulty. Nine grade 3 items, three grade 4 items, and one grade 6 item were lagged for low levels of difficulty.

The last column in Table 6.9 presents the number of items at each grade level flagged for not passing infit and outfit tests. This statistic indicates that these items demonstrated student response patterns that were not as expected given the item difficulty. Three grade 3 items and one item each from grades 4, 5, and 7 were flagged. Overall, the available Rasch item statistics indicate that the 2016-17 operational SC READY items measured student achievement in mathematics at appropriate levels of difficulty, and that items functioned as intended.

Table 6.9 Rasch Item Statistics: SC READY Math

Grade	Item Type	Rasch Empirical Items Difficulty							Item Fit
		N	Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)	Not passing infit/outfit tests[a] % (N)
3	MC	50	-2.752	1.101	-1.185	0.901	0	18.00 (9)	6.00 (3)
4	MC	56	-2.700	0.820	-0.529	0.736	0	5.36 (3)	1.79 (1)
5	MC	54	-1.764	1.499	-0.020	0.835	0	0	1.85 (1)
5	MS	2	1.061	3.089	2.075	1.434	50.00 (1)	0	0
6	MC	54	-2.166	1.528	-0.090	0.753	0	1.85 (1)	0
6	SR	3	-0.843	0.197	-0.441	0.559	0	0	0
6	MS	3	0.281	0.830	0.551	0.274	0	0	0
6	TE	1	1.172	1.172	1.172	--	0	0	0
7	MC	54	-1.445	1.912	0.476	0.756	0	0	1.85 (1)
7	SR	3	0.118	1.180	0.541	0.563	0	0	0
7	MS	3	1.252	2.593	1.744	0.738	33.33 (1)	0	0
7	TE	1	2.736	2.736	2.736	--	100.00 (1)	0	0
8	MC	55	-0.890	1.949	0.759	0.713	0	0	0
8	SR	3	2.021	3.017	2.480	0.503	100.00 (3)	0	0
8	MS	4	2.096	4.073	2.826	0.864	100.00 (4)	0	0
8	TE	3	1.442	1.886	1.643	0.225	0	0	0

Note. MC= Multiple choice; SR= Short response; MS= Multiple select; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation.

[a] Infit/outfit tests based on criteria of >|3.6|.

English 1 EOCEP

Table 6.10 presents a summary of CTT item difficulty statistics for operational English 1 items from the fall/winter 2016-17 and spring 2017 administrations. Items with p-values greater than .95 were very easy for this group of examinees, while items with p-values less than .10 were very difficult. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so ideally item p-values will fall between these values. As Table 6.10 shows, only one English 1 item was flagged for p-values falling outside of this acceptable range. A closer look indicates that this item's p-value was just outside of the acceptable range (p-value = .951).

Table 6.10 Item Difficulty Analysis: English 1 (fall/winter 2016-17 and spring 2017)

Item Type	Item p-values					Item Difficulty Flags % (N)		Distractor Flags % (N)
	N	Min	Max	Mean	SD	P value above .95	P value below .10	Distractor Selected more often than Correct Response
MC	106	.283	.951	.630	.130	1.00 (1)	0	0.94 (1)
EB	4	.465	.668	.592	.093	0	0	NA
TE	4	.445	.724	.586	.117	0	0	NA

Note. MC= Multiple choice; EB = Evidence based; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.10 also presents analysis of multiple-choice item distractors that further inform item difficulty. Items with one or more appealing distractors may have lower p-values. Slightly less than one percent (0.94%) of multiple-choice items were flagged for having a distractor that was selected more often than the correct response.

Table 6.11 presents a summary of CTT item discrimination statistics for operational English 1 items from the fall/winter 2016-17 and spring 2017 administrations. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in English 1. As Table 6.11 shows, no items were flagged for low item-total correlations.

Table 6.11 also presents analysis of the correlation between multiple-choice item distractors and total test score. A distractor-total correlation that is higher than the key-total correlation would indicate that higher ability students are selecting the distractor more frequently than the correct response. Only about 3% of English 1 multiple-choice items were flagged for having a distractor-total correlation higher than the key-total correlation.

Table 6.11 Item Discrimination Analysis: English 1 (fall/winter 2016-17 and spring 2017)

Item Type	Item-Total Correlations					Item Discrimination Flags % (N)	Distractor Flags % (N)
	N	Min	Max	Mean	SD	Item-total correlation below .10	MC distractor-total correlation higher than key-total correlation
MC	106	.117	.588	.377	.091	0	2.83 (3)
EB	4	.339	.619	.514	.121	0	NA
TE	4	.391	.600	.462	.095	0	NA

Note. MC= Multiple choice; EB = Evidence based; TE= Technology enhanced N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.12 presents a summary of results from DIF analyses for operational English 1 EOCEP items from the fall/winter 2016-17 and spring 2017 administrations. As Table 6.12 demonstrates, only three items were flagged for DIF. Both were flagged for differences between Black and White student groups. The presence of DIF is not sufficient for bias, but rather is a trigger for further scrutiny of an item. The small number of items flagged for DIF indicates that that were no systematic fairness issues with the operational English 1 EOCEP items.

Table 6.12 Differential Item Functioning (DIF) Analysis: English 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	DIF Flags % (N)	
		Female/Male	Black/White
MC	106	0	2.83 (3)
EB	4	0	0
TE	4	0	0

Note. MC= Multiple choice. TE= Technology enhanced.

Table 6.13 summarizes Rasch item statistics from the spring 2017 administration. No items were flagged for item difficulty that fell outside of the acceptable range. Only two multiple-choice items were flagged for not passing infit and outfit tests. This indicates that these items demonstrated student response patterns that were not as expected given the item difficulty. Overall, the available Rasch item statistics indicate that spring 2017 operational English 1

EOCEP items measured student achievement in English 1 at appropriate levels of difficulty, and that items functioned as intended.

Table 6.13 Rasch Item Statistics: English 1 (spring 2017)

		Rasch Empirical Items Difficulty						Item Fit
Item Type	N	Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)	Not passing infit/outfit tests[a] % (N)
MC	53	-1.521	1.379	0.186	0.750	0	0	3.77 (2)
EB	2	0.511	0.512	0.512	0.001	0	0	0
TE	2	-0.3756	0.617	0.121	0.702	0	0	0

Note. MC= Multiple choice; EB = Evidence based; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation.
[a] Infit/outfit tests based on criteria of >|3.6|.

Biology 1 EOCEP

Table 6.14 presents a summary of CTT item difficulty statistics for operational Biology 1 items from the fall/winter 2016-17 and spring 2017 administrations. Items with *p*-values greater than .95 were very easy for this group of examinees, while items with *p*-values less than .10 were very difficult. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so ideally item *p*-values will fall between these values. As Table 6.14 shows, no Biology 1 items were flagged for *p*-values falling outside of this acceptable range.

Table 6.14 also presents analysis of multiple-choice item distractors that further inform item difficulty. Items with one or more appealing distractors may have lower *p*-values. Slightly less than one percent (0.83%) of multiple-choice items were flagged for having a distractor that was selected more often than the correct response.

Table 6.14 Item Difficulty Analysis: Biology 1 (fall/winter 2016-17 and spring 2017)

		Item p-values				Item Difficulty Flags- % (N)		Distractor Flags % (N)
Item Type	N	Min	Max	Mean	SD	P value above .95	P value below .10	Distractor Selected more often than Correct Response
MC	120	.353	.871	.587	.114	0	0	0.83 (1)
MS	1	.647	.647	.647	--	0	0	NA
TE	9	.330	.715	.538	.120	0	0	NA

Note. MC= Multiple choice; MS = Multiple select; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation.

Table 6.15 presents a summary of CTT item discrimination statistics for operational Biology 1 EOCEP items from the fall/winter 2016-17 and spring 2017 administrations. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in Biology 1. As Table 6.15 shows, only two items were flagged for low item-total correlations, one of which was a negative correlation. Negative item-total correlations are of concern because they negatively impact score reliability. The item-total correlations for both flagged items were based on a very small number of students, and so these item statistics should be interpreted with caution.

Table 6.15 also presents analysis of the correlation between multiple-choice item distractors and total test score. A distractor-total correlation that is higher than the key-total correlation would indicate that higher ability students are selecting the distractor more frequently than the correct response. Approximately 4% of Biology 1 multiple-choice items were flagged for having a distractor-total correlation higher than the key-total correlation.

Table 6.15 Item Discrimination Analysis: Biology 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	Item-Total Correlations				Item Discrimination Flags % (N)	Distractor Flags % (N)
		Min	Max	Mean	SD	Item-total correlation below .10	MC distractor-total correlation higher than key-total correlation
MC	120	-.249	.804	.389	.115	1.67 (2)	4.17 (5)
MS	1	.551	.551	.551	--	0	NA
TE	9	.258	.526	.413	.096	0	NA

Note. MC= Multiple choice; MS = Multiple select; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation.

Table 6.16 presents a summary of results from DIF analyses for operational Biology 1 items from the fall/winter 2016-17 and spring 2017 administrations. DIF statistics provide an indication of whether items are functioning differently for different student groups, after taking into account underlying ability. DRC calculated the Mantel-Haenszel (MH) chi-square for dichotomous items and then identified items that demonstrated large DIF. As Table 6.16 demonstrates, only two items were flagged for DIF. Both were flagged for differences between Black and White student groups. The presence of DIF is not sufficient for bias, but rather is a trigger for further scrutiny of an item. The small number of items flagged for DIF indicates that that were no systematic fairness issues with the operational Biology 1 items.

Table 6.16 Differential Item Functioning (DIF) Analysis: Biology 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	DIF Flags % (N)	
		Female/Male	Black/White
MC	120	0	0.83 (1)
MS	1	0	0
TE	9	0	11.11 (1)

Note. MC= Multiple choice; MS = Multiple select; TE= Technology enhanced; N = number of items.

Table 6.17 summarizes Rasch item statistics from the spring 2017 administration. No items were flagged for item difficulty that fell outside of the acceptable range. Five multiple-choice items were flagged for not passing infit and outfit tests. This indicates that these items demonstrated student response patterns that were not as expected given the item difficulty. Overall, the available Rasch item statistics indicate that spring 2017 operational Biology 1 items measured student achievement in Biology 1 at appropriate levels of difficulty, and that items generally functioned as intended.

Table 6.17 Rasch Item Statistics: Biology 1 (spring 2017)

Item Type	N	Rasch Empirical Items Difficulty						Item Fit
		Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)	Not passing infit/outfit tests[a] % (N)
MC	60	-1.123	1.023	0.055	0.564	0	0	8.33 (5)
MS	0	NA	NA	NA	NA	NA	NA	NA
TE	5	-0.666	1.004	0.222	0.605	0	0	0

Note. MC= Multiple choice; MS = Multiple select; TE= Technology enhanced.
[a] Infit/outfit tests based on criteria of >|3.6|.

EOCEP Algebra 1

Table 6.18 presents a summary of CTT item difficulty statistics for operational Algebra 1 items from the fall/winter 2016-17 and spring 2017 administrations. Items with p-values greater than .95 were very easy for this group of examinees, while items with p-values less than .10 were very difficult. Items that are very easy or very difficult contribute little information to our understanding of student achievement, and so, ideally, item p-values will fall between these values. As Table 6.18 shows, only one Algebra 1 item was flagged for being too difficult. A closer look at this item showed that its CTT item statistics are based on only six students; it should be noted that very few students complete the paper assessments, particularly for the EOCEP assessments The CTT item difficulty for this item, which informed its placement on the operational test form, fell in the acceptable range.

Table 6.18 also presents analysis of multiple-choice item distractors that further inform item difficulty. Items with one or more appealing distractors may have lower p-values. Nine percent (9%) of multiple-choice items were flagged for having a distractor that was selected more often than the correct response. However, taken into consideration with the low number of item difficulty flags and the overall distribution of p-values that include a range of values, it does not appear that the Algebra 1 distractors are of concern.

Table 6.18 Item Difficulty Analysis: Algebra 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	Item p-values				Item Difficulty Flags % (N)		Distractor Flags % (N)
		Min	Max	Mean	SD	P value above .95	P value below .10	Distractor Selected more often than Correct Response
MC	100	.000	.890	.548	.164	0	1.00 (1)	9.00 (9)
TE	4	.187	.296	.232	.046	0	0	NA

Note. MC= Multiple choice. TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.19 presents a summary of CTT item discrimination statistics for operational Algebra 1 items from the fall/winter 2016-17 and spring 2017 administrations. Items with item-total correlations less than .10 do not help differentiate between students who are low performing and students who are high performing in Algebra 1. As Table 6.19 shows, only two items were flagged for low item-total correlations. One of these items was the same item that demonstrated very high difficulty among a very small sample of students taking the paper form.

Table 6.19 also presents analysis of the correlation between multiple-choice item distractors and total test score. A distractor-total correlation that is higher than the key-total correlation would indicate that higher ability students are selecting the distractor more frequently than the correct response. Only 3% of Algebra 1 multiple-choice items were flagged for having a distractor-total correlation higher than the key-total correlation.

Table 6.19 Item Discrimination Analysis: Algebra 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	Item-Total Correlations				Item Discrimination Flags % (N)	Distractor Flags % (N)
		Min	Max	Mean	SD	Item-total correlation below .10	MC distractor-total correlation higher than key-total correlation
MC	100	.000	.728	.332	.108	2.00 (2)	3.00 (3)
TE	4	.318	.517	.453	.092	0	NA

Note. MC= Multiple choice; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.

Table 6.20 presents a summary of results from DIF analyses for operational Algebra 1 items from the fall/winter 2016-17 and spring 2017 administrations. DIF statistics provide an indication of whether items are functioning differently for different student groups, after taking into account underlying ability. DRC calculated the Mantel-Haenszel (MH) chi-square for dichotomous items and then identified items that demonstrated large DIF. As Table 6.20 demonstrates, only two items were flagged for DIF. Both were flagged for differences between Black and White student groups. The presence of DIF is not sufficient for bias, but rather is a trigger for further scrutiny of an item. The small number of items flagged for DIF indicates that there were no systematic fairness issues with the operational Algebra 1 items.

Table 6.20 Differential Item Functioning (DIF) Analysis: Algebra 1 (fall/winter 2016-17 and spring 2017)

Item Type	N	DIF Flags % (N)	
		Female/Male	Black/White
MC	100	0	2.00 (2)
TE	4	0	0

Note. MC= Multiple choice; TE= Technology enhanced; N = number of items.

Table 6.21 summarizes Rasch item statistics from the spring 2017 administration. One multiple-choice item was flagged for low item difficulty, and one technology enhanced item was flagged for high item difficulty. Both items were relatively close to the absolute value of 2 criterion. One multiple-choice item was flagged for not passing infit and outfit tests. This indicates that this item demonstrated student response patterns that were not as expected given the item difficulty. Overall, the available Rasch item statistics indicate that spring 2017 operational Algebra 1 items measured student achievement in Algebra 1 at appropriate levels of difficulty, and that items functioned as intended.

Table 6.21 Rasch Item Statistics: Algebra 1 (spring 2017)

		Rasch Item Difficulty						Item Fit
Item Type	N	Min	Max	Mean	SD	Rasch difficulty above 2 % (N)	Rasch difficulty below -2 % (N)	Not passing infit/outfit tests[a] % (N)
MC	50	-2.12	1.59	0.111	0.756	0	1.00 (1)	1.00 (1)
TE	2	1.833	2.011	1.922	0.126	50.00 (1)	0	0

Note. MC= Multiple choice; TE= Technology enhanced; N = number of items; Min = Minimum; Max = Maximum; SD = Standard Deviation; NA = Not applicable.
[a] Infit/outfit tests based on criteria of >|3.6|.

Task 6: Discussion

Review of Psychometric Processing

Experts in the field of psychological testing recognize that independent replication is an integral component of quality control in test analysis, scoring, and reporting (Allalouf, 2007; International Test Commission, 2011). As test scores are increasingly used to inform decision-making, testing programs recognize the importance of ensuring that these scores are free of error that could be introduced during complex data processing and statistical modeling procedures. National and state educational assessment programs (e.g., Partnership for Assessment of Readiness for College and Careers, State of Texas Assessments of Academic Readiness) routinely incorporate independent replication of scaling, equating, and scoring into their psychometric processing activities.

As we emphasized at the beginning of this chapter, fully independent replication requires very detailed documentation. This level of detailed documentation is also helpful should there ever be a change in testing vendor, substantial staffing changes within the current test vendor, or other needs for revisiting earlier processes. However, formal documentation at this level of detail may not be something that testing contractors regularly maintain. Rather, documentation may be less formal and decisions may be made during discussions but without formal documentation of the detailed steps. Per the *Test Standards*, “test documents should provide sufficient detail to permit reviewers and researchers to evaluate important analyses published in the test manual or technical report” (AERA, APA, & NCME, 2014, p. 125). Developing comprehensive, detailed documentation that would allow for an independent party or new testing vendor to replicate psychometric processes exceeds the guidance in the *Test Standards* for documentation. Thus, it is not our intent to suggest that DRC failed to follow best practices. We were able to understand the logic of item calibration and scaling processes and procedures; however, more detailed documentation of step-by-step processes would have facilitated independent replication of results to the fourth decimal place.

The request for the data and documentation required to conduct our review did uncover an internal quality control issue for the testing contractor. Specifically, there was an error during the data cleaning process that resulted in duplicate student records being output into the student data file used to calibrate item parameters. Although DRC concluded that this error did not have any impact on item parameter estimation, it does highlight the benefit of having quality control mechanisms in place during operational psychometric processing.

It is important to note that DRC staff were responsive to our questions during a follow-up teleconference meeting to discuss clarifications around the initial documentation provided, and they were able to quickly pull together additional documentation and clarification to assist our efforts in matching item parameters and linking constants as independently as possible. As mentioned previously, we were able to understand the logic of item calibration and scaling processes and procedures, which is arguably more important than, or at least is a necessary precursor to, being able to exactly replicate results. Our recommendations for improvement are therefore focused on increasing available documentation to facilitate future knowledge transfers and quality control efforts.

We offer three recommendations based on our review:

- **Expand existing internal quality control procedures**: SCDE may want to request expanded internal quality procedures from their testing contractor to minimize the potential for errors during operational psychometric processing. This might include multiple staff members conducting the same analyses concurrently and then comparing at predefined points in the process. We did notice that in documentation on EOCEP item development (*EOCEP Forms Construction Guidelines_101817.pdf*), there was mention of "estimation… duplicated by the SCDE" (DRC, 2017, p. 4), but it is not clear what steps are duplicated and at what stage in the overall process. If some amount of duplicating is in place, consider clearly documenting it and expanding upon it.
- **Incorporate independent third-party replication into established procedures for producing test scores:** SCDE should consider requiring the testing contractor to coordinate with a third-party to independently replicate scaling, equating, and scoring (e.g., the production of scoring tables) to help further ensure accuracy in scores.
- **Expand technical documentation of processes and procedures for test scaling, equating, and scoring:** Regardless of whether third-party replication is adopted, SCDE should consider requesting that DRC combine existing psychometric processing documentation into a single, streamlined technical document. This document should include expanded detail about psychometric processing steps.

Review of Item Parameters

DRC provided item statistics for operational items for the 2016-17 SC READY assessments and for the fall/winter 2016-17 and spring 2017 EOCEP assessments. Our analysis reflects an independent process of flagging items based on the statistics provided for the purposes of detecting patterns that would raise concerns about the psychometric validity of test scores. We do not have, nor did we request, documentation of the final decisions made regarding these items. The remainder of this section discusses separately our findings for each assessment under review.

SC READY ELA

Our review of the item-level data from the 2016-17 administration of the SC READY ELA assessments indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer.

Analysis of Rasch IRT statistics did reveal a pattern in which non-traditional item types (e.g., multiple-select, evidence-based) at the middle school level had more items flagged for difficulty parameters that fell outside of the ideal range. Based on this finding, we offer the following recommendation:

- **Examine item content and format of SC READY ELA non-traditional item types at the middle school grade levels:** DRC should consider taking a closer look at items flagged for high levels of difficulty to determine if there were any characteristics of these items that may have influenced student responses. At minimum, further scrutiny of these items could inform subsequent item development activities.

SC READY Math

Our review of the item-level data from the 2016-17 administration of the SC READY math assessments indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer.

Analysis of Rasch IRT statistics did reveal a pattern in which non-traditional item types (e.g., multiple-select, technology enhanced) were more frequently flagged for difficulty parameters that fell outside of the ideal range. Based on this finding, we offer the following recommendation:

- **Examine item content and format of SC READY math non-traditional item types:** DRC should consider taking a closer look at items flagged for high levels of difficulty to determine if there were any characteristics of these items that may have influenced student responses. At minimum, further scrutiny of these items could inform subsequent item development activities.

English 1 EOCEP

Our review of the item-level data from the fall/winter 2016-17 and spring 2017 administrations of the English 1 assessment indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer. We have no recommendations for improving English 1 based on the results of this Task.

Biology 1 EOCEP

Our review of the item-level data from the fall/winter 2016-17 and spring 2017 administrations of the Biology 1 assessment indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer. We have no recommendations for improving Biology 1 based on the results of this Task.

Algebra 1 EOCEP

Our review of the item-level data from the fall/winter 2016-17 and spring 2017 administrations of the Algebra 1 assessment indicate that overall, items (a) are appropriately difficult, (b) discriminate among student ability levels, and (c) were not written in such a way as to enable students to easily guess the correct answer. We have no recommendations for improving Algebra 1 based on the results of this Task.

Summary and Conclusions from Part I

Andrea Sinclair (HumRRO)

The technical evaluation of the SC READY and EOCEP assessments included a comprehensive, external evaluation of the documentation and data available for these assessments. The technical evaluation entailed six tasks related to the design, administration, scoring, and reporting of the assessments:

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

Overall, the findings from these tasks indicate that the South Carolina assessments mostly adhere to sound testing practices as described in *The Standards for Educational and Psychological Testing*, and thereby support the validity of the test scores for their intended uses and purposes. No critical concerns were identified from the technical evaluation of the South Carolina assessments. Nonetheless, several recommendations are provided in Part I of this report to further strengthen and improve the quality of the assessments. We applaud South Carolina for securing an external evaluation of its assessments to help ensure their quality. Periodic evaluations of testing practices will help to ensure their continued technical soundness.

The evaluation included in Part I does not constitute a statement on the legal requirements of the South Carolina assessments, as compliance with the *Test Standards* is not synonymous with compliance with legal requirements. Part II of this report provides an evaluation of the minimum legal requirements of the SC READY assessments specified in Section 59-18-325 of the South Carolina Code of Laws.

References

Allalouf, A. (2007). Quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement: Issues and Practice, 26*, pp. 36-46.

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.

Dickinson, E.R., Chen, J., & Swain, M. (2017). *South Carolina Assessment Evaluation Report #1.* (2017 No. 019). Alexandria, VA: Human Resources Research Organization.

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

Huynh, H. (1979). Computational and statistical inference for two reliability indices based on the beta-binomial model. *Journal of Educational Statistics,* 4, pp. 231–246.

Huynh, H., & Saunders, J.C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement,* 17, pp. 351–358.

International Test Commission (2014). ITC guidelines for quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing, 14*, pp. 195-217.

Linacre, J. M. (2014). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

Nemeth, Y., Purl, J., & Smith, E. (2016). *Independent alignment review of the Florida Standards Assessment (FSA) in English language arts and mathematics.* Alexandria, VA: Human Resources Research Organization.

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

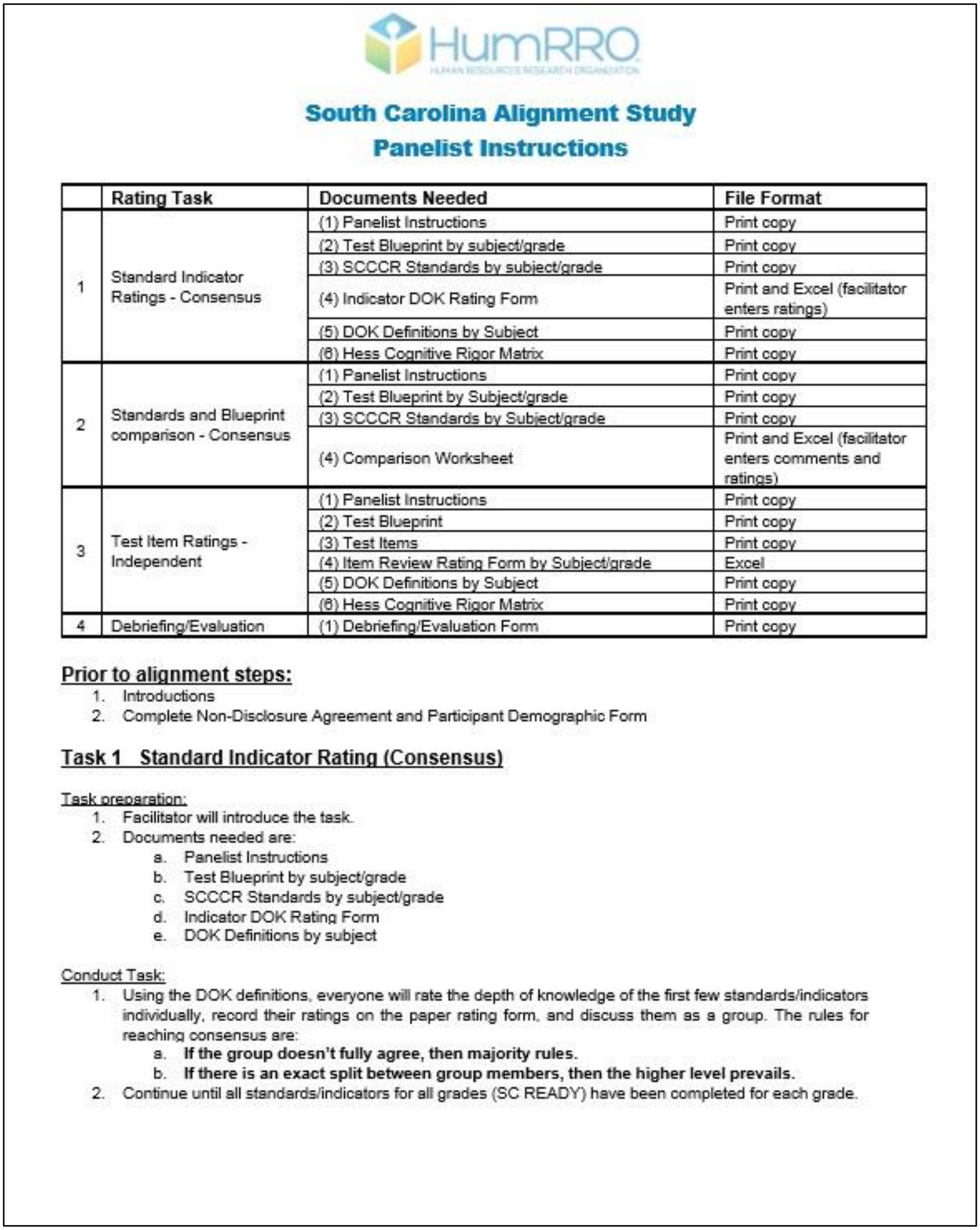
Webb, N. L. (1999). *Alignment of Science and Science standards and assessments in four states. (Research Monograph 18)*. Madison, WI: National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).

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: Panelist Instructions

Panelists were provided an instruction document as a reference to use through the workshop containing information regarding alignment tasks processes and rating code definitions.



South Carolina Alignment Study
Panelist Instructions

	Rating Task	Documents Needed	File Format
1	Standard Indicator Ratings - Consensus	(1) Panelist Instructions	Print copy
		(2) Test Blueprint by subject/grade	Print copy
		(3) SCCCR Standards by subject/grade	Print copy
		(4) Indicator DOK Rating Form	Print and Excel (facilitator enters ratings)
		(5) DOK Definitions by Subject	Print copy
		(6) Hess Cognitive Rigor Matrix	Print copy
2	Standards and Blueprint comparison - Consensus	(1) Panelist Instructions	Print copy
		(2) Test Blueprint by Subject/grade	Print copy
		(3) SCCCR Standards by Subject/grade	Print copy
		(4) Comparison Worksheet	Print and Excel (facilitator enters comments and ratings)
3	Test Item Ratings - Independent	(1) Panelist Instructions	Print copy
		(2) Test Blueprint	Print copy
		(3) Test Items	Print copy
		(4) Item Review Rating Form by Subject/grade	Excel
		(5) DOK Definitions by Subject	Print copy
		(6) Hess Cognitive Rigor Matrix	Print copy
4	Debriefing/Evaluation	(1) Debriefing/Evaluation Form	Print copy

Prior to alignment steps:

1. Introductions
2. Complete Non-Disclosure Agreement and Participant Demographic Form

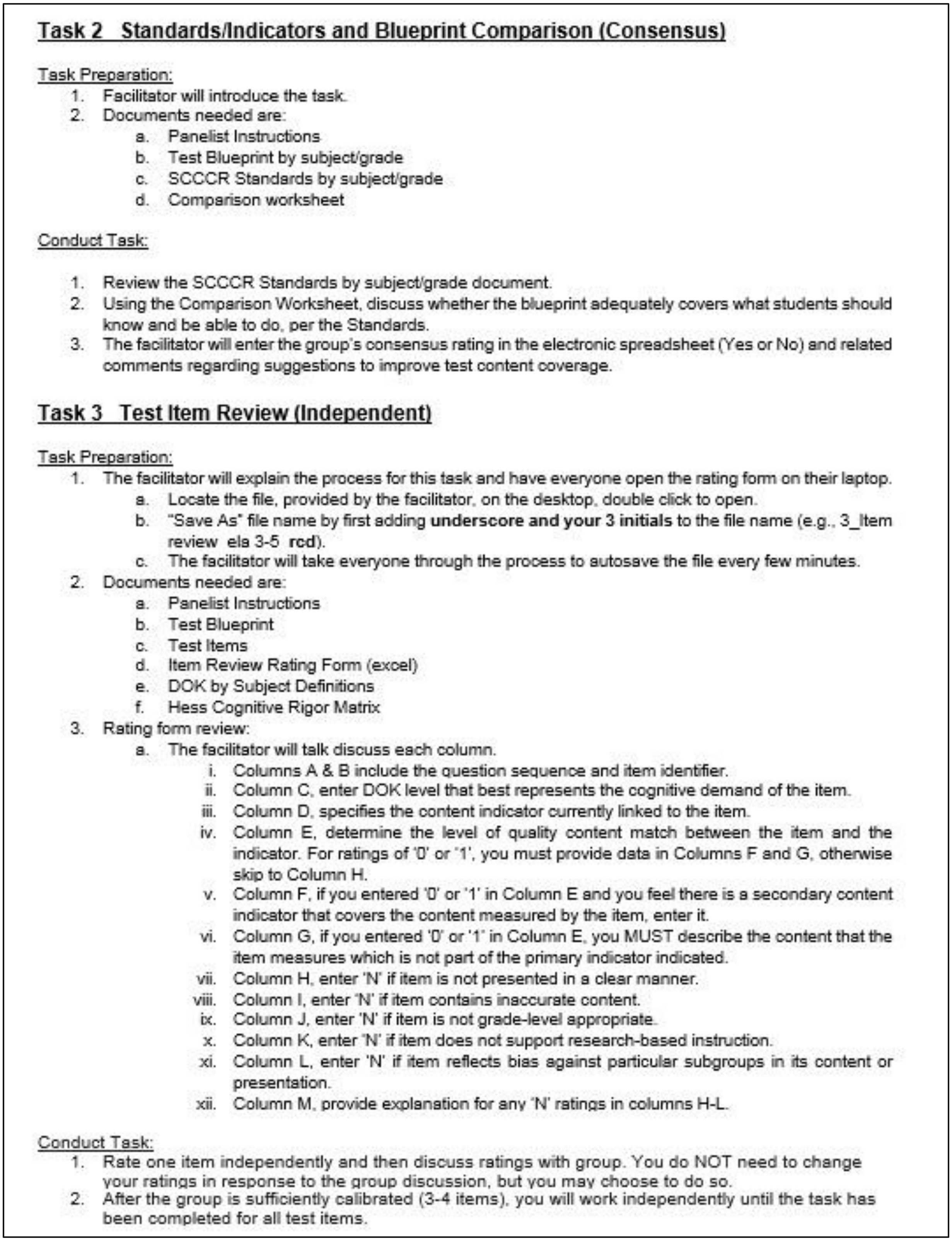
Task 1 Standard Indicator Rating (Consensus)

Task preparation:

1. Facilitator will introduce the task.
2. Documents needed are:
   a. Panelist Instructions
   b. Test Blueprint by subject/grade
   c. SCCCR Standards by subject/grade
   d. Indicator DOK Rating Form
   e. DOK Definitions by subject

Conduct Task:

1. Using the DOK definitions, everyone will rate the depth of knowledge of the first few standards/indicators individually, record their ratings on the paper rating form, and discuss them as a group. 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.**
2. Continue until all standards/indicators for all grades (SC READY) have been completed for each grade.



Task 2 Standards/Indicators and Blueprint Comparison (Consensus)

Task Preparation:

1. Facilitator will introduce the task.
2. Documents needed are:
   a. Panelist Instructions
   b. Test Blueprint by subject/grade
   c. SCCCR Standards by subject/grade
   d. Comparison worksheet

Conduct Task:

1. Review the SCCCR Standards by subject/grade document.
2. Using the Comparison Worksheet, discuss whether the blueprint adequately covers what students should know and be able to do, per the Standards.
3. The facilitator will enter the group's consensus rating in the electronic spreadsheet (Yes or No) and related comments regarding suggestions to improve test content coverage.

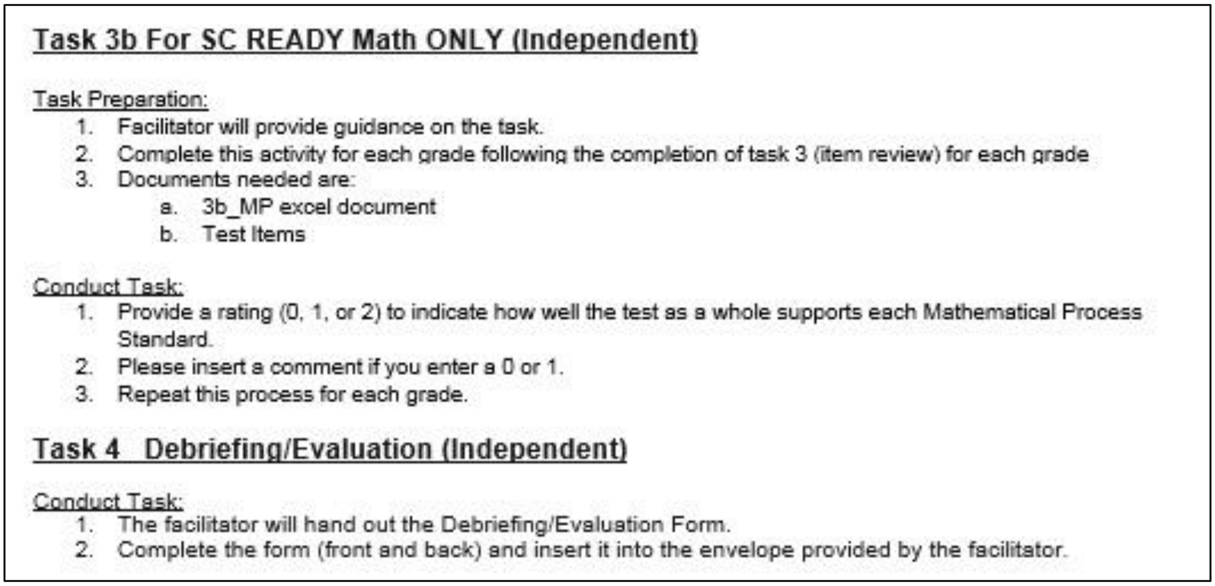
Task 3 Test Item Review (Independent)

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., 3_Item review ela 3-5 **rcd**).
   c. The facilitator will take everyone through the process to autosave the file every few minutes.
2. Documents needed are:
   a. Panelist Instructions
   b. Test Blueprint
   c. Test Items
   d. Item Review Rating Form (excel)
   e. DOK by Subject Definitions
   f. Hess Cognitive Rigor Matrix
3. Rating form review:
   a. The facilitator will talk discuss each column.
      i. Columns A & B include the question 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 indicator currently linked to the item.
      iv. Column E, determine the level of quality content match between the item and the indicator. For ratings of '0' or '1', you must provide data in Columns F and G, otherwise skip to Column H.
      v. Column F, if you entered '0' or '1' in Column E and you feel there is a secondary content indicator that covers the content measured by the item, enter it.
      vi. Column G, if you entered '0' or '1' in Column E, you MUST describe the content that the item measures which is not part of the primary indicator 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 (3-4 items), you will work independently until the task has been completed for all test items.



Task 3b For SC READY Math ONLY (Independent)

Task Preparation:

1. Facilitator will provide guidance on the task.
2. Complete this activity for each grade following the completion of task 3 (item review) for each grade
3. Documents needed are:
   a. 3b_MP excel document
   b. Test Items

Conduct Task:

1. Provide a rating (0, 1, or 2) to indicate how well the test as a whole supports each Mathematical Process Standard.
2. Please insert a comment if you enter a 0 or 1.
3. Repeat this process for each grade.

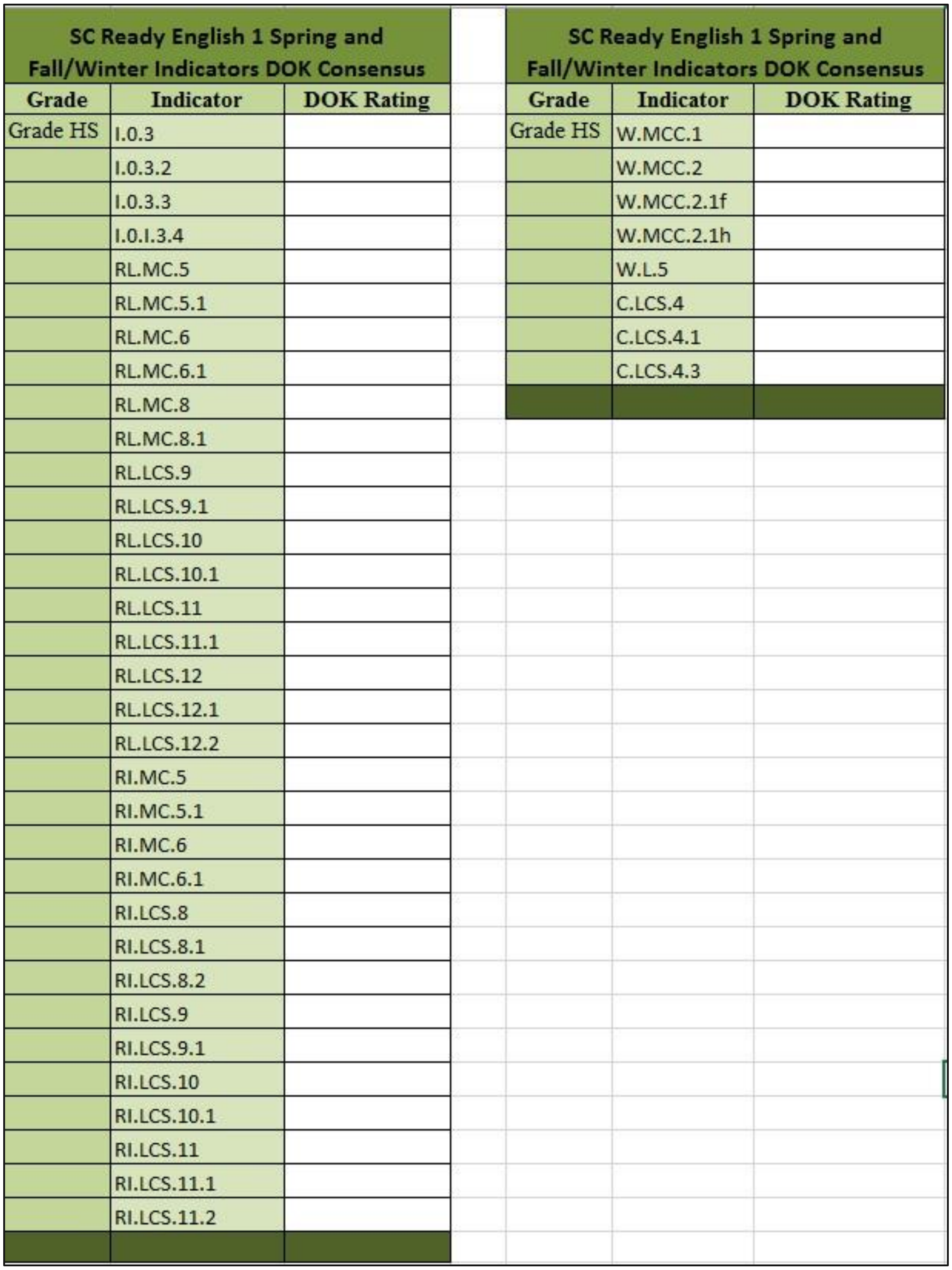
Task 4 Debriefing/Evaluation (Independent)

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.

Appendix B: Example Panelist DOK Rating Form

This form was used by panelists to enter their individual DOK ratings for content standards or indicators in preparation for discussions to determine the group's consensus rating (captured by the group facilitator).

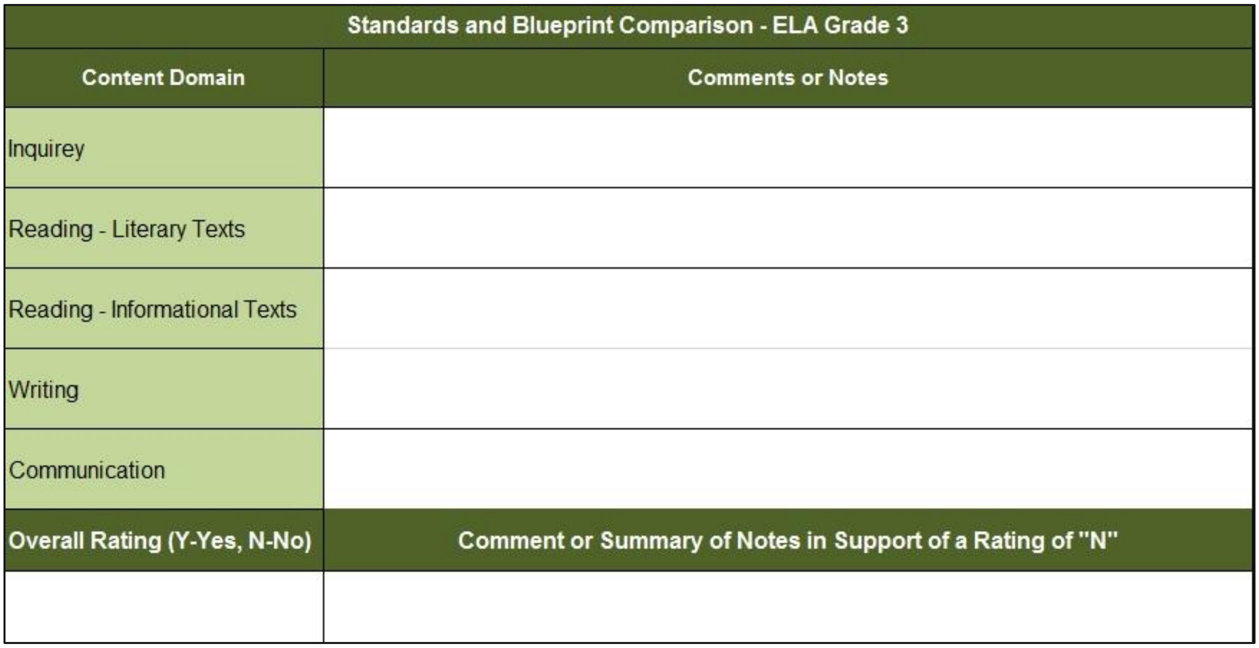


SC Ready English 1 Spring and Fall/Winter Indicators DOK Consensus		
Grade	Indicator	DOK Rating
Grade HS	I.0.3	
	I.0.3.2	
	I.0.3.3	
	I.0.I.3.4	
	RL.MC.5	
	RL.MC.5.1	
	RL.MC.6	
	RL.MC.6.1	
	RL.MC.8	
	RL.MC.8.1	
	RL.LCS.9	
	RL.LCS.9.1	
	RL.LCS.10	
	RL.LCS.10.1	
	RL.LCS.11	
	RL.LCS.11.1	
	RL.LCS.12	
	RL.LCS.12.1	
	RL.LCS.12.2	
	RI.MC.5	
	RI.MC.5.1	
	RI.MC.6	
	RI.MC.6.1	
	RI.LCS.8	
	RI.LCS.8.1	
	RI.LCS.8.2	
	RI.LCS.9	
	RI.LCS.9.1	
	RI.LCS.10	
	RI.LCS.10.1	
	RI.LCS.11	
	RI.LCS.11.1	
	RI.LCS.11.2	

SC Ready English 1 Spring and Fall/Winter Indicators DOK Consensus		
Grade	Indicator	DOK Rating
Grade HS	W.MCC.1	
	W.MCC.2	
	W.MCC.2.1f	
	W.MCC.2.1h	
	W.L.5	
	C.LCS.4	
	C.LCS.4.1	
	C.LCS.4.3	

Appendix C: Example Comparison of Standards and Blueprint Form

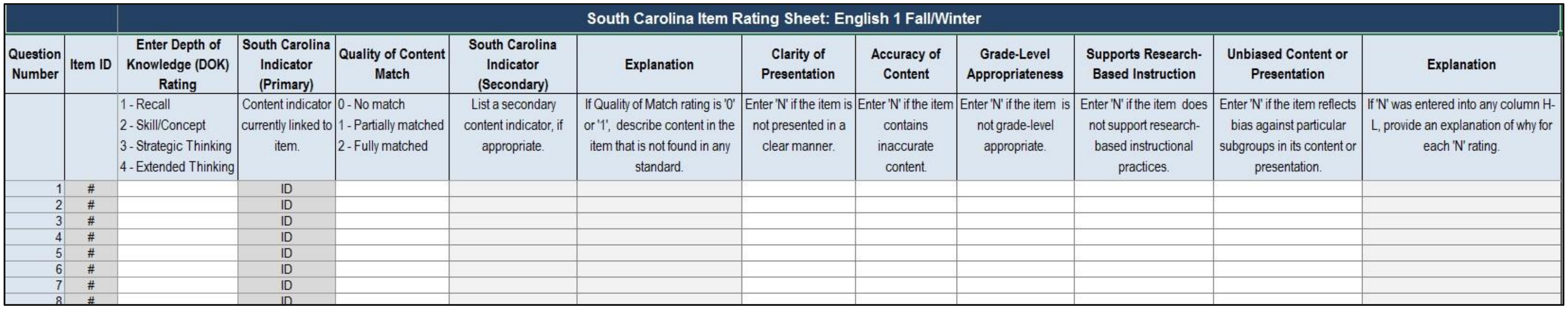
This form was used as an organizer for panelists to guide their discussion regarding the coverage of standards/indicators by the test blueprint. The facilitator recorded their comments by content domain, the group's overall rating, and summary notes supporting that decision.



Standards and Blueprint Comparison - ELA Grade 3	
Content Domain	Comments or Notes
Inquirey	
Reading - Literary Texts	
Reading - Informational Texts	
Writing	
Communication	
Overall Rating (Y-Yes, N-No)	Comment or Summary of Notes in Support of a Rating of "N"

Appendix D: Example Panelist Item Rating Form

Panelist used the Item Rating sheet in Excel™ to record their individual ratings for each test item for DOK; the quality of match of the linked standard or indicator; provide a secondary standard if applicable and explanation; and item quality ratings for clarity of presentation, accuracy of content, grade-level appropriateness, supports research-based instruction, and unbiased content of presentation.



South Carolina Item Rating Sheet: English 1 Fall/Winter

Question Number	Item ID	Enter Depth of Knowledge (DOK) Rating	South Carolina Indicator (Primary)	Quality of Content Match	South Carolina Indicator (Secondary)	Explanation	Clarity of Presentation	Accuracy of Content	Grade-Level Appropriateness	Supports Research-Based Instruction	Unbiased Content or Presentation	Explanation
		1 - Recall 2 - Skill/Concept 3 - Strategic Thinking 4 - Extended Thinking	Content indicator currently linked to item.	0 - No match 1 - Partially matched 2 - Fully matched	List a secondary content indicator, if appropriate.	If Quality of Match rating is '0' or '1', describe 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 does not support research-based instructional practices.	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.
1	#		ID									
2	#		ID									
3	#		ID									
4	#		ID									
5	#		ID									
6	#		ID									
7	#		ID									
8	#		ID									

Appendix E: Panelist Feedback

Panelists completed an evaluation of the alignment workshop after all panelist tasks were completed. The table provides the evaluation questions with the percentage of panelists in each group who responded with Strongly Agree or Agree. There were 5 panelists in ELA Grades 3-5, English 1, and Biology 1, and there were 6 panelists in ELA Grades 6-8, Math Grades 3-5, and Math Grades 6-8.

	Percent Strongly Agree or Agree					
Evaluation Question	ELA G3-5	ELA G6-8	Math G3-5	Math G6-8	Eng 1	Bio 1[a]
The training presentation in the large group provided useful information about the OSDE assessment systems and the alignment method used.	100%	83%	83%	100%	80%	100%
After the additional training in my small group, I felt prepared to be a panelist.	100%	100%	100%	100%	100%	100%
HumRRO staff seemed knowledgeable of the OSDE assessment systems and alignment steps.	100%	100%	100%	100%	100%	100%
The Panelist Instruction document was clear, understandable, and useful in performing the alignment steps.	100%	83%	100%	100%	80%	100%
The excel files were relatively easy to use to enter data.	100%	100%	100%	100%	60%	100%

[a] One panelist did not complete the form

There were some suggestions for improvement provided by the panelists. Two panelists felt that the content provided in the large group training was redundant to training provided in their panel group. There were 3 comments that the excel file rating sheets were not ideal as one person indicated that he or she was unfamiliar with excel and two other stated that GoogleSheets are easier to manage. Finally, two panelists felt the stacks of support materials (i.e., standards, panelist instructions, items, DOK definitions) were difficult to use.

Appendix F: DOK Consistency Results for SC READY ELA, by Reporting Category within Domain and Grade

Grade	Domain	Reporting Category	% below standard level		% at standard level		% above standard level	
			Mean	SD	Mean	SD	Mean	SD
ELA 3	Reading - Literary Text	Meaning and Context	70.90	10.00	29.10	10.00	0.00	0.00
ELA 3		Language, Craft, and Structure	0.00	0.00	73.10	6.00	26.90	6.00
ELA 3	Reading - Informational Text	Meaning and Context	78.00	4.50	22.00	4.50	0.00	0.00
ELA 3		Language, Craft, and Structure	16.10	6.90	61.10	7.90	22.80	1.20
ELA 3	Writing/ Inquiry	Meaning, Context, and Craft	91.40	6.00	8.60	6.00	0.00	0.00
ELA 3		Language	11.40	6.40	88.60	6.40	0.00	0.00
ELA 3		Inquiry	74.00	11.40	26.00	11.40	0.00	0.00
ELA 4	Reading - Literary Text	Meaning and Context	37.80	9.90	62.20	9.90	0.00	0.00
ELA 4		Language, Craft, and Structure	12.00	4.50	78.00	4.50	10.00	0.00
ELA 4	Reading - Informational Text	Meaning and Context	88.90	0.00	11.10	0.00	0.00	0.00
ELA 4		Language, Craft, and Structure	22.70	6.00	77.30	6.00	0.00	0.00
ELA 4	Writing/ Inquiry	Meaning, Context, and Craft	63.10	3.40	36.90	3.40	0.00	0.00
ELA 4		Language	25.00	0.00	66.70	0.00	8.30	0.00
ELA 4		Inquiry	100.00	0.00	0.00	0.00	0.00	0.00
ELA 5	Reading - Literary Text	Meaning and Context	92.70	7.60	7.30	7.60	0.00	0.00
ELA 5		Language, Craft, and Structure	52.50	5.60	40.00	10.50	7.50	6.80
ELA 5	Reading - Informational Text	Meaning and Context	96.00	5.50	4.00	5.50	0.00	0.00
ELA 5		Language, Craft, and Structure	91.10	5.00	8.90	5.00	0.00	0.00
ELA 5	Writing/ Inquiry	Meaning, Context, and Craft	75.30	4.90	24.70	4.90	0.00	0.00
ELA 5		Language	0.00	0.00	100.00	0.00	0.00	0.00
ELA 5		Inquiry	100.00	0.00	0.00	0.00	0.00	0.00
ELA 6	Reading - Literary Text	Meaning and Context	71.20	10.60	28.80	10.60	0.00	0.00
ELA 6		Language, Craft, and Structure	67.60	10.00	32.40	10.00	0.00	0.00
ELA 6	Reading - Informational Text	Meaning and Context	55.20	8.30	42.70	7.30	2.10	3.20
ELA 6		Language, Craft, and Structure	30.80	6.90	29.50	7.60	39.70	5.80

Grade	Domain	Reporting Category	% below standard level		% at standard level		% above standard level	
			Mean	SD	Mean	SD	Mean	SD
ELA 6	Writing/ Inquiry	Meaning, Context, and Craft	29.30	17.90	63.60	16.90	7.10	6.30
ELA 6		Language	22.20	14.10	59.30	9.10	18.50	20.70
ELA 6		Inquiry	56.30	23.40	39.60	18.40	4.20	6.50
ELA 7	Reading - Literary Text	Meaning and Context	70.60	12.40	29.40	12.40	0.00	0.00
ELA 7		Language, Craft, and Structure	61.10	13.60	38.90	13.60	0.00	0.00
ELA 7	Reading - Informational Text	Meaning and Context	91.40	10.00	8.60	10.00	0.00	0.00
ELA 7		Language, Craft, and Structure	78.00	16.50	22.00	16.50	0.00	0.00
ELA 7	Writing/ Inquiry	Meaning, Context, and Craft	71.90	8.50	28.10	8.50	0.00	0.00
ELA 7		Language	16.70	3.70	69.70	12.40	13.60	14.90
ELA 7		Inquiry	88.90	20.20	11.10	20.20	0.00	0.00
ELA 8	Reading - Literary Text	Meaning and Context	87.20	9.30	12.80	9.30	0.00	0.00
ELA 8		Language, Craft, and Structure	83.30	10.20	16.70	10.20	0.00	0.00
ELA 8	Reading - Informational Text	Meaning and Context	77.10	11.60	22.90	11.60	0.00	0.00
ELA 8		Language, Craft, and Structure	43.60	12.60	56.40	12.60	0.00	0.00
ELA 8	Writing/ Inquiry	Meaning, Context, and Craft	79.20	15.10	20.80	15.10	0.00	0.00
ELA 8		Language	25.00	0.00	43.80	27.10	31.30	27.10
ELA 8		Inquiry	76.20	14.80	23.80	14.80	0.00	0.00

Appendix G: Forms Construction Meeting Observation Checklist

Rating Codes for Fidelity Ratings

Score Level	Description of Score Level
1	Documented procedure was not followed; Actual procedure did not resemble documented procedure.
2	Documented procedure was rarely followed, or was followed incompletely or mostly incorrectly.
3	Documented procedure was followed some of the time, but not all the time. Aspects/steps of the procedure may have been missing or may not have been documented.
4	Documented procedure was mostly followed most of the time. Extraneous aspects/steps were rarely included.
5	Documented procedure was followed; there were no additional aspects/steps taken than what was planned.

Test Construction Process

The main steps of the SC READY ELA and Mathematics operational test construction process are outlined below.

Documented Procedure	ELA	Math	Consensus
1. The Content Specialist selects the initial set of operational items using the Excel-based file of the item pool for each subject. *Although a sequence of the operational items is preferred, the initial pull of test items can be left unsequenced for the purposes of the initial psychometric review.*	**Notes:** DRC Content Specialists looked at item *p*-values, point-biserials, keyed responses, and content areas (in their spreadsheet) to determine which items to remove from last year's form. Approximately 25% of items are refreshed each year, so ideally, every item is refreshed every 4 years. However, it's not clear whether this is systematic. Inquired whether they keep track of how long an item has been on the form (e.g., do they have limitations to how many years in a row an item can appear on the form?), and it doesn't seem like they do; they have only 2016 and	**Notes:** The content specialists organized print-outs of each item in administration order on a long table, and spread in FT/new items (i.e., sequencing) based on the required items per content category and item statistics. The first time they did this, it seemed a bit more based on content. After receiving feedback from the psychometrician the first time and being told that he was focusing on key distribution, *p*-value distribution (largely normally distributed), etc. they focused on these about equally as the content.	**Fidelity Rating: 3** **Notes:**

Documented Procedure	ELA	Math	Consensus
	2017 in their spreadsheet. Their goal is to have *p*-values between .4 and .8 and to have point-biserials greater than .25. Items not meeting both criteria are used only if necessary (e.g., exceptions are made to have a certain number of items for a passage). They looked at the median of point-biserials for last year's form and compared it to this year's proposed form. They said they try to keep items positioned similarly to last year's form. SCDE reviewed the proposed operational items and made comments about content and about standards (e.g., the SCDE Content Specialist asks DRC which standard a particular item is linked to, and if he/she disagrees, the standard is changed).		

Documented Procedure	ELA	Math	Consensus
2. The preliminary item selection and related item performance data is sent to the Senior Psychometrician for psychometric review.	**Notes:** The DRC Content Specialists send a spreadsheet with the proposed items to the psychometrician.	**Notes:** The content specialists sent their item choices to the psychometrician in an Excel spreadsheet (posting them on the DRC network).	**Fidelity Rating: 5**
3. As the Senior Psychometrician is reviewing the set of operational items (regardless of sequencing), a comparison of the psychometric requirements and the proposed form is documented.	**Notes:** The psychometrician provided a handout with information about the proposed form – *p*-value distribution, mean/SD/median/min/max *p*-value and point-biserial, key distribution, DIF distribution (across A, A+, A-, B+, B-, C+, C- categories), and new items. Neither the handout nor the discussion included a reference to particular rules/requirements that the psychometrician was following.	**Notes:** The DRC psychometrician took the item selections and input them into an Excel template that summarized the item statistics in terms of: 1) a histogram of the distribution of *p*-values, 2) the mean, median, sd, min, max of *p*-values and point-biserials for the items, 3) the distribution of keyed answers, and 4) DIF, categorized by A+, A-, B+, B-, C+, C-, indicating different levels of DIF on either the majority/minority side for gender and ethnicity.	**Fidelity Rating: 3**

Documented Procedure	ELA	Math	Consensus
4. The psychometric feedback is sent to the Content Specialist.	**Notes:** The psychometrician gave the handout and verbal feedback to the group (SCDE and DRC Content Specialists) about *p*-value distribution (e.g., bimodal), *p*-value mean (too low/difficult), key distribution (too many A and C keys), DIF (codes – making sure they're counter-balanced).	**Notes:** For each grade, the DRC psychometrician printed out the Excel template, and provided copies to the content experts and discussed his interpretation of the psychometrics of the forms. The psychometrician communicated whether the forms were an acceptable deviation from an ideal psychometric distribution, or whether content specialists should attempt to correct particularly high/low *p*-values or unbalanced key distributions.	**Fidelity Rating: 5**
5. The Content Specialist adjusts the set of operational items as necessary based on feedback from the Senior Psychometrician.	**Notes:** The DRC Content Specialists went through their item banks to search for items with specific qualities (e.g., keys, *p*-value) to insert on a form to address issues identified by the psychometrician (e.g., unequal key distribution, *p*-value distribution). Observed development of 5 forms, and this step only happened for one form; four forms were approved after the first submission, and grade 6 was approved after one additional revision.	**Notes:** As noted above, there was deference to the content specialists. However, when stronger objections were made, the content specialists went back to the available items and swapped in acceptable replacements.	**Fidelity Rating: 4**

Documented Procedure	ELA	Math	Consensus
6. **The Content Specialist submits revised version of the form to the Senior Psychometrician.** ***Note: In the event that adjustments cannot be made, the Content Specialist must provide rationales.***	**Notes:** As noted above, this happened for one grade, but all other forms were approved after the first submission.	**Notes:** As with step 2, they submitted any revisions to the DRC network where the psychometrician would download them and re-run the Excel template. Didn't see an instance where a strongly objected item/form was not able to be revised and an official rationale had to be made. However, a few times in the discussion for step 4, the psychometrician would say he wasn't particularly happy with an element of the form, the content specialists would say there weren't many alternatives, and it would stop there (so, it didn't rise to the level of step 5).	**Fidelity Rating: 4**

Documented Procedure	ELA	Math	Consensus
7. Repeat steps 3–6 until agreement is reached between the Content Specialist and Senior Psychometrician on the set of operational items.	**Notes:** The psychometrician defered to the DRC Content Specialists. For example, he said he didn't like the key distribution of the revised grade 6 form, and SCDE expressed concern as well, but the DRC Content Specialists said they couldn't do any better, so the form stood. The psychometrician also didn't like the DIF distribution, but the DRC Content Specialists said that DIF was low on their list of priorities, and they couldn't do better, so the form stood.	**Notes:** From the few instances where revisions needed to be made, they didn't go through the whole process. Usually it was only one or two items that needed to be replaced, so, the content specialists submitted the form, and the psychometrician accepted it without providing feedback.	**Fidelity Rating: 5**
8. The Content Specialist sends a list of operational items (or "test map") and corresponding item cards to the SCDE for feedback and/or approval. *The above steps are repeated until an approved form is created. Two scrambled versions are created to complete the Form Set.*	**Notes:** This happens earlier in the process. SCDE reviewed the items before the items were sent to the psychometrician (i.e., as part of step 1). This seems reasonable from a logistics perspective. Otherwise, the SCDE Content Specialists wouldn't have had anything to do throughout most of the workshop (e.g., on the	**Notes:** This happened during Step 1 in conjunction with the content specialist. Didn't see a scrambled version being created.	**Fidelity Rating: 5 – but happens during step 1** **Notes:** Did not observe creation of scrambled forms.

Documented Procedure	ELA	Math	Consensus
	first day, the psychometrician didn't do a review of the first grade – step 2 – until about 4pm). Did not observe them creating the scrambled versions of the forms.		
9. The operational items are also reviewed by the SCDE, APH, and Dr. Mickey Jones for appropriateness for the visually and hearing impaired. If any items are deemed inappropriate, the item is replaced. If an item is replaced, the updated operational form must be re-approved.	**Notes:** Sheila reviewed printed versions of the items independently and gave the items back to the group with sticky notes. Did not see what the sticky notes said, and the group did not look at them during the sessions observed. Note: edited the description of step 9 to reflect the handout received at the meeting.	**Notes:** Sheila reviewed the printed copies of the items, organized in proposed order. She wrote post-it notes on some item sheets. She would ask questions wondering whether items were TE or not, or ask about alternatives in some cases.	**Fidelity Rating: 5 – but did not observe the entire step**

Documented Procedure	ELA	Math	Consensus
10. The approved sequenced form is replicated as a Form Set within IDEAS, and a PDF of a draft test booklet is created. Placeholders are added if the form is to include appended field-test items. Form components, such as the test booklet cover, directions, and other relevant non-test-item content will also be added to IDEAS during the test booklet production process. *A copy of the final sequenced and scrambled forms are provided to SCDE before the end of the face to face Forms Construction Meeting.*	**Notes:** did not observe this step.	**Notes:** did not observe this step.	**Fidelity Rating: Did Not Observe**
11. The SCDE reviews the composed operational and field-test forms and provides guidance on item usage, format, and placement. Test maps will be provided to the SCDE along with the composed operational test forms. Both print and online modes of the test forms may be reviewed by the SCDE during this process.	**Notes:** did not observe this step.	**Notes:** did not observe	**Fidelity Rating: Did Not Observe** **Notes:**

Appendix H: Psychometric Replication Steps for ELA Grade 5

Steps:	Circumstances that Contributed to a Less Rigorous Replication:
1. Imported test map and arranged items in calibration order	– Had to identify calibration item order based on order in DRC's Winsteps output file. It was not documented in the materials we reviewed.
2. Imported student data and scored items	– Initial differences in data files led to the discovery that duplicate students were accidentally included in the 2017 ELA G5 spring calibration. DRC re-estimated parameters using a corrected datafile and determined that the estimation was invariant. HumRRO proceeded to replicate using the file with duplicates so we could identify differences if they existed. – Used the same student data file DRC input into Winsteps for their calibration and did not replicate documented exclusion rules or any data formatting. DRC mentioned (via the conference call) that their tech team applies the exclusion rules when they pull data for calibration and thus there are no exclusion rules to be replicated. However, one exclusion rule was identified in DRC's control file on the PSELECT command that had to be implemented in order to replicate DRC's counts. The exclusion rule could not be identified and was not documented as a necessary step in calibration. – Relied on DRC's Winsteps control file to identify the location of items in the file. The file layout for the calibration file was not documented in the materials we reviewed. – Relied on a combination of DRC's Winsteps control file and the data layout for the state data file to determine how to score items. Scoring rules for the calibration file were not documented in the materials we reviewed.
3. Conducted free calibration of operational items (except TDA item)	– Relied on DRC's Winsteps control file to identify Winsteps options used in calibration. Winsteps options (e.g., EXTRSC, UDECIM, STBIAS) were not documented in the materials we reviewed.
4. Put G4 vertical linking items on operational G5 scale	– Relied on DRC's Winsteps control file to identify the items to use in vertical linking, the position of the items in the data file, and the item type for scoring. These items could not be identified in the test map.
5. Compute vertical linking constant	– There was limited documentation in the initial materials we reviewed on the decision criteria around estimation and execution of Robust Z as well as the method of calculating the constant (results indicated the mean difference method was used). Additional documentation provided specifically for this task included the details necessary to replicate the final decision of items to be included in linking as well as the estimates.
6. Put operational items on the 2017 vertical scale	– Relied on DRC's Winsteps control file to understand how this step was done (e.g., use of UAMOVE). – Used DRC's documented vertical linking constant as opposed to the constant HumRRO produced (there were differences at the fourth decimal place).
7. Calibrate the TDA item on the vertical scale	– Relied on DRC's Winsteps control files to understand how this step was done (i.e., one calibration to estimate the TDA parameters anchored to the 2017 parameters on the vertical scale, then another to weight the contribution of the TDA item for conversion table estimation).
8. Compute horizontal linking constant	– Similar to step 5, there was limited documentation in the initial materials we reviewed about the Robust Z decision rules, the use of the t-test, and the steps in computing the Wright and Bell unweighted linking constant. Additional documentation was specifically created for this task by DRC in order to replicate Robust Z decision rules and resulting linking constant (which was the final linking estimate). The additional documentation did not review the Wright and Bell linking method (an alternative explored); thus, that could not be replicated.

South Carolina Assessment Evaluation Report #2
Part II: Legal Evaluation

Chapter 7: Review of Minimum Legal Requirements of SC READY (Task 7)[1]

Table of Contents

[1] *Note:* Consistent with legal citation conventions, reference citations in Part II are presented in footnotes rather than in the APA citation format used in Part I to assist the reader in connecting the information presented with its sources.

Table of Contents (Continued)

Table of Contents (Continued)

Table of Contents (Continued)

List of Exhibits

List of Figures

List of Charts

Table of Contents (Continued)

List of Tables

Table of Contents (Continued)

South Carolina Assessment Evaluation Report #2
Part II: Legal Evaluation

Chapter 7: Review of Minimum Legal Requirements of SC READY (Task 7)[2]

S.E. Phillips (Consultant, Assessment Law & Psychometrics)

Executive Summary

In its Request for Proposals for an assessment system evaluation, the Education Oversight Committee (EOC) included a requirement that the responder evaluate the minimum statutory requirements for the SC READY assessments after the 2017 administration. SC READY is a system of assessments that measure student achievement of the South Carolina state content standards in English language arts (ELA) and Mathematics in Grades 3 through 8.

In response, HumRRO contracted with Dr. S. E. Phillips, PhD, JD, a nationally recognized assessment law expert, for consultation on this legal evaluation (Task 7). The legal evaluation was completed following the 2017 administration of the SC READY assessments and consisted of three phases: review of written materials, follow-up inquiries to key personnel, and analysis and evaluation of the collected evidence. This final report for Task 7 details the findings of the legal evaluation, determines whether the minimum requirements of Section 59-18-325 of the South Carolina Code of Laws have been met, and makes recommendations for strengthening the legal and psychometric defensibility of the SC READY assessment system in the future.

Task 7: Results

The results of the legal evaluation are presented by criterion in the order in which the eight criteria appear in Section 59-18-325. After stating each criterion, relevant SC READY evidence supporting that criterion is presented followed by evaluative commentary on the quality and sufficiency of that evidence.

1. Comparison of Student SC READY Performance to Score Scales of Assessments of Comparable Standards in Other States

Evidence. SC READY comparison scores include user percentile ranks from "other states with comparable standards" and MetaMetrics'® lexile®/quantile® scores. Evidence relevant to Legislative Criterion 1 includes an Achieve Report discussing the comparability of South Carolina ELA and Mathematics content standards to the Common Core State Standards (Common Core) and other states' college and career readiness (CCR) content standards adapted after an original adoption of the Common Core, the composition of the user group contributing data for the "other states" percentile ranks, and linking studies used to map SC READY scores to the lexile® and quantile® frameworks.

Evaluation. The comparability of the content standards and representativeness of the three user states contributing data for the "other states" percentile ranks is unclear because no demographic or concordance information has been documented. Although the lexile®/quantile® user sample of over 3.5 million students is much larger and more geographically diverse, it still

[2] *Note:* Consistent with legal citation conventions, reference citations in Part II are presented in footnotes rather than in the APA citation format used in Part I to assist the reader in connecting the information presented with its sources.

may not be representative of students nationally and no claim is made about the similarity of users' content standards. In sum, comparative information is available for two volunteer user groups from two different contractors. Limited information about the composition of these user samples makes it difficult to judge their comparability or representativeness. On the other hand, these data may be the best available and do provide some useful comparative information.

2. Development of a System of Summative, Vertically-Scaled, Benchmarked, Standards-Based Assessments

Evidence. The SC READY assessments are a system of grade level, standards-based assessments administered at the end of the school year. HumRRO evaluations confirmed that the 2017 SC READY assessments demonstrated very good alignment between the content standards, test blueprints and test items for ELA and good to acceptable alignment for Mathematics. Vertical scale scores are reported and the tests are directly benchmarked to performance by students in relatively large and small user norm groups from two contractors.

Evaluation. The lexile® and quantile® trajectories to Grade 12 CCR ranges provide useful evidence for claims of *on track performance for CCR*, particularly for students who *meet expectations*, but the accuracy of such predictions for South Carolina students has not yet been documented. As an alternative, the state might consider using South Carolina data to validate a chain of performance linking each grade level to preparedness for the following grade level with a culminating prediction of sufficient content knowledge in Grade 8 to be prepared for CCR courses in high school that are in turn linked to appropriate CCR measures such as college admissions tests' CCR benchmarks.

Reliability estimates for SC READY were generally high and met the Assessment TAC recommendation of .85 for all subjects, grade levels and groups except students with disabilities in Grades 7 and 8 Mathematics. Similar reliability estimates are not yet available for ELA Reading and some reliability evidence is needed for the reporting category indicator scores.

The 2017 vertical score scale was developed from 2017 data for which lower grade items were administered in adjacent upper grades. A major issue with the 2017 SC READY vertical scale is the potential for confusion and distress when students with equivalent scale scores are compared or negative growth is reported. Alternatively, if one assumed (purely for illustration purposes) that the 2017 vertical scale grade level distributions exhibited the same minimal overlap as the within-grade-level scale scores reported for SC READY in 2016, the potential for misinterpretation and anxiety would be greatly reduced.

3. Creation of SC READY Scores for Achievement of State Standards, Preparation for the Next Grade Level, and Student Growth in ELA (reading, writing) and Mathematics

Evidence. Individual student score reports for the SC READY ELA and Mathematics tests include several different types of scores designed to provide evidence of student achievement of state standards. For the ELA total score, the ELA Reading subscore, and the Mathematics total score, the student receives a performance level designation of *exceeds expectations, meets expectations, approaches expectations, or does not meet expectations* as defined by the South Carolina grade-level content standards and standard setting activities. One might logically conclude that students who score at or above the *meets expectations* performance level cut score on their grade level SC READY ELA or Mathematics tests have sufficient prerequisite knowledge and skills to be adequately prepared for the material covered at the next grade level. Students can demonstrate growth in ELA and Mathematics by maintaining a *meets or exceeds* *expectations* performance level in the prior and current testing years, exceeding the prior year's lexile® or quantile® scores, or increasing their vertical scale scores.

Evaluation. There is substantial evidence that the SC READY assessments provide appropriate scores indicating achievement of state standards and preparation for the next grade level. The evidence for growth measures is less convincing. It is unfortunate that the 2017 vertical scale score model does not provide traditional growth scores with reasonable interpretations. Its contradictory properties for scores that are supposed to be comparable and potential for reporting negative growth may make its scale scores troublesome for important audiences such as parents, educators and the public.

This leaves only the lexile® and quantile® scores as reasonable measures of growth over time. However, these scores are incomplete growth measures for ELA because they include reading but not writing. Moreover, the samples used to link the SC READY scores to the lexile® and quantile® scales were quite small relative to the student population, and student motivation for the separate linking tests may have been diminished because students likely knew it was a research study with no reporting of individual student scores.

4. Measurement of Student Progress Toward National College- and Career-Ready Benchmarks Derived from Empirical Research and State Standards

Evidence. MetaMetrics® conducted empirical research to develop direct links to lexile® and quantile® CCR ranges by analyzing typical reading texts and mathematical materials used in postsecondary education and the workplace. The reported lexile® and quantile® predicted growth trajectories are selected from among a set of typical student growth curves from a North Carolina norm group that best fit the current (and earlier grade level, if available) point estimate(s). If the estimated growth trajectory ends within the CCR interval, the student is predicted to achieve CCR by the end of Grade 12. If not, the score report provides a recommended growth trajectory that reflects the proportional accelerated improvement across the remaining grades that will be needed to reach the CCR interval by the end of Grade 12.

The vertical moderation procedure used in standard setting for the SC READY assessments provided an indirect link to national CCR standards. Panelists were provided with impact data from students' 2015 ACT Aspire® test series scores linked to the ACT Assessment college admissions test when they made their cut score adjustments.

Evaluation. It is difficult to identify a single, appropriate, national benchmark for CCR. Many states have used college admissions test benchmarks, but they apply only to high school students and are problematic because they assess content that does not align very well with most state content standards. MetaMetrics® has taken a different approach by quantifying the complexity of reading text or mathematical materials typically encountered in entry-level college courses or jobs requiring a high school diploma. The validity data linking SC READY *meets expectations* performance intervals to the lexile® and quantile® *on track for CCR* target ranges provide persuasive evidence that longitudinal data yet to be collected for South Carolina will support current CCR predictions.

5. Establishment of at Least Four Student Achievement Levels

Evidence. Evidence relevant to Legislative Criterion 5 includes the policy definitions and performance level descriptors for four student achievement (performance) levels and the standard setting activities that delimited the test score intervals corresponding to each of the four performance levels for the SC READY ELA and Mathematics assessments in Grades 3-8.

Evaluation. The SC READY assessments include four performance levels, two that signify proficiency and two that do not. Each of the performance levels is described by general policy statements related to the subject matter and by more specific performance level descriptors related to the state content standards. There is good documentation of the standard setting activities that recommended cut scores to delimit the four performance levels on the test score scales.

The consistency with which the SC READY assessments are predicted to classify students in the same performance level if they were to retest under similar conditions is quantified by estimates of decision consistency. Decision consistency estimates for SC READY were high, especially for classifying students into two performance categories (proficient and not proficient).

6. Inclusion of a Variety of Question Types that Test Student Understanding of the Content

Evidence. There are six different question types utilized in the SC READY assessments. Each is designed to address a different type of student understanding of the content. The question types include multiple choice (recognize a correct answer), multi-select (distinguish multiple correct and incorrect answers), evidence-based selected response (use evidence from a text to justify and support an answer), short answer or gridded response (supply a correct answer by typing or blackening ovals in a number grid), technology enhanced (online only: drag and drop, click on a spot, graph, or arrange options correctly) and a text-dependent analysis essay item (written response supported by text evidence) scored holistically by two raters.

Evaluation. The SC READY assessments are composed of a variety of item types that measure student understanding of the content in different ways. For some items, students select a correct answer and for others, the student must produce the answer. Some items require distinguishing multiple correct and incorrect answers and some require identification of evidence that best supports an answer. For students testing online, a few items utilize some of the unique features of the technology. There is also an extended essay item that requires students to combine text analysis, writing skill and use of evidence to support an answer.

Several studies conducted by HumRRO support the quality of the SC READY items.
The evidence for the content validity, alignment, differential functioning, reliability and quality control all supports the appropriateness and quality of the SC READY items and test forms.
No indicators of text complexity, such as readability indices or passage/form word lengths, are reported for the SC READY assessments.

DIF statistics are within normal limits for a standards-based achievement test but ethnic DIF is reported only for African-Americans. There appear to be enough Hispanic students to also calculate DIF statistics for that group. Psychometric best practice is to ask the fairness/sensitivity committee to re-evaluate items exhibiting DIF to determine if the committee members can identify anything about the items likely to have caused the DIF. If yes, the item is revised; if not, it is assumed the result occurred by chance and the item is retained for use if needed to satisfy the test blueprint.

7. Test Administration in Paper-Based and Computer-Based Formats

Evidence. Evidence relevant to Legislative Criterion 7 includes mode administration data, the district waiver policy, test forms, a mode comparability study, separate scale score tables, test accommodations policies, and test security policies.

Overall in 2016 about 35% of students tested online and 65% tested on paper. In 2017, the percent of students testing online improved substantially, ranging from nearly 60% in Grade 3 to almost 85% in Grade 8. Waivers of the requirement to test all students online are granted by the State Board of Education (SBE). In 2017, the SBE granted 55 waivers, primarily for lack of sufficient infrastructure and testing devices.

At the request of SCDE, the contractor completed a mode comparability study for the online and paper/pencil forms using the Spring 2016 field test data. Only two of 449 (about ½%) of the SC READY ELA operational items exhibited mode DIF (one each in Grades 5 and 8). For Mathematics, no mode DIF items were identified. The mode comparability study also examined p-value differences for online and paper/pencil tests. Summed across all the items, the study found an advantage for paper/pencil of about 1½ to 3⅓ raw score points for ELA and .03 to .62 raw score points for Mathematics.

Evaluation. The mode comparability study did not account for overall differences in the ability of online and paper/pencil test takers to manage the logistics of responding to entire test forms. In addition, the observed raw score differences occurred in groups of unequal ability. To evaluate whether there is a true mode advantage for paper/pencil ELA test takers, a linking study using matched samples could be conducted. A useful methodology for doing so annually is to create matched groups by selecting representative samples from the larger group that match the smaller group to create reference and focal groups of equal size and ability.

In other applications, decisions to report mode equated scores have been made when the average difference is more than one raw score point or when differential advantages were observed in specific segments of the test score distribution. The purpose for conducting mode equating when empirical studies detect *practically significant* differential *test form* performance is to be fair to all students and remove any performance incentives for educators to prefer administering paper/pencil tests. Conducting mode comparability equating should remain a priority as long as a considerable number of students continue to be tested via paper/pencil.

Test Administration and Test Security Policies for SC READY are detailed and strict. Reporting of violations is mandatory and the statutory provisions and administrative rules provide clear guidelines for investigations and sanctions for violators.

South Carolina also has a clear and detailed Testing Accommodations Policy. Testing accommodations decisions are made by the student's individualized education program (IEP) team and it is considered a security violation if they are not administered as prescribed. There are appropriate procedures for requesting accommodated testing forms and the online test engine has several useful features available to all students. Testing accommodations have been appropriately classified as standard when the tested skills are congruent with those specified by the content standards and the resulting test scores are comparable to test scores obtained under standardized conditions.

South Carolina has made substantial progress moving schools and districts to online testing, but there are still substantial numbers of students testing paper/pencil in the lower grades. Providing support and incentives for meeting the 100% online goal (except for accommodations) will likely remain a challenge.

8. Information Reported That Can Assist Educators to Align Assessment, Curriculum, and Instruction

Evidence. Educators have several tools available to assist them in using SC READY assessment information to align assessment, curriculum and instruction. Evidence relevant to Legislative Criterion 8 includes the South Carolina ELA and Mathematics content standards, Performance Level Descriptors (PLDs), test blueprints and sample items, SC READY Individual Student Reports (ISRs), District and School Roster Reports and labels, the eDirect Information Portal and Lexile® and Quantile® Score Reports.

Evaluation. The SC READY assessments include informative score reports and user information to aid educators in utilizing the test results to align their curriculum and instruction with the tested content from the state content standards. Appropriate interpretive cautions are also included with the reported scores on the individual student score reports.

Task 7: Ratings

The Task 7 legal review examined and evaluated the available evidence to determine whether the 2017 SC Ready assessment system meets the eight minimum legislative criteria prescribed in Section 59-18-325. Based on this review, the eight legislative criteria were rated using the rating scale presented in Table A.

Table A. Rating Scale for Legislative Criteria

RATING	DESCRIPTION
Meets +	Robustly meets minimum legislative criteria; evidence is extensive for all aspects
Meets	Meets minimum legislative criteria; evidence is adequate for all aspects
Meets –	Barely meets minimum legislative criteria; evidence is limited for some aspects
Does Not Meet	Fails to meet minimum legislative criteria; evidence is missing or inadequate

The ratings of each of the legislative criteria reflect an assessment of the adequacy and strength of the evidence presented and the degree to which the evidence is consistent with professional psychometric standards and supports the legal defensibility of the assessment program. The ratings for each of the eight legislative criteria with key comments are presented in Table B.

Summary:* *Overall, the SC READY ELA and Mathematics assessment system meets all of the eight minimum legislative criteria prescribed in Section 59-18-325. Policymakers, educators and the public can have confidence that the scores South Carolina students obtain on the SC READY assessments accurately reflect their current achievement of state standards and provide meaningful guidance about their readiness for the academic content of the next grade level. The assessment system effectively utilizes a variety of item types and a comprehensive development and review process to screen, assemble and analyze items aligned to the state content standards. Psychometrically appropriate standard setting procedures were used to establish four student achievement levels labeled *does not meet expectations*, *approaches expectations*, *meets expectations,* and *exceeds expectations*. Online and paper/pencil Test Administration, Testing Accommodations and Test Security policies are detailed, clear and designed to produce psychometrically valid and reliable student scores. Individual student reports present test information clearly and concisely and contain appropriate caveats for interpreting test scores. The best available evidence links the test performance of South Carolina students to the performance of students in other states and to college- and career-readiness. Useful information is provided for aligning curricula/instruction with the assessments.

Table B. Ratings and Comments for the Eight SC READY Legislative Criteria

RATING	LEGISLATIVE CRITERIA Comments
Meets	**1. LINKED SCALES FOR COMPARISON TO OTHER STATES WITH COMPARABLE STANDARDS** comparison groups are best available but may be nationally unrepresentative, of inadequate size, or have insufficiently aligned content standards
Meets	**2. VERTICALLY-SCALED, BENCHMARKED, STANDARDS-BASED, SUMMATIVE ASSESSMENT SYSTEM** system of grade level, standards-aligned, end-of-year tests with potentially confusing vertical scale scores and *on track for CCR* benchmarks
Meets –	**3. PERFORMANCE AGAINST STATE STANDARDS IN ELA, READING, WRITING AND MATHEMATICS; PREPAREDNESS FOR THE NEXT GRADE; GROWTH** validity studies linking test scores to performance at the next grade level not yet done; vertical scale scores may show negative growth and other growth evidence is indirect; writing is part of ELA but no subscores with achievement levels are reported
Meets –	**4. PROGRESS TOWARD NATIONAL CCR BENCHMARKS FROM EMPIRICAL RESEARCH AND STATE STANDARDS** available CCR evidence is indirect but persuasive; direct CCR predictions for elementary students are ill-advised due to imprecision and unproven validity; inchoate validity studies linking Grade 8 test scores to admissions test CCR benchmarks
Meets +	**5. ESTABLISHMENT OF AT LEAST FOUR STUDENT ACHIEVEMENT LEVELS** appropriate and well-documented standard setting procedures and performance level descriptors for 4 levels (*does not meet, approaches, meets*, & *exceeds expectations*)
Meets +	**6. USE OF A VARIETY OF ITEM TYPES REQUIRING DEMONSTRATION OF CONTENT UNDERSTANDING** mixture of item types; multiple-select, evidence-based & text-dependent analysis essay items simulate the type of thinking and analysis typically associated with CCR
Meets	**7. AVAILABILITY OF ONLINE AND PAPER/PENCIL ADMINISTRATIONS** paper form and easy-to-use online testing platform with appropriate accommodations; online testing goals and capabilities (e.g., TE items; adaptive testing) not yet fully attained
Meets	**8. REPORTS INFORMATION TO ASSIST EDUCATORS IN ALIGNING CURRICULA WITH ASSESSMENTS** summative assessments useful for global curricular alignment; reporting categories guide educators to areas for more in-depth evaluation

As with any new testing program, there are many supporting research studies and procedural decisions yet to be finalized for future test administrations to maintain the quality, equivalence, alignment and usefulness of the test forms. The SCDE has a knowledgeable Assessment TAC and experienced contractor staff to aid them in appropriately constructing and analyzing future test forms and in designing and conducting useful research studies. In the spirit of improving and strengthening the assessment program as these future actions are deliberated, the next section provides specific recommendations related to each legislative criterion. Addressing these recommendations and the suggestions provided in prior sections of this report will further support the psychometric and legal defensibility of the SC READY assessment system.

Task 7: Recommendations

Recommendations for improvement are listed below. Each recommendation is associated with one of the eight legislative criteria and has been assigned a priority rating of *urgent*, *high*, *medium* or *low* as described in Table C. In addition to improving legal defensibility, many of these recommendations also support improved psychometric defensibility.

Table C. Priority Ratings for Recommendations

PRIORITY	DESCRIPTION
Urgent	Definitely needs to be considered and addressed now
High	Needs to be considered and addressed as soon as possible
Medium	Should be considered and addressed as time and circumstances permit
Low	Might be considered and addressed as part of long term planning

Urgent Priority

Legislative Criteria 1 & 2: Request that the contractor provide South Carolina with additional validity information about the participating states and the methods used to derive the reported *other states with comparable standards* percentile rank norms. Consider requesting that the contractor organize alignment information similar to a textbook crosswalk (e.g., from the Achieve Report or published state content standards) to confirm the comparability of the other states' standards to those of South Carolina. Also consider exploring the option of reporting percentile ranks for *other states* independent of South Carolina data.

Legislative Criteria 2 & 3: Weigh the advantages against the potential misinterpretations of using the current, vertical scale, and consider adopting a more traditional vertical scale before reporting 2018 SC READY scores to provide reasonable growth score interpretations and avoid the appearance of negative growth. Now is an ideal time to make this change before a second year of comparative data is reported. Score reports for 2018 could report revised 2017 scale scores on the new vertical scale for comparison.

Legislative Criterion 5: Urge the State Board of Education (SBE), with the advice and consent of the Education Oversight Committee (EOC) per Section 59-18-320(D), to officially adopt the SC READY cut scores.

Legislative Criterion 7: Create a backup test form for each grade/subject to be held in reserve in case the operational test form is compromised before all schools have finished testing.

Legislative Criterion 8: Provide additional explanatory text in the Score Report User's Guide identifying the standard error of measurement (SEM) type and size actually used to calculate the scale score ranges reported on the individual student reports, and if necessary, revise the sample reports to be consistent with the actual data.

High Priority

Legislative Criteria 1-8: Consolidate scattered program documents and information into a single, expanded Technical Manual with summarized material and data, relevant appendices, and references to supporting documents.

Legislative Criterion 2: For the Grades 3-8 ELA Reading subscores, report decision consistency estimates and reliabilities using the same methodology and statistics as for the total ELA scores. Revise, if necessary, when scores become more stable.

Legislative Criterion 2: To be consistent with the 2014 *Test Standards*, report preliminary reliability estimates for the reporting category indicator scores (low, middle, high) now and then revisit and revise them later, as appropriate, when scores are more stable.

Legislative Criterion 4: Consider creating an ELA Writing subscore and reporting performance levels and statistics similar to what is currently being done for ELA Reading.

Legislative Criterion 6: Document the frequency of item usage across years and use this information to target items for replacement based on prior exposure.

Legislative Criterion 6: Calculate ethnic differential item functioning (DIF) for Hispanics which represent about 9% of the South Carolina Grades 3-8 student population. Special rules/procedures for small samples may be appropriate for some grade/subject combinations.

Legislative Criterion 6: Consider routine replication of psychometric processing by an independent third party as an additional quality check. This will require more detailed documentation of procedures.

Legislative Criteria 6 & 7: As long as significant numbers of schools continue to census test with paper/pencil, conduct annual mode equating studies for ELA to ensure comparable scores and deter incentives for avoiding online testing. Also do so at least once for Mathematics to confirm that the differences are too small to warrant adjustment.

Legislative Criterion 7: Reconsider whether oral test administrations of the ELA Reading subtest should continue to be classified as standard accommodations in Grades 4-8 given the skill differences between reading and listening comprehension, the Achieve Report finding that reading fluency skills are included in the state content standards through the upper grades, and the removal of students tested orally from the lexile® linking study calibrations.

Medium Priority

Legislative Criterion 2: Design and conduct empirical research studies to validate CCR benchmarks using South Carolina data.

Legislative Criterion 3: Print numerical values next to point estimates on the lexile® and quantile® score report graphs to make year-to-year growth comparisons easier.

Legislative Criterion 3: Conduct research studies to empirically confirm that SC READY proficiency scores indicate adequate preparation for the next grade level for South Carolina students.

Legislative Criteria 3 & 4: Consider placing error bands around the reported lexile® and quantile® growth trajectories using ± 1 SEM estimated from the longitudinal sample. Also consider strengthening the cautionary statements at the bottom of the score reports. Develop a research plan to collect validity evidence to support CCR claims for South Carolina students.

Legislative Criterion 5: For future standard settings, select a wider representation of stakeholders to serve on the vertical moderation panels.

Legislative Criterion 6: Use an index of readability or total word counts to track the reading load for ELA passages and ELA and Mathematics test forms within and across grade levels.

Legislative Criterion 6: Ask the fairness/sensitivity educator committee to re-examine items with gender or ethnic DIF when deciding whether to retain or revise them.

Legislative Criterion 6: Report demographic information for fairness/sensitivity and content review committees similar to that reported for standard setting committees.

Legislative Criterion 7: Expand the number of annual site visits to increase coverage and deterrence. Develop a site visit plan and seek Assessment TAC advice. Select schools where violations are suspected and randomly select others so each District receives at least one unannounced visit over a several year period.

Low Priority

Legislative Criteria 2 & 6: Consider convening an experienced educator panel to reconsider the assessment of inquiry skills for ELA and blueprint weights for Mathematics.

Legislative Criterion 6: Consider specifying target depth of knowledge (DOK) levels in the test blueprints to support greater consistency with the content standards, especially for ELA where the greatest variability was observed.

Legislative Criterion 6: Superimpose cut scores on the Rasch item maps and identify the content of the items within each performance level to refine the PLDs and further strengthen the standards-based validity evidence for the SC READY assessment system.

Legislative Criterion 7: Continue to expand the availability of accommodated practice materials. Develop a plan for monitoring the provision of accommodations using school/district testing coordinators and/or site visits.

Legislative Criterion 7: Continue to explore item formats that take full advantage of the technological capabilities of online testing. Consider computer adaptive testing to shorten test lengths and administration times, and speed score reporting while maintaining score accuracy.

Chapter 7: Review of Minimum Legal Requirements of SC READY (Task 7)[3]

Task 7: Introduction

SC READY is a system of assessments that measure student achievement of the South Carolina state content standards in English language arts (ELA) and Mathematics in Grades 3 through 8. In its Request for Proposals for an assessment system evaluation, the Education Oversight Committee (EOC) included a requirement that the responder evaluate the minimum statutory requirements for the SC READY assessments by analyzing "whether [SC READY] meets the minimum legal requirements of Section 59-18-325" after the 2017 administration.[4] Section 59-18-325 of the South Carolina Code of Laws enumerates, in part, the eight minimum requirements described below.

> The summative assessment must be administered to all students in grades three through eight … . The summative assessment must assess students in English/language arts and mathematics, including those students as required by the federal Individuals with Disabilities Education Act and by Title I of the Elementary and Secondary Education Act. For purposes of this subsection, "English/language arts" includes English, reading, and writing skills as required by existing state standards. The assessment must be a rigorous, achievement assessment that measures student mastery of the state standards, that provides timely reporting of results to educators, parents, and students, and that measures each student's progress toward college and career readiness. Therefore, the [assessments] must meet all of the following **minimum requirements**:
>
> (a) compares performance of students in South Carolina to other students' performance on comparable standards in other states with the ability to link the scales of the South Carolina assessment to the scales from other assessments measuring those comparable standards;
>
> (b) [is] a vertically scaled, benchmarked, standards-based system of summative assessments;
>
> (c) measures [students'] preparedness for the next level of their educational matriculation and individual student performance against the state standards in English/language arts, reading, writing, and mathematics and student growth;
>
> (d) documents student progress toward national college and career readiness benchmarks derived from empirical research and state standards;
>
> (e) establishes at least four student achievement levels;
>
> (f) includes various test questions including, but not limited to, multiple choice, constructed response, and selected response, that require students to demonstrate their understanding of the content;
>
> (g) [is available for administration in] paper-based … [and] computer-based format[s] …; and
>
> (h) [reports information which can assist] school districts and schools in aligning assessment, curriculum, and instruction.[5]

[3] *Note:* Consistent with legal citation conventions, reference citations in Part II are presented in footnotes rather than in the APA citation format used in Part I to assist the reader in connecting the information presented with its sources.
[4] State of South Carolina Request for Proposal (RFP), *Evaluation of State Assessments*, Aug. 12, 2016, Scope of Work Section III (f).
[5] South Carolina Code of Laws on Educational Assessment and Accountability, Section 59-18-325, College and career readiness assessment; Summative assessment, Subsection C(1), emphasis added.

In response, HumRRO contracted with Dr. S. E. Phillips, PhD, JD, a nationally recognized assessment law expert, for consultation on this legal evaluation (Task 7). The legal evaluation was completed following the 2017 administration of the SC READY assessments and consisted of three phases: review of written materials, follow-up inquiries to key personnel, and analysis and evaluation of the collected evidence. This final report for Task 7 details the findings from the legal evaluation, determines whether the minimum requirements of Section 59-18-325 of the South Carolina Code of Laws have been met, and makes recommendations for strengthening the legal and psychometric defensibility of the SC READY assessment system in the future.

Task 7: Methods

The work of Task 7 was conducted in three phases. These phases included collection and review of written documentation, additional requests to the EOC, the South Carolina Department of Education (SCDE) and the testing contractor, Data Recognition Corporation (DRC) for additional information and clarifications, and analysis and evaluation of the available evidence for compliance with the statutory requirements. The activities in each of these three phases are described more fully in the next sections.

Phase I

In the first phase of the legal review, the legal requirements of Section 59-18-325, the testing program documentation for SC READY that specifically addressed the eight statutory minimum requirements listed in Section III (f) (pages 16-17) of the RFP, and any related topics identified as also covered by the statute were carefully reviewed. This review of written materials included the findings and analyses from Tasks 1 through 6 discussed in previous chapters, but with a specific focus on consistency with the minimum legal requirements specified in Section 59-18-325. In addition to the specific requirements listed in Section 59-18-325, the related legal and psychometric defensibility matters of a) the reliability of test scores, b) testing accommodations for students with disabilities (SDs) and English language learners (ELs), c) appropriateness and completeness of test security policies, d) fairness and sensitivity considerations, including item bias reviews and differential performance statistics, e) subgroup performance related to potential disparate impact, and f) the alignment of content standards, test blueprints and test items were also considered.

The review of program documentation for legal defensibility of the SC Ready assessment system included the following:

- Information collected and analyzed for Tasks 1 – 6 of this Report (see lists in Chapters 1-6)
- Paper/pencil and online test forms
- SC READY Technical Manual (TM)
- Item level sensitivity and differential performance review information
- Alignment data
- SC READY Test Administration Manual (TAM)
- Testing Accommodations and Test Security policies
- Reported procedures for setting performance level standards, including impact data
- Reported methods for developing preparedness and college- and career-ready (CCR) benchmarks
- Vertical equating and scaling documentation

- SC READY Score Report User's Guide (SRUG), including sample student and school reports
- Subgroup performance data by grade, subject and year
- Other documented procedures, studies and website information.

In addition to consistency with legislative requirements, consideration was also given to whether the implementation of the documented policies and procedures has been consistent with relevant federal laws and regulations that govern aspects of state testing, such as the Individuals with Disabilities Education Act (IDEA), the Americans with Disabilities Act (ADA), and the Every Student Succeeds Act (ESSA). Also evaluated was whether the procedures used were consistent with current professional best practices as embodied in the 2014 *Standards for Educational and Psychological Assessment* (*Test Standards*) and the 1998 *Code of Fair Testing Practices*.[6] Finally, specific questions were identified for follow up with state and contractor personnel to determine whether any additional or related materials were available for further review and clarification.

Phase II

During the second phase of the legal review, written inquiries and targeted phone-based conversations were conducted with key state and contractor personnel to supplement information in the written materials and files and to further explore key issues, clarify details, and gather additional information. Some evidence for the legal review was also gathered concurrently with inquiries related to the other six tasks addressed in this report.

Phase III

The final phase of the legal review consisted of a detailed analysis and evaluation of the written documentation and responses to the inquiries of key personnel to determine whether the minimum requirements of Section 59-18-325 had been met for the 2017 administration of SC READY and to assess the quality and sufficiency of the available evidence. Finally, recommendations are offered, where appropriate, for adjustments to the SC READY assessments that could strengthen adherence to statutory requirements and psychometric standards. The next sections detail these findings, ratings and recommendations.

Task 7: Results

The results are presented by criterion in the order in which the eight criteria appear in Section 59-18-325. After stating each criterion, relevant SC READY evidence supporting that criterion is presented followed by evaluative commentary on the quality and sufficiency of that evidence.

1. Comparison of Student SC READY Performance to Score Scales of Assessments of Comparable Standards in Other States

Evidence

Evidence relevant to Legislative Criterion 1 includes comparability of South Carolina ELA and Mathematics content standards to the Common Core State Standards (Common Core) and other states' college and career readiness (CCR) content standards, reported percentile ranks

[6] APA, AERA, NCME (2014). *Standards for Educational and Psychological Testing.* Washington, DC: APA [*Test Standards*]; Joint Committee (1998). *Code of Fair Testing Practices*, Washington, DC: APA.

from other states with comparable standards, MetaMetrics'® lexile® and quantile® scores, and information from other achievement tests.

Comparability of South Carolina Content Standards to CCSS and Other States' Standards

Percentile ranks for SC READY ELA and Mathematics total scores are reported for two norm groups: South Carolina students and students from other states with similar content standards. The degree to which this information indicates competitiveness in the College and Career Ready (CCR) marketplace depends in part on whether the content standards for South Carolina and the other states with similar content standards remained consistent with acknowledged national CCR principles when these states replaced their originally-adopted Common Core content standards with their own versions of CCR content standards.

One answer to this question is provided by a recent study undertaken by Achieve to review changes to state standards since their original adoption of the Common Core State Standards. In 2010 when the CCSS were first introduced, 45 states and DC adopted the Common Core. Subsequently, under increased political pressure, 24 states have reviewed and revised their ELA and Mathematics content standards. Achieve reviewed and rated the revised ELA and Mathematics content standards in these 24 states against nine key ELA/Literacy and 7 key Mathematics elements identified by research as necessary foundations for effective CCR. The following 3-point scale provided the basis for the ratings.[7]

Rating	Description
2 = STRONG	The CCR element is clearly and fully addressed
1 = MODERATE	The CCR element is not clearly or completely addressed
0 = WEAK / ABSENT	The CCR element is weak or nonexistent

Achieve's ratings for South Carolina and two sets of comparison states, three lower scoring states and three higher scoring states, are summarized in Table 1.

> ***"South Carolina does not define grade-level text or detail any of the factors that should be considered to determine grade-level complexity."***
>
> ***"The South Carolina Disciplinary Literacy practices … consist of three broad-based recommendations – not sufficient detail to lead to effective instruction on disciplinary literacy … South Carolina explicitly states that the disciplinary practices 'are not standards' and that they therefore should not be assessed. Expectations that are not assessed often take a back seat in instruction to content and skills that will be assessed."***
>
> — Achieve Report, p. 17, 23

For ELA, South Carolina's content standards received strong ratings in every category except *Analysis of Text Complexity & Guidance* and *Disciplinary Literacy.* Achieve's comments explaining the moderate ratings for these two key elements are shown at left.

[7] Achieve (2017). *Strong Standards: A Review of Changes to State Standards Since the Common Core*, www.achieve.org/state-standards-remain-strong.

Table 1. Achieve Report Ratings of States' ELA and Mathematics CCR Content Standards

ELA Key CCR Elements	South Carolina	Lower Scoring			Higher Scoring		
		MO	OK	AZ	FL	OH	ID
Foundational Skills	2 = Strong	2	1	2	2	2	2
Reading Literary and Informational Texts	2 = Strong	1	1	2	2	2	2
Evidence Drawn from Text	2 = Strong	2	1	2	2	2	2
Academic Vocabulary Acquisition & Use	2 = Strong	1	2	2	2	2	2
Writing from Sources and Research	2 = Strong	1.5	2	2	2	2	2
Oral Communication and Collaboration	2 = Strong	2	2	2	2	2	2
Grammar and Conventions	2 = Strong	2	2	2	2	2	2
Analysis of Text Complexity & Guidance	1 = Moderate	0	0	1	2	2	2
Disciplinary Literacy	1 = Moderate	0	2	*	2	2	2

Mathematics Key CCR Elements		South Carolina	Lower Scoring			Higher Scoring		
			PA	OK	IN	IA	NJ	ID
Structure		1 = Moderate	0	0	1	2 2	2	
Mathematical Practices		2 = Strong	1	1	2	2 2	2	
Procedures, Conceptual Understandings, and Applications		2 = Strong	0	0	1	2 2	2	
Sequencing		1 = Moderate	1	0	0	2 2	2	
Grades K-5	Focus on arithmetic	1 = Moderate	2	1	0	2 2	2	
	Memorize single-digit sums & products	0 = Weak/Absent	0	0	0	2 2	2	
Grades 6-8	Address critical topics; G8 solve pairs linear equations algebraically	2 = Strong	1	0.5	1.5	2 2	2	
High School	Modeling emphasized; Statistics through Algebra II	0 = Weak/Absent	0	0	1	2 2	2	

* Under development; *Source:* Achieve (2017). *Strong Standards: A Review of Changes to State Standards Since the Common Core*.

For Mathematics, South Carolina's content standards rated strong in mathematical practices; procedures, conceptual understanding and applications; and Grades 6-8 topics. But the ratings were only moderate for structure, sequencing, and Grades K-5 focus on arithmetic. The lowest ratings were given for Grades K-5 memorization of single digit sums and products and high school content. Achieve's comments explaining the moderate and weak ratings are shown at the top of the next page.

WEAK RATINGS

"South Carolina add[s] primary-grade standards related to patterns that are not connected to numbers (e.g., shapes and sounds) … that might detract from the emphasis on arithmetic in grades K-5."

"Some of the statistics topics appear in high school but only in a fourth-year course, which de-emphasizes the importance of statistics for all students."

— Achieve Report, p. 36, 41

MODERATE RATINGS

"South Carolina

- *lack[s] an intermediate level of organization and thus lose[s] how the standards are clustered for specific purposes under domain titles …*
- *expect[s] work with angles in grade 3 before introducing and defining angles and their measures in grade 4 …*
- *add[s] primary-grade standards related to patterns that are not connected to numbers [and] might detract from the emphasis on arithmetic."*

— Achieve Report, p.30, 34

Percentile Ranks from Other States

South Carolina has contracted with Data Recognition Corporation (DRC) to lease items from DRC's item bank of college and career ready (CCR) items. According to DRC, these CCR items are also utilized by three other states with *comparable academic content standards*.[8] Like South Carolina, two of these other states have adapted their state content standards from an earlier adoption of the Common Core. The Common Core was developed by the Council of Chief State School Officers and a consortium of state governors to reflect the content and skills in English language arts (ELA) and mathematics that students are expected to learn at each grade level, culminating in high school graduates who are sufficiently prepared academically to be successful in postsecondary education and the workplace (i.e., CCR).

The contractor provides user norms for SC READY that include South Carolina plus the three other states that use the contractor's CCR item bank and are said to have "comparable content standards." The contractor has calculated percentile ranks that quantify the percent of students in the four-state user norm group that score below each possible scale score on the SC READY ELA and Mathematics grade-level tests. For example, a third grade student whose ELA percentile rank is 75 has performed better than three quarters of the students in the user norm group. Percentile ranks comparing student performance to South Carolina students who were administered the same SC READY test are also reported. Together, the two percentile ranks describe the relative performance of each South Carolina student to students in South Carolina and to students in the user norm group. For example, if the third grade student described above earned a South Carolina percentile rank of 70, the student's ELA performance would be higher compared to students from other states than compared to students from South Carolina. The percentile ranks from other states are one indicator that allows South Carolina student performance to be compared to that of students in other states with *comparable content standards* in ELA and Mathematics.[9]

[8] DRC (Dec. 13, 2017). *Evaluation of Minimum Legal Requirements_Questions_DRC_SCDE 121317,* p. 1 [DRC or SCDE Response to Questions].
[9] SCDE (2017a). *SC READY Score Report User's Guide* [SRUG], Columbia, SC: Author, p. 11.

However, as can be seen from the comparison data in Table 1, it matters which three other states are providing the data. The Achieve Report evaluated 24 states that had revised their content standards after having previously adopted the Common Core. According to Achieve, most, but not all, of the revisions in these states retained critical CCR competencies.[10] There were states that scored both higher and lower than South Carolina in this regard. For ELA and Mathematics, three regionally-diverse states fully addressing the key CCR elements identified by Achieve and three regionally-diverse states with significant deficiencies identified by Achieve are included in the right-hand columns of Table 1. Given the differences in ratings presented in Table 1, the content standards in neither group of higher- or lower-rated states would likely be judged to align well with South Carolina's content standards with respect to the key CCR elements identified by Achieve. As illustrated by these examples, without knowing which other states are included in the user norms, or having alignment and descriptive demographic data, one cannot fully evaluate the sufficiency of the comparability of those states' content standards to those of South Carolina or the regional diversity of those states compared to South Carolina. This information is currently unavailable because the contractor considers it proprietary.

Lexiles® and Quantiles®

Other information reported for students administered the SC READY ELA and Mathematics tests provides a different comparison of student achievement relative to the performance of students at the same grade level in other states. MetaMetrics® has mapped a variety of reading texts from each grade level to a common scale called lexiles® that are reported as a number followed by an "L" designation. A student's lexile® score estimates the level of reading material where the student can expect to achieve approximately 75% comprehension. In South Carolina, separate ELA Reading subscores are reported and used to calculate a student's corresponding lexile® interval of 100L below to 50L above the student's lexile® measure, a range of reading texts most appropriate for the student's current level of reading comprehension. A student's lexile® interval can be compared to the range of lexile® scores for typical materials at the student's grade level to evaluate whether the student's reading level is sufficient for the nationally-representative, grade-level texts calibrated by MetaMetrics® that the student is likely to encounter.[11]

MetaMetrics® has also developed similar scores called quantiles® for quantifying the complexity of mathematics instructional materials typically encountered by students at each grade level. Quantiles® are calculated based on SC READY Mathematics Total scores and reported as a number followed by a "Q" designation. A student's quantile® interval, indicating the range of mathematical materials that are most appropriate for that student, consists of 50Q above and 50Q below the student's quantile® measure. For example, if a student's quantile® measure is 1050Q, the student's quantile® interval is (1000Q-1100Q). Students can expect to achieve approximately 50% success with mathematics materials at their quantile® scores.[12]

Based on volunteer user norms, percentile ranks (PRs) corresponding to students' lexile® and quantile® scores are also reported. The map in Figure 1 from the MetaMetrics® website

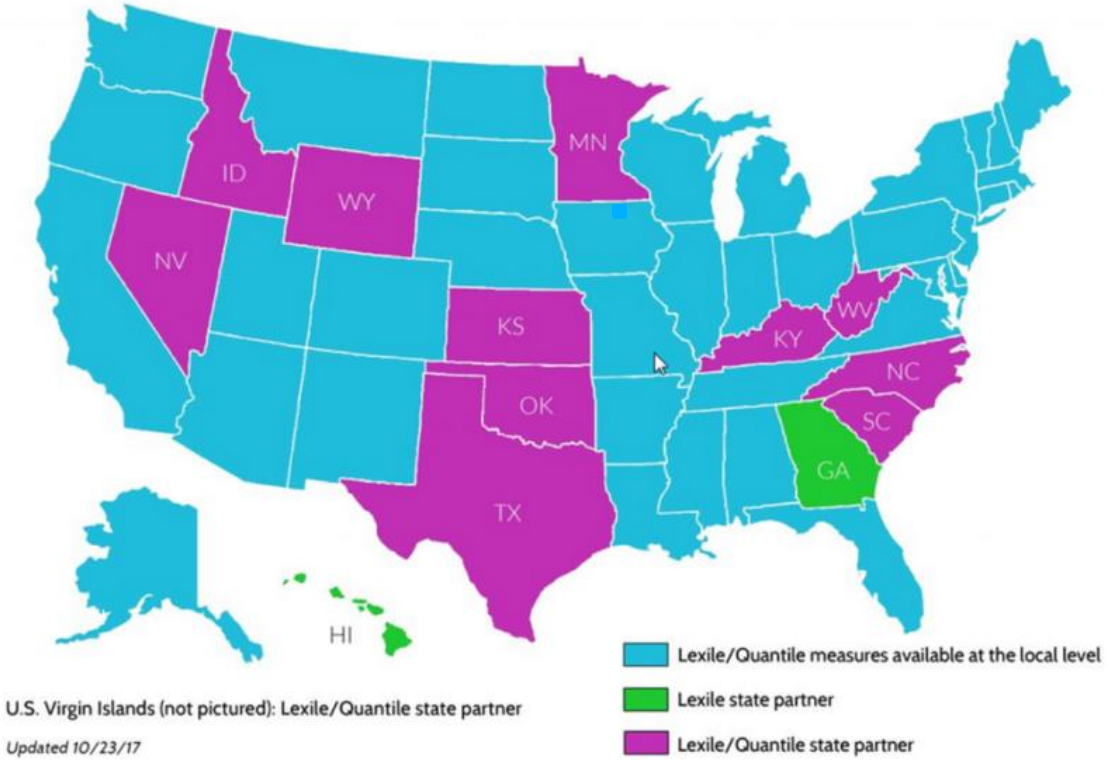
[10] Achieve Report, *supra* note 5, p. 4, 28.

[11] MetaMetrics® (Nov. 2017). *Linking the South Carolina Ready Reading and South Carolina EOCEP English I Assessments with The Lexile® Framework for Mathematics,* Durham, NC: Author [*Lexile® Linking Study*]; www.Lexile.com; *Score Report User's Guide,* p. 14; *see* Exhibit C.

[12] MetaMetrics® (Oct. 2017). *Linking the South Carolina Ready Mathematics and South Carolina EOCEP Algebra I Assessments with The Quantile® Framework for Mathematics,* Durham, NC: Author [*Quantile® Linking Study*]; *Score Report User's Guide,* p. 14; www.Quantiles.com; *see* Exhibit C.

indicates the partner states that likely were included in the norm groups used to derive the lexile® (green plus purple states) and quantile® (purple states) reported percentile ranks.[13]

Figure 1. Likely States in Lexile®/Quantile® Norm Groups


ID
WY
MN
NV
KS
OK
TX
KY
WV
NC
SC
GA
HI
🟦 Lexile/Quantile measures available at the local level
🟩 Lexile state partner
🟪 Lexile/Quantile state partner

U.S. Virgin Islands (not pictured): Lexile/Quantile state partner

Updated 10/23/17

A graphical presentation is used on the individual score report to show the student's lexile® and quantile® scores relative to the ranges identified for each grade level. Sample SC READY Lexile® and Quantile® score reports showing lexile®/quantile® ranges, percentile norms, and grade level ranges are presented in Exhibit A.

On the Lexile® Sample Report, the lexile® range is (1115L-1265L), the norm percentile is 79%, and the Grade 6 range is shown in yellow shading above the Grade 6 label on the horizontal axis (approximately 950L-1050L). Analogous statistics on the Quantile® Sample Report for the same student indicate a quantile® range of (815Q-915Q), a norm percentile of 60%, and a Grade 6 range of approximately (700Q-900Q). From this information, one can infer that Edward's reading and Mathematics abilities are above average (PRs above 50%) and more than sufficient for grade level work.

Linking studies were conducted to derive the corresponding lexile® and quantile® scores for the SC READY Reading and Mathematics scores. SCDE selected a sample of about 2,000 students (<5%) per grade from 100 schools to be administered a separate linking form of 30-40 items in one class period within two weeks of operational testing. Gender and ethnic demographics were similar to state students in Grades 3-8. The SC READY and linking form scores for the sampled students correlated 0.84 for Reading and 0.88 for Mathematics providing sufficient similarity for linking the two score scales. A small number of students with scores at

[13] *Source:* www.MetaMetricsInc.com.

the extremes or misfitting were removed from the analysis. Concurrent calibrations with linear equating methods were used to produce the score correspondences.[14]

Performance on Other Achievement Tests

According to the SC READY Technical Manual, "Efforts were made to align South Carolina standards with the national standards of the National Assessment of Educational Progress (NAEP), the National Council of Teachers of Mathematics, the National Council of Teachers of English, the Third International Mathematics and Science Standards … The Common Core State Standards, the 2014 ACT College and Career Readiness Standards, and the SAT test specifications."[15] Because South Carolina was originally part of the Smarter Balanced Assessment Consortium and had adopted the Common Core content standards for English Language Arts and Mathematics, the new 2014 South Carolina ELA Content Standards and the new 2016 South Carolina Mathematics Content Standards substantially overlap with the content of the Common Core that had been adopted by a majority of states. Thus, SC READY assessments measure content that is similar to that of many other states, but not identical as indicated in the Achieve Report described earlier and summarized in Table 1.

ACT Aspire® and NAEP. With respect to linkage to other assessment scales, some information from nationally-recognized assessments was utilized in the standard setting process for the SC READY assessments, including ACT Aspire® and the National Assessment of Educational Progress (NAEP). ACT Aspire® scores are linked to ACT Assessment scores that include CCR benchmarks. NAEP measures reading and mathematics achievement in 4th and 8th grades. A census administration of ACT Aspire® and administration of NAEP assessments to a sample of South Carolina students were conducted in 2015.

During the vertical moderation phase of the SC READY standard setting, impact data from ACT Aspire® and NAEP were considered together with impact data for SC READY as panelists made their adjustments. Table 2 presents the ACT Aspire® and NAEP impact data provided to the vertical moderation panelists for Grades 4 and 8 along with the SC READY 2016 impact data from the educator panels' recommended cut scores and the actual SC READY 2017 impact data using the final SC READY cut scores. The impact data presented in Table 2 are the percents of students scoring in each of the labeled levels. The impact data for Level 3+4 are presented graphically in Chart 1.

When adjustments to the estimated impact results (shown in Table 2 as SC READY 2016) were considered, of most importance was the similarity between the ACT Aspire® Levels 3 and 4 (Ready and Above) and SC READY Levels 3 and 4 (Meets and Exceeds Expectations) because the ACT Aspire® Level 3 (Ready) cut scores for Grade 8 had been linked to being on track for achieving the ACT Assessment CCR benchmarks.[16] When the vertical moderation panel made adjustments to the impact data for the cut scores recommended by the educator panels, the new impact estimates generally moved closer to the ACT Aspire® values. As indicated in the final column of Table 2 and Chart 1, the actual SC READY 2017 Level 3+4 impact data were within 7-9 percentage points of the corresponding ACT Aspire® values for ELA and within 3-4 percentage points for Mathematics. Differences from NAEP impact data were slightly larger.

[14] *Lexile® Linking Study, supra* note 9, p. 22-28; *Quantile® Linking Study, supra* note 10, p. 32-39.

[15] SCDE (2017b). *Technical Documentation for the 2017 South Carolina College- and Career-Ready Assessments – ELA and Mathematics* [Technical Manual], Columbia, SC: Author, p. 7-8.

[16] SCDE (2016a). *Standard Setting Report Addendum,* Columbia, SC: Author.

Table 2. Comparison of ACT Aspire®, NAEP and SC READY 2016 & 2017 Impact Data

	LEVEL 1	LEVEL 2	LEVEL 3 + 4
ELA Grade 4			
ACT Aspire® 2015 Reading	36	31	**32**
NAEP 2015 Reading	35	31	**33**
SC READY 2016 ELA	23	22	**54**
SC READY 2017 ELA	*30*	*30*	***41***
ELA Grade 8			
ACT Aspire® 2015 Reading	29	24	**47**
NAEP 2015 Reading	29	44	**28**
SC READY 2016 ELA	26	23	**51**
SC READY 2017 ELA	*28*	*32*	***40***
MATH Grade 4			
ACT Aspire® 2015	9	42	**50**
NAEP 2015	21	43	**37**
SC READY 2016	32	29	**39**
SC READY 2017	*24*	*30*	***46***
MATH Grade 8			
ACT Aspire® 2015	39	29	**32**
NAEP 2015	35	40	**25**
SC READY 2016	41	34	**25**
SC READY 2017	*32*	*34*	***35***

Source: DRC, Document C1a.pdf, TAC Webinar, June 28, 2016; Technical Manual, p. 35.

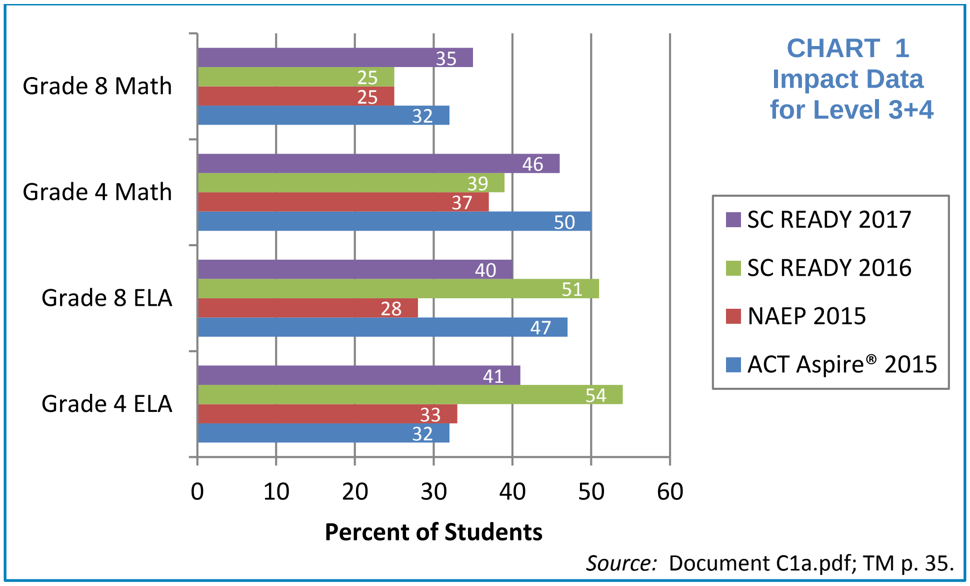


CHART 1
Impact Data
for Level 3+4

Source: Document C1a.pdf; TM p. 35.

NWEA Study. A 2015 study of South Carolina students who were administered SC READY and the NWEA Measures of Academic Progress (MAP) test created concordance tables. The samples consisted of 78,320 ELA students and 78,063 Mathematics students in Grades 3-8 (samples of approximately 20% to 25% of South Carolina students). MAP proficiency classifications matched SC READY proficiency levels (meets + exceeds expectations) 84-86% for ELA and 86-89% for Mathematics. Data on the representativeness of the samples, content similarity of the tests and test reliabilities were not reported.[17]

Evaluation

The comparability of the content standards of the three “other states with comparable standards” to the South Carolina content standards is based primarily on the similarity of the current content standards in these states to the Common Core content standards they originally adopted and then adapted (except one that kept Common Core). It is also a convenient sample of performance from concurrent users of the items in the contractor’s CCR item bank. However, other than the Achieve Report, no current alignment studies appear to be available to confirm the degree of similarity between the South Carolina content standards and the Common Core or the content standards of the other user states. Judging by the data presented in the Achieve Report for the 24 states that adopted the Common Core content standards and then revised them, it matters which three states constitute the user group with South Carolina. In addition, South Carolina contributes approximately 25% of the user group data, so strictly speaking, the resulting “other states” percentile ranks do not reflect results independent of South Carolina. In any case, a sample from four user states is probably too small and unrepresentative to derive percentile ranks that accurately reflect national or Common Core CCR norms.

Basically, the percentile ranks reported for SC READY represent user norms for a small, volunteer sample of states that is undefined and whose characteristics are currently unknown. As a result, it is difficult to interpret with any certainty what the reported percentile ranks represent in terms of performance relative to states with Common-Core-like content standards or students in the United States as a whole. It would be helpful for the contractor to provide South Carolina with additional demographic information about the participating states and descriptions of the concurrent calibrations used to derive the reported percentile rank norms. It might also be more informative if the “other states with similar standards” percentile ranks were calculated independently using only the data from the other three states in the user group.

The lexile® and quantile® linking studies describe the user norm groups as including students from 51 (reading) or 38 (mathematics) states (full state/districts/territories) who tested from 2010 to 2016. The number of states represented likely includes the 13 partner states shown on the map in Figure 1. The other states that are represented are contributing an unknown number of students from only certain districts. Although this sample of over 3.5 million students is much larger and more geographically diverse compared to the “other states” user sample, it still may not be representative of students nationally and no claim is made about the similarity of content standards. Less than 50% (reading) or 30% (mathematics) of the students in the sample provided demographic information for comparison with national or South Carolina statistics.

The linking studies also state that the user norms were validated with a longitudinal sample of over 100,000 students. This sample may have been the same 2007 census data from North Carolina followed longitudinally for several years and used to develop the reported CCR growth trajectories discussed in the section for Legislative Criterion 4.[18] If so, the percentile norms are valuable indicators to the extent North Carolina students are judged to be similar to South Carolina students and/or to students nationally. It would be helpful if gender, ethnic, SD and EL data were available to judge the representativeness of the full lexile®/quantile® user samples.

[17] Chapter 5 (Task 5).

[18] *Lexile Linking Study, supra* note 9, p. 39; *Quantile Linking Study, supra* note 10, p. 51; *see* Legislative Criterion 4.

In sum, comparative information is available for two volunteer user groups from two different contractors. Limited information about the demographics of these user samples makes it difficult to judge their representativeness. The contractors appear to consider detailed information about the specific states included in the samples and the procedures used to develop the reported percentile norms proprietary information unavailable to customers. With incomplete information, it is difficult to evaluate the quality or sufficiency of the evidence for Legislative Criterion 1. On the other hand, these data may be the best available and do provide some useful comparative information. The primary available alternative, participation in a consortium of states using common content standards and common assessments (e.g., Smarter Balanced, PARCC), has already been attempted and discarded, and it may no longer be feasible politically or financially.

Similarly, the NWEA study provides comparative data based on yet another and different volunteer sample of users. While its findings may have some usefulness for those districts that administer the MAP tests, there appears to be neither alignment data relating MAP test content to the South Carolina state content standards nor any claim that the MAP and SC READY tests are comparable. The reported predictions may be more a function of common ELA or mathematics ability and less an indicator of achievement of the specific knowledge and skills embodied in the South Carolina state content standards.

2. Development of a System of Summative, Vertically-Scaled, Benchmarked, Standards-Based Assessments

Evidence

Evidence relevant to Legislative Criterion 2 includes a description of the SC READY system of summative assessments, vertical scaling of the assessments, benchmarking of the assessments and the standards-based feature of the SC READY assessments.

System of Summative Assessments

The glossary from the 2014 *Test Standards* defines *summative assessment* as follows:

> **summative assessment:** The assessment of a test taker's knowledge and skills typically carried out at the completion of a program of learning, such as the end of an instructional unit.[19]

In this case, the typical test taker is a student who has completed a grade level, standards-based curriculum for a full school year. The SC READY ELA and Mathematics tests are summative assessments because they test the knowledge and skills from the appropriate grade level content standards (the school-year curriculum) and are required to be given within the last 30 days of a school district's calendar.[20] Scheduling the SC READY assessments within the last month of school allows for the maximum possible instructional time for teachers to cover the tested state content standards. Ensuring maximum curriculum coverage prior to testing allows the SC READY assessments to measure the sum of the student's learning for that school year.

The test blueprints demonstrate a systematic plan for representing content with similar numbers of items and subarea content in adjacent grades. The SC READY ELA and Mathematics test blueprints presented in Exhibit B provide target ranges for the number of desired test items for each

[19] *Test Standards, supra* note 4, p. 224.
[20] *Note:* According to the SCDE website, beginning in 2018 the SC READY tests must be administered within the last 20 days of a school district's calendar.

reporting category. These blueprints, created by SCDE and the testing contractor, were derived from the state content standards for the respective subjects and grades. Information from the HumRRO alignment studies discussed below provides support for the proposition that the system of SC READY tests for six grades (3-8) and two fundamental school subjects (ELA and Mathematics) systematically covers the breadth (sum) of the corresponding state content standards.

Standards-Based

For a standards-based test to be consistent with professional standards, the test must be valid and reliable for its intended score interpretations. The primary source of validity evidence for a standards-based test is content validity. Content validity evidence includes alignment of the test items to the state standards and test blueprints, and item quality data.

To validly measure the intended content, a standards-based test must also be reliable. Reliability data for a standards-based test typically include reliability estimates, standard errors, decision consistency estimates, and conditional standard errors at the cut scores. Supplementary validity evidence for a standards-based test may include subscore intercorrelations that quantify the degree to which the variation among subscores is attributable to common versus unique variance.

The following sections describe the alignment, reliability, and intercorrelation validity evidence for the SC READY assessments. Validity evidence for item quality is presented in the section addressing Legislative Criterion 6.

Alignment. Content representation is the primary factor used to select items for each SC READY standards-based assessment. The test blueprint for each subject and grade level is based on the state content standards describing the knowledge and skills students are expected to learn in that subject at that grade level. The weights assigned to each subarea within a subject/grade test blueprint generally reflect the relative importance and emphasis placed on that content within the corresponding state standards. An alignment review provides one type of evidence supporting the validity of the content representation of an assessment by evaluating the degree to which these goals have been achieved.

The purpose of an alignment review is to determine whether the content of the test items appropriately matches the depth and complexity of the knowledge and skills specified in the test blueprints and state content standards. To make this determination, HumRRO convened a series of educator panels to provide expert judgments for the following three alignment criteria for content standards, test blueprints and test items:

1. **Alignment between the *test blueprint* and the *state content standards* –** qualitative judgments of the degree to which the test blueprint adequately covers the knowledge and skills contained in the state content standards;
2. **Alignment between the *test items* and the *test blueprint* –** a comparison of the actual numbers of test items measuring each reporting category with the ranges specified in the test blueprint;
3. **Alignment between the *test items* and the *state content standards* –** qualitative judgments of whether the content of an item is fully, partially or not aligned to the content standard it is intended to measure.

Four different educator panels of 5-6 educators each rated the items on the 2017 operational test forms for ELA Grades 3-5, ELA Grades 6-8, Mathematics Grades 3-5 and Mathematics

Grades 6-8. Detailed descriptions of the qualifications and training of panel members and the methods used to obtain their judgments are provided in Chapter 2 of this report.

Ratings were averaged across panelists for reporting. An indicator of interrater reliability is provided by the correlation coefficient for independent panel member depth of knowledge ratings (discussed below in the section for Legislative Criterion 6). Across subjects and grades, and except for Grade 6 Mathematics at 0.75, these correlations ranged from 0.81 to 0.98.[21] Values greater than 0.70 are considered acceptable; values greater than 0.80 are very good. The remainder of this section summarizes the results from these alignment reviews that are most relevant to Legislative Criterion 2.

Alignment of the ELA test blueprint to the content standards. Based on a holistic discussion of the link between the test blueprint and the content standards, panelists agreed that overall the Grades 3-8 test blueprints adequately cover what students should know and be able to do as specified by the state content standards. However, several suggestions for improvement were offered and are summarized in Table 3.

Table 3. Alignment Results for SC READY Grades 3-8 ELA‡

ELA	Blueprint to Standards	Items to Blueprint			Items to Standards
			Mean # linked	Target #	Items partially + fully aligned
Grade 3	**Overall adequate link –** Inquiry difficult to assess with test format; delete and redistribute items to word analysis and phonics	Read Lit Read Info Writing Inquiry	19.2 18.8 21.0 10.0	19 19 30*	100%
4	**Overall adequate link –** Inquiry difficult to assess with test format; delete and redistribute items to word analysis and phonics	Read Lit Read Info Writing Inquiry	19.0 18.8 25.0 6.0	19 19 30*	97%
5	**Overall adequate link –** Inquiry difficult to assess with test format; delete and redistribute items to word analysis and phonics	Read Lit Read Info Writing Inquiry	19.0 19.0 24.0 7.0	19 19 30*	99%
6	**Overall adequate link –** Inquiry difficult to assess with test format; add communication skills; vary weights in Grades 6-8 to reflect growing skills; 6/7 similar	Read Lit Read Info Writing Inquiry	21.8 29.0 22.0 8.0	21 29 30*	96%
7	**Overall adequate link –** Inquiry difficult to assess with test format; add communication skills; vary weights in Grades 6-8 to reflect growing skills; 6/7 similar	Read Lit Read Info Writing Inquiry	20.8 28.8 23.5 6.7	21 29 30*	94%
8	**Overall adequate link –** Inquiry difficult to assess with test format; add communication skills; vary weights Grades 6-8 to reflect growing skills	Read Lit Read Info Writing Inquiry	21.0 29.0 24.0 7.0	21 29 30*	99%

* The TDA essay item (16 points) was not evaluated in this analysis and the total number of objective items is 30;
‡ HumRRO recommended that South Carolina content experts re-examine the themes from panel comments; *Source:* Chapter 2 (Task 2).

[21] *See* Chapter 2 (Task 2).

Alignment of the ELA items to the test blueprint. When rounded, the mean number of items linked to each domain (reading literary text, reading informational text, writing plus inquiry) by panelists was equal to the number of items specified in the test blueprint. In addition, when analyzed by the seven reporting categories, the mean number of linked items for each grade level fell within the range specified by the test blueprint.

Alignment of the ELA items to the content standards. For the ELA tests, as indicated in Table 3, nearly all the items were judged to be partially or fully aligned to the content standards. For Grades 3-5 and 8, 96%-99% of the items were rated fully aligned. Only in Grades 6 and 7 did the percent of fully aligned items drop slightly below 90%, with 4% and 6% of the items, respectively, judged not aligned.[22]

Similar analyses were conducted for the Mathematics tests. The results are presented in Table 4.

Table 4. Alignment Results for SC READY Grades 3-8 Mathematics‡

MATH	Blueprint to Standards	Items to Blueprint			Items to Standards
			Mean # linked	Target #	Items partially + fully aligned
Grade 3	**Overall weak link –** More emphasis on foundational numbers and fractions; greater variety of graphing data items; overuse of interpreting bar graphs	Numbers Fractions Alg Ideas Geometry Data Anal	7 8 13 9 13	7 - 9 7 - 9 13-16 7 - 9 13-16	96%
4	**Overall adequate link –** Covers what students should know and be able to do as specified in the content standards	Numbers Fractions Alg Ideas Geometry Data Anal	12 12 12 9 11	10-12 11-14 11-14 8-10 11-14	100%
5	**Overall weak link –** Increase items to 11-14 for first three categories and reduce to 10-12 for last two categories to reflect relative number and complexity of standards	Numbers Fractions Alg Ideas Geometry Data Anal	10 12 13 10 11	10-13 10-12 10-13 10-12 11-14	100%
6	**Overall weak link –** Weight first three categories more (25% each) and last two categories less (12.5% each)	Numbers Ratio/Prop Alg Eq/Ineq Geom/Meas Data/Stat	14 10 14.8 9 11.7	12-15 8-10 12-15 8-10 11-13	100%
7	**Overall weak link –** Weight first three categories more (25% each) and last two categories less (12.5% each)	Numbers Ratio/Prop Alg Eq/Ineq Geom/Meas Data/Stat/Prob	13 10 12 12 13	13-15 8-10 12-14 11-13 13-15	98%
8	**Overall weak link –** Less weight first and last categories and more weight on the middle three categories	Numbers Functions Alg Eq/Ineq Geom/Meas Data/Stat/Prob	9 13.8 16.2 14 9	9-11 11-14 12-16 12-16 9-11	97%

‡ HumRRO recommended that South Carolina content experts re-examine the themes from panel comments; *Source:* Chapter 2 (Task 2).

[22] Chapter 2 (Task 2).

Alignment of the Mathematics test blueprint to the content standards. Based on a holistic discussion of the link between the test blueprint and the content standards, panelists agreed that overall the Grade 4 blueprint adequately covers what students should know and be able to do as specified by the standards. However, panelists judged the link to be weak for the other grades. Suggestions for improvement were offered for each grade and are summarized in Table 4. Specifically, the panelists felt the weighting of items by reporting categories did not adequately reflect the number and complexity of the standards in each category and suggested alternative weightings.

Alignment of the Mathematics items to the test blueprint. When rounded, the mean number of items linked to each reporting category (e.g., for Grade 3, number sense/base ten, fractions, algebraic thinking/operations, geometry, measurement/data analysis) by panelists was within the target range of items specified in the test blueprint.

Alignment of the Mathematics items to the content standards. For the Mathematics tests, as indicated in Table 4, nearly all the items were judged to be partially or fully aligned to the content standards. Across grades, 90% or more of the items were fully aligned. There were no nonaligned items in Grades 4-6 and no more than 4% of the items were nonaligned in the other three grades.

Reliability. To be valid indicators of mastery of the standards-based content embodied in the state content standards, test scores must also be reliable. Reliability estimates quantify the degree to which scores are replicable, that is, the confidence one has that if a student were to retest under similar conditions, the student's new score would be substantially similar to the original score.

The metric used to quantify the reliability of SC READY ELA and Mathematics test scores is based on a single administration of the test. Reliability estimates are decimal numbers that range from 0 to 1, with larger values indicating greater reliability. The South Carolina Assessment TAC recommended a minimum reliability of .85 for the SC READY assessments, a commonly-cited target when the test scores are being used to make decisions about individual students. The reliability estimates for the 2017 SC READY assessments by group (total, gender, ethnic, English learners, students with disabilities) are presented in Table 5.[23]

For all students administered an SC READY ELA test, the average estimated reliability was .94; for Mathematics it was .93. The range of average reliability estimates across groups was .90 to .94 for ELA and .86 to .93 for Mathematics. Out of 96 reliability estimates reported, only eight (8%) fell below .90. Of those eight, only two (Mathematics for students with disabilities in Grades 7 and 8) fell below the recommended .85. Overall, Mathematics reliabilities tended to be slightly lower than those for ELA. The average test reliabilities by group are presented graphically in Chart 2.

All of the total group reliabilities exceeded .90. All of the gender reliabilities also exceeded .90 and were nearly identical by subject and grade level. Ethnic reliabilities were also nearly all above .90 and very similar. Only the reliabilities for African-Americans in Grades 4, 5, 7, and 8 Mathematics were slightly below .90. All reliabilities for English learners were at or above .90. However, reliabilities for students with disabilities were somewhat lower, especially in Grades 7 and 8 Mathematics where they fell below 0.85.

[23] Technical Manual, *supra* note 13, p. 42. Total group conditional standard errors at the cut scores are also reported and ranged from 23.6 to 27.9 for ELA and 27.4 to 32.0 for Mathematics on the vertical scale score metric.

*Table 5. 2017 SC READY Reliabilities by Subject, Grade and Group**

	Total	Gender		Ethnic				
		F	M	AA	H	W	EL	SD
ELA Grade 3	**.92**	.92	.92	.90	.91	.92	.91	.90
4	**.93**	.93	.93	.92	.92	.93	.92	.92
5	**.94**	.93	.94	.91	.93	.93	.93	.90
6	**.95**	.95	.95	.93	.94	.95	.94	.89
7	**.94**	.94	.94	.92	.94	.94	.93	.89
8	**.94**	.94	.95	.93	.94	.94	.93	.90
Average	**.94**	**.94**	**.94**	**.92**	**.93**	**.94**	**.93**	**.90**
MATH Grade 3	**.92**	.92	.93	.90	.91	.92	.91	.91
4	**.93**	.92	.93	.89	.91	.92	.92	.89
5	**.93**	.92	.93	.89	.91	.93	.92	.87
6	**.93**	.93	.94	.90	.92	.93	.93	.86
7	**.92**	.92	.92	.87	.90	.92	.91	***.79***
8	**.92**	.92	.92	.88	.91	.92	.90	***.81***
Average	**.93**	**.92**	**.93**	**.89**	**.91**	**.92**	**.92**	**.86**

* AA=African-American; H=Hispanic; W=White; EL=English learners; SD=students with disabilities; Raw score reliabilities were estimated using WINSTEPS' Rasch Student Reliability. *Source:* Technical Manual, p. 42.

ELA Reading. Reliabilities for the ELA Reading subscore, reported for all grades and used in part in Grade 3 for deciding whether students should attend a remedial summer camp, are presented in Table 6. As indicated in Table 6, the ELA Reading score is based on 38 items in the lower elementary grades and 50 items in the upper middle school grades, a subset of 56% and 63%, respectively, of the total ELA objective test items. Because subscores are based on fewer items, they typically have somewhat lower reliabilities than total test scores. Nonetheless, except for students with disabilities at .84, the average group reliabilities shown at the bottom of Table 6 meet the Assessment TAC guideline of .85, and many of the individual grade/group values also do. The reliabilities for African-Americans, Hispanics and English learners in Grades 3 and 4, and students with disabilities in Grades 3, 5, and 7, are all very close, ranging from .82 to .84. The average SC READY Reading reliabilities for these groups are summarized graphically in Chart 2.

CHART 2
SC READY Grades 3-8 Average Test Reliabilities by Group

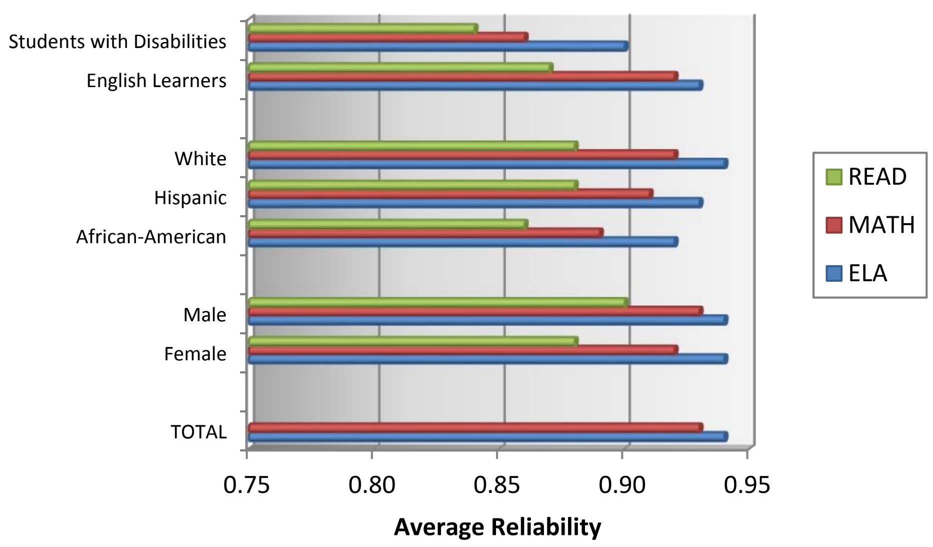


*Table 6. 2016 SC READY Reliabilities For ELA Reading**

ELA READING	NUMBER OF ITEMS	GENDER		ETHNIC				
		F	M	AA	H	W	EL	SD
Grade 3	38	.86	.87	***.82***	***.84***	.86	***.84***	***.84***
4	38	.85	.87	***.83***	***.84***	.85	***.84***	.85
5	38	.87	.89	.85	.87	.87	.87	***.84***
6	50	.91	.92	.88	.90	.91	.90	.85
7	50	.90	.91	.87	.90	.90	.89	***.82***
8	50	.90	.92	.89	.90	.90	.89	.85
	Average	**.88**	**.90**	**.86**	**.88**	**.88**	**.87**	***.84***

* 2017 data were not available; AA=African-American, H=Hispanic, W=White, EL=English learners, SD=students with disabilities; Reliabilities are raw score Kuder-Richardson formula 21 (KR_{21}) internal consistency estimates. *Source:* DRC, Table CE3.1A.2b.

Rater Agreement. The reliability estimates presented in Table 5, Table 6 and Chart 2 are for the objectively scored items. The SC READY ELA assessment also includes a text-dependent analysis (TDA) essay item that is scored by two raters. The reliability of ratings supplied by human raters is quantified by the percent of exact and adjacent agreement between the two scores for the same responses. Ratings that differ by more than one point (nonadjacent scores) are resolved by a third rater. Rater agreement data by grade level are presented in Table 7.[24]

As the data in Table 7 indicate, the SCDE requirement for at least 70% exact agreement was met in all grades and 98-99% of the scored TDA items required no resolution. These data confirm that the quality control scoring procedures utilized by the contractor for the TDA items were successful and produced reliable scores. However, mean scores were quite low, ranging between *1=minimal text analysis with inadequate writing* and *2=limited text analysis with inconsistent writing*.

For subgroups, the SCDE requirement for at least 70% exact agreement was met in all cases and often significantly exceeded. Exact agreement was consistently a bit lower than the grade-level average for females and Whites, but less than 2% of all responses in all groups and grades required resolution by a third rater.

Subscore Intercorrelations. The term *subscore* usually refers to any subset of items reported as a separate score. However, for the SC READY tests, the ELA Reading score is considered a subscore and the other scores formed by subsets of items are referred to as reporting category scores.

SC READY reporting category scores are expected to share some common variance because they are part of a unidimensional construct of ELA or Mathematics. The Rasch model used to analyze the SC READY assessments assumes unidimensionality of the construct that is being tested. This assumption is usually verified with factor analyses that confirm a large first factor and much smaller subsequent factors. If subscores are to be meaningfully interpreted as indicating relative strengths and weaknesses, they should exhibit sufficient unique variance to be considered distinguishable. For example, if two subscores intercorrelate at .95, 90% of the variation measured is common. This indicates that they are measuring almost the same skills and having two scores is redundant. Alternatively, if two subscores intercorrelate at .50, only 25% of the variation in their scores is common and they are measuring markedly different skills.

Average intercorrelations and the percent of common variance for the major reporting category scores for the SC READY ELA and Mathematics assessments in Grades 3-8 are presented in Tables 8, 9 and 10. Pearson correlations are reported on the upper diagonal and the percent of common variance is reported on the lower diagonal. The percent of common variance is calculated by squaring the intercorrelation value for two test scores, multiplying by 100 and rounding to the nearest whole number. For example, using the data from Table 8 for the ELA writing and inquiry reporting categories, the percent of common variance equals

$$\text{(writing/inquiry intercorrelation)}^2 \text{ x } 100 = (.66)^2 \text{ x } 100 = 44\%.$$

[24] Technical Manual, *supra* note 13, p. 31.

*Table 7. Rater Agreement for SC READY ELA Text-Dependent Analysis (TDA) Essay Items**

Grade Group	Mean	SD	Exact Agreement	Adjacent Agreement	*Exact+Adjacent Agreement*	Resolved by a Third Rater
Grade 3	**1.6**	0.6	75%	24%	***99%***	1.6%
M	1.5	0.6	77%	22%	***99%***	1.1%
F	1.6	0.6	73%	25%	***98%***	1.6%
H	1.5	0.6	76%	23%	***99%***	0.8%
AA	1.4	0.5	78%	21%	***99%***	1.1%
W	1.6	0.6	73%	25%	***98%***	1.5%
SD	1.3	0.5	84%	16%	***99%***	0.5%
EL	1.5	0.6	77%	22%	***99%***	0.9%
Grade 4	**1.2**	0.4	85%	14%	***99%***	0.6%
M	1.2	0.4	88%	12%	***99%***	0.4%
F	1.3	0.5	83%	16%	***99%***	0.8%
H	1.2	0.4	88%	12%	***99%***	0.4%
AA	1.1	0.3	89%	10%	***99%***	0.3%
W	1.3	0.5	83%	16%	***99%***	0.8%
SD	1.1	0.3	93%	7%	***99%***	0.2%
EL	1.2	0.4	88%	12%	***98%***	0.5%
Grade 5	**1.4**	0.6	73%	26%	***99%***	1.0%
M	1.4	0.5	76%	23%	***99%***	0.7%
F	1.5	0.6	70%	29%	***99%***	1.3%
H	1.4	0.5	75%	25%	***99%***	0.8%
AA	1.3	0.5	78%	22%	***99%***	0.6%
W	1.5	0.6	70%	28%	***98%***	1.3%
SD	1.1	0.3	88%	12%	***99%***	0.2%
EL	1.4	0.5	75%	24%	***99%***	0.9%
Grade 6	**1.4**	0.5	78%	21%	***99%***	1.3%
M	1.3	0.5	80%	19%	***99%***	1.0%
F	1.4	0.6	75%	23%	***98%***	1.6%
H	1.3	0.5	80%	18%	***98%***	1.3%
AA	1.2	0.4	83%	16%	***99%***	0.8%
W	1.5	0.6	74%	24%	***98%***	1.6%
SD	1.1	0.3	90%	9%	***99%***	0.4%
EL	1.3	0.5	80%	19%	***99%***	1.3%
Grade 7	**1.7**	0.7	75%	24%	***99%***	0.7%
M	1.6	0.6	77%	22%	***99%***	0.6%
F	1.8	0.7	73%	26%	***99%***	0.9%
H	1.6	0.6	77%	22%	***99%***	0.6%
AA	1.5	0.6	78%	22%	***99%***	0.6%
W	1.8	0.7	74%	25%	***99%***	0.9%
SD	1.2	0.4	86%	13%	***99%***	0.3%
EL	1.5	0.6	78%	21%	***99%***	0.4%
Grade 8	**2.0**	0.8	71%	27%	***98%***	1.6%
M	1.8	0.8	73%	26%	***99%***	1.4%
F	2.1	0.8	70%	28%	***98%***	1.9%
H	1.9	0.8	72%	26%	***98%***	1.6%
AA	1.7	0.7	73%	25%	***98%***	1.5%
W	2.1	0.8	70%	28%	***98%***	1.7%
SD	1.4	0.6	81%	18%	***99%***	0.7%
EL	1.8	0.7	72%	26%	***98%***	1.3%

* These more recent data differ slightly from that presented in the Technical Manual, p. 31; percents may not sum to 100 due to rounding; M=male, F=female, H=Hispanic, AA=African-American; W=White, SD=students with disabilities, EL=English learners;

Source: Response to Questions, Dec. 13, 2017.

Table 8. Average Pearson Correlations (Upper Diagonal) and Percents of Common Variance (Lower Diagonal) for SC READY Grades 3-8 ELA Reporting Category Scores

ELA	Reading Literary Text	Reading Informational Text	Writing	Inquiry
Reading Literary Text		.82	.77	.66
Reading Informational Text	67%		.77	.66
Writing	59%	59%		.66
Inquiry	44%	44%	44%	

Source: DRC Statistical Printout, Dec. 12, 2017.

The data in Table 8 indicate that the two types of reading correlate the highest with 67% common variance but both also exhibit fairly high correlations with writing with 59% common variance. The Literary Text and the Informational Text scores are both heavily influenced by general reading ability but about ⅓ of what they each measure is unique. Similarly, about 40% of the skills measured by the writing items are unique. The inquiry items correlate the lowest with both reading and writing scores but still share 44% common variance. Of all the reporting categories listed in Table 8, the inquiry category is the most unique with more than half the variability in its scores accounted for by skills other than reading or writing.

Table 9. Average Pearson Correlations (Upper Diagonal) and Percents of Common Variance (Lower Diagonal) for SC READY Grades 3-5 Mathematics Reporting Category Scores

MATH Grades 3-5	Numbers	Fractions	Algebraic Thinking	Geometry	Measurement Data Analysis
Numbers		.67	.74	.63	.68
Fractions*	45%		.69	.61	.67
Algebraic Thinking	55%	48%		.65	.71
Geometry	40%	37%	42%		.64
Measurement Data Analysis	46%	45%	50%	41%	

* Grade 3 does not include operations; *Source:* DRC Statistical Printout, Dec. 12, 2017.

The correlations of Grades 3-5 Mathematics reporting category scores presented in Table 9 indicate that algebraic thinking and numbers share the greatest common variance at 55% with algebraic thinking and measurement/data analysis close behind at 50%. The geometry reporting category has the most unique variance, ranging from 58% compared with algebraic thinking to 63% when compared with fractions. Most of the reporting categories exhibit common variances of 50% or less indicating that these scores share some common mathematics ability but also are distinguishable by significant amounts of unique variance.

Table 10. Average Pearson Correlations (Upper Diagonal) and Percents of Common Variance (Lower Diagonal) for SC READY Grades 6-8 Mathematics Reporting Category Scores

MATH Grades 6-8	Number System	Algebra‡	Geometry Measurement	Data Statistics Probability	Ratio Proportion (Grades 6-7)	Functions (Grade 8)
Number System		.74	.66	.67	.73	.66
Algebra	55%		.68	.71	.74	.75
Geometry Measurement	44%	46%		.65	.63	.70
Data / Statistics Probability*	45%	50%	42%		.68	.71
Ratio / Proportion (Grades 6-7)	53%	55%	40%	46%		
Functions (Grade 8)	44%	56%	49%	50%		

* Grade 6 does not include probability; ‡ Algebra includes expressions, equations and inequalities;
Source: DRC Statistical Printout, Dec. 12, 2017.

The data for Grades 6-8 Mathematics in Table 10 exhibit similar patterns to those observed for Grades 3-5 Mathematics. Again, the highest correlations are between algebraic skills and numerical skills with algebra (expressions, equations and inequalities) and numbers and algebra and ratio/proportion (Grades 6 & 7) correlating .74 and sharing 55% common variance. Not surprisingly, algebra correlates highest (.75) with functions (Grade 8), a topic typically taught along with more advanced algebraic skills. Also again, the geometry/measurement reporting category exhibits the most unique score variation ranging from 60% unique variance when compared with ratio/proportion (Grades 6 & 7) to 54% compared with algebra. Interestingly, its highest correlation is with functions (Grade 8) at .70, indicating about half shared and half unique score variation. Again, nearly half of the score variation for the Mathematics reporting category scores is unique indicating that they are distinguishable scores worth reporting separately.

SC READY and EOCEP Relationships. Correlations between the SC READY Grade 8 ELA and Mathematics tests and the End of Course Examination Program (EOCEP) English I and Algebra I tests provide evidence of convergent and divergent validity and are presented in Table 11. Demonstrating convergent validity, the ELA/English I and Mathematics/Algebra I tests correlate highly, at .72 and .78, respectively, and share 52% and 61% common variance, respectively. The remaining 48% and 39% of variance, respectively, is unique to each test, likely in part because the EOCEP tests are aligned to more complex content standards than the SC READY tests for the same subjects. These relationships indicate that proficiency on the Grade 8 SC READY tests is likely predictive of success on the corresponding EOCEP tests because convergent validity indicates similar constructs are being measured by both tests.

Table 11. Pearson Correlations (Upper Diagonal) and Percents of Common Variance (Lower Diagonal) for SC READY Grade 8 Tests and End-of-Course (EOCEP) Tests

	SC READY ELA	SC READY Mathematics	EOCEP English I	EOCEP Algebra I
SC READY ELA		.60	.72	.58
SC READY Mathematics	36%		.54	.78
EOCEP English I	52%	29%		
EOCEP Algebra I	34%	61%		

Source: DRC, *SC READY & EOCEP Relationships,* Dec. 8, 2017.

Evidence of divergent validity can be seen in the correlations between SC READY Mathematics/ EOCEP English I and SC READY ELA/EOCEP Algebra I which are lower and exhibit shared variances of only 29% and 34%, respectively. Not surprisingly, these values indicate that the majority of skills assessed by these different subject matter tests are unique. Similarly, the correlation between SC READY ELA and SC READY Mathematics tests is also lower, indicating only 36% shared variance and again demonstrating assessment of largely unique skills. Though relatively small, these disparate subjects still share some common variance that is most likely attributable to general academic ability.

Benchmarked

The glossary from the 2014 *Test Standards* defines *benchmark assessments* as follows:

> **benchmark assessments:** Assessments administered in educational settings at specified times during a curriculum sequence, to evaluate students' knowledge and skills relative to an explicit set of longer-term learning goals.[25]

SC READY test forms for each grade/subject combination are developed to match the test blueprints that in turn align with the state content standards. Items on each test form have been reviewed to ensure a match to the content intended to be measured (content validity described above), universal design to provide accessibility to the widest possible range of test takers, and freedom from any characteristics that might unfairly disadvantage or contain sensitive content for students from different demographic groups. Content tested at each grade level reflects the prerequisite knowledge and skills necessary for success at the next grade level. The state content standards to which the test forms are aligned represent a progression of content that is designed to prepare students to achieve CCR expectations in high school. As already described in the section for Legislative Criterion 1, SC READY score reports provide percentile ranks linking student performance to user group norms for four states with relatively similar content standards and states also reporting lexiles® and quantiles®.

Vertically Scaled

A true vertical scale places the scores from a series of content-related tests (e.g., ELA, Mathematics) across adjacent grade levels (e.g., 3-8) on a common scale so that the scores are comparable from year-to-year as students progress from one grade to the next and take

[25] *Test Standards, supra* note 4, p. 216.

different content-related tests. A common method for developing vertical scales is to administer the same, small set of items to samples of students at adjacent grade levels. It is also common for state testing programs to develop within-grade-level scale scores for reporting student progress across years. Although not true vertical scales, the properties of within-grade-level score scales with carefully-chosen anchor points and boundaries may resemble those of a vertical scale. Both types of score scales have been used with the SC READY assessments.

A three-digit, grade-level scale was developed in Spring 2016 for reporting test results at the student, district and state levels. A four-digit, vertical scale was developed in Spring 2017 for reporting test results at the student, district and state levels. In 2017, the SC READY assessments were horizontally equated to the grade-level scale for reporting only at the state level. Current plans are to continue reporting the four-digit vertical scale scores in 2018 and beyond.[26] The next sections describe the SC READY 2017 vertical scale and the 2016 grade-level scale in greater detail and compare their respective properties.

2017 Vertical Scale. A vertical scaling study for SC READY was conducted in which students from Grades 4-8 were administered a sample of ELA and Mathematics items from the adjacent lower grade level. About 15-18 items were chosen per grade level that were representative of the content in the lower grade level and assessed skills likely to have been reviewed and practiced at the adjacent upper grade level.[27] Using a Rasch model analysis,[28] a vertical scale was constructed for the 2017 SC READY tests reflecting the relationships between the performance on those common items at the lower and next higher grade levels.

Grades 5 and 6 were scaled together first, and then the other grades were linked in turn to the common scale via appropriate equating constants. The ability measures for each grade on the common scale were then transformed so that the range of scale scores for each grade began at 100. The maximum scale score for third grade was fixed at 825 and increased by 25 scale score points at each successive grade. Thus, the range of vertical scale scores was 100-825 for Grade 3, 100-850 for Grade 4, 100-875 for Grade 5, 100-900 for Grade 6, 100-925 for Grade 7 and 100-950 for Grade 8. One member of the South Carolina Assessment Technical Advisory Committee (TAC) worked closely with the contractor to implement this vertical scaling model for the SC READY tests.[29]

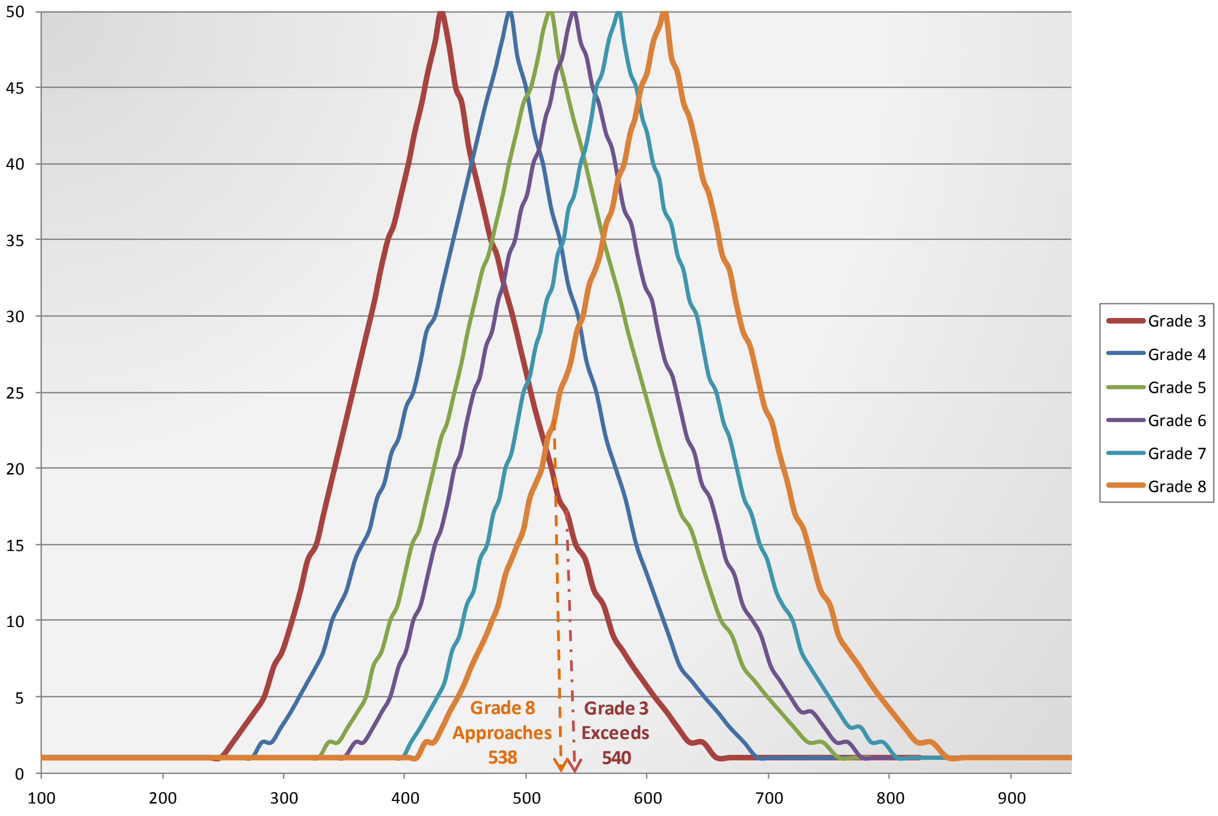
Although the *meets expectations* cut scores increased across the grade levels from lowest in Grade 3 to highest in Grade 8 (e.g., 452, 509, 558, 576, 615, and 643, respectively, for ELA), the grade level scale score distributions overlapped substantially because the minimum vertical scale score was identical (100) for all grade levels and the maximum score increased only 25 points from one grade level to the next on a scale with a maximum range of 850 (Grade 8 maximum of 950 points minus the all grades minimum of 100 points). These relationships are illustrated in Figure 2 for the ELA 2017 vertical scale.

As Figure 2 illustrates, the grade level distributions overlap significantly on this vertical scale, especially Grades 5 and 6 (the green and purple distributions in the center of the figure). Just as one example, consider two hypothetical siblings in Grades 3 and 8, Chris and Pat, whose SC READY ELA scores are at the *exceeds expectations* (540) and *approaches expectations* (538)

[26] DRC (Dec. 14, 2017). *Response to Questions and Requests for Additional Information/Data for Report #2*, Communication to HumRRO; Technical Manual, *supra* note 13, p. 37.
[27] Technical Manual, *supra* note 13, p. 36-37.
[28] Rasch, Georg. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute for Educational Research.
[29] *See* Technical Manual, *supra* note 13, p. 37.

cut scores, respectively, as shown in Figure 2. An uninformed observer might conclude from these data that Chris and Pat have similar ELA knowledge and skills. But Chris' *exceeds expectations* score is based on an assessment of Grade 3 ELA content standards while Pat's *approaches expectations* score is based on an assessment of the more complex Grade 8 ELA content standards. No doubt, neither Chris, nor Pat, nor their parents or teachers believe their ELA skills are similar. Such comparisons are unwarranted because most Grade 3 students have not yet been taught any Grade 8 content and no Grade 8 ELA items were administered to any Grade 3 students. Yet, one of the purposes of a vertical scale is to produce comparable scores that facilitate meaningful interpretations for tests administered at different grade levels.

Figure 2. SC READY ELA Distributions on 2017 Vertical Scale



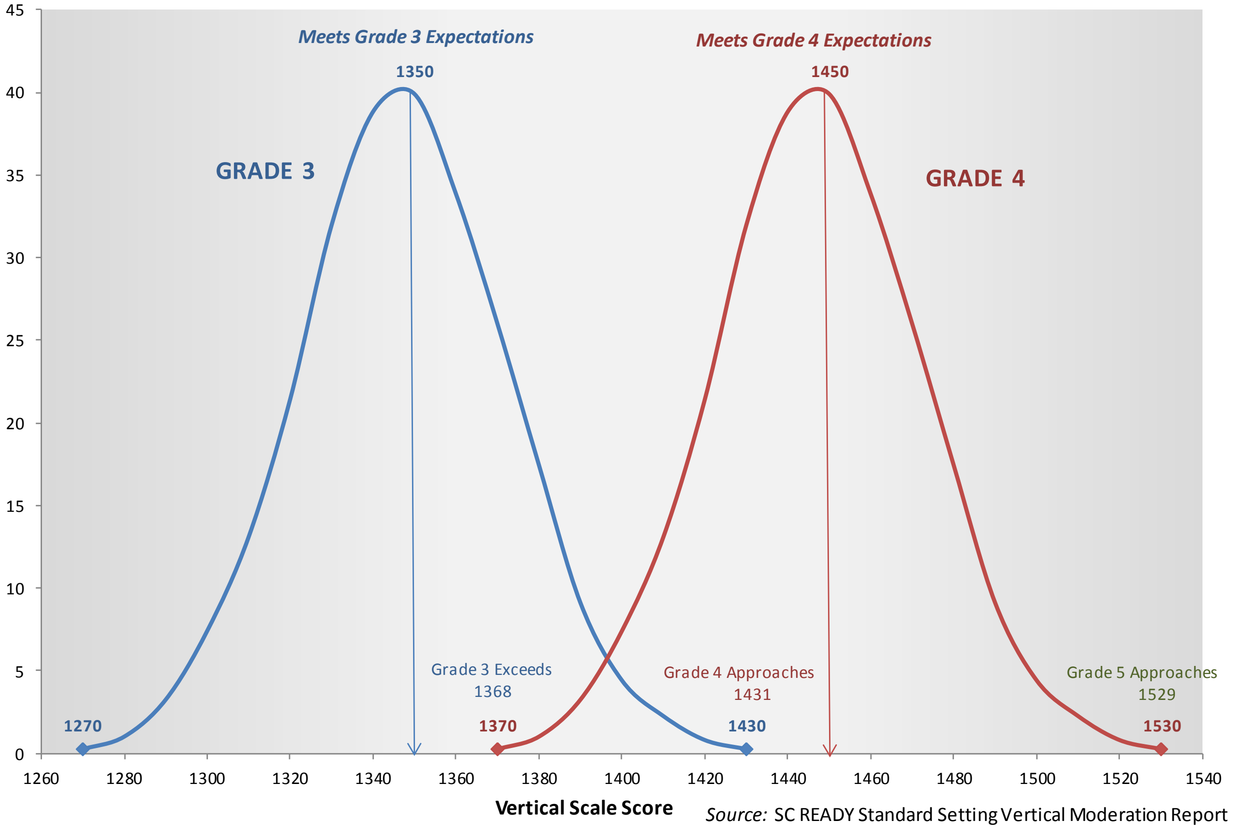
Source: SC READY Spring 2017 Conversion Tables

2016 Grade Level Scale. For the initial administration of the operational SC READY tests in 2016, a within grade-level scale was developed. This within grade-level scale was constructed by fixing the minimum score, maximum score and *meets expectations* cut score using the ability scale from the Rasch model calibrations. The ability corresponding to the *meets expectations* cut score was fixed at 1X50 where X=the grade level (3, 4, 5, 6, 7, or 8). For example, the Grade 3 ELA *meets expectations* cut score on this scale was 1350, the Grade 4 *meets expectations* cut score was 1450, and so on. The range of scale scores was then set at ± 2.5 standard deviations on the Rasch ability scale. The minimum and maximum scale scores for each grade level range were then fixed at 1(X-1)70 and 1(X+1)30, respectively, where X= 3, 4, 5, 6, 7, or 8. For example, the range of possible within-grade-level scale scores was 1270-1430 for Grade 3, 1370-1530 for Grade 4 … 1770-1930 for Grade 8.[30] Although not on a common

[30] *See* DRC (2016a). *SC READY Standard Setting Vertical Moderation Report*, Maple Grove, MN: Author, p. 2.

scale, these scale score distributions exhibited much less numerical overlap between adjacent grade level distributions than the 2017 vertical scale as illustrated in Figure 3 for SC READY ELA Grades 3 and 4.

Figure 3. SC READY ELA Distributions for 2016 Within-Grade-Level Scale Scores



Source: SC READY Standard Setting Vertical Moderation Report

As indicated in Figure 3, the Grade 3 *exceeds expectations* cut score of 1368 is near the minimum Grade 4 score of 1370 on the scale. The Grade 4 *approaches expectations* cut score of 1431 is near the maximum Grade 3 score of 1430 on the scale. Similarly, the Grade 5 *approaches expectations* cut score of 1529 is near the maximum Grade 4 score of 1530 on the scale. If an uninformed observer concluded from these data that a student who *exceeds expectations* in Grade 3 based on the Grade 3 content standards is in the range of *does not meet expectations* for Grade 4 based on the Grade 4 content standards, this would be a reasonable conclusion given that most Grade 3 students have not yet been instructed on the Grade 4 content standards and no Grade 3 student was administered any Grade 4 ELA items.

In more global terms, these relationships can be generalized as follows. With *approaches expectations* cut scores near 1X30 and exceeds expectations cut scores near 1X70 for each grade level (3-8), the within-grade-level scale scores for each grade level ranged from an approximate minimum of the *exceeds expectations* cut score for the next lower grade to an approximate maximum of the *approaches expectations* cut score for the next higher grade. For example, the range of vertical scale scores for Grade 4 ELA, 1370-1530, ran from near the Grade 3 *exceeds expectations* cut score (1368) to near the Grade 5 *approaches expectations* cut score (1529). These scores are marked on the ELA within-grade-level scale score distributions shown in Figure 3.

Again, consider the two siblings in Grades 3 and 8 who scored at the *exceeds expectations* and *approaches expectations* cut scores, respectively. Unlike the corresponding 2017 vertical scale scores (540 and 538), the 2016 within-grade-level scale scores for the Grade 3 ELA *exceeds expectations* cut score (1368) and the Grade 8 ELA *approaches expectations* cut score (1830) are far enough apart that a casual observer would be highly unlikely to conclude that they represented similar levels of ELA achievement.

The 2016 within-grade-level scaling model was apparently replaced by the 2017 vertical scaling model described above sometime after the conclusion of the SC READY standard setting vertical moderation activities. The reason(s) for the change in scale score models and the purpose(s) for developing vertical scale scores for reporting are unclear from the documentation available to us.

Consistency with Relevant Standards. HumRRO researchers identified and rated 10 Standards from the 2014 *Test Standards* judged to be most relevant to the scaling and equating of the SC READY tests. Based on a review of the available documentation, the consistency of SC READY procedures with these Standards was rated on a scale of 1=no evidence to 5=fully covered. A summary of the results is presented in Table 12.[31]

As indicated in Table 12, ratings were high, with three 5s and seven 4s across the 10 identified Standards. Comments suggested that PLDs and use of post-equating procedures for Grade 3 Reading be reported in more detail, equating error metrics be reported, and more detail for the study linking SC READY scores to lexiles® and quantiles® be provided.

Evaluation

The evaluation of Legislative Criterion 2 includes comments on the system of SC READY assessments, alignment, benchmarking, reliability and vertical scaling.

System of SC READY Assessments

The SC READY assessments are a system of summative, standards-based assessments aligned to state content standards for ELA and Mathematics in Grades 3 through 8. The content standards are designed to cover progressively more difficult content required for success in subsequent grades and leading to sufficient content knowledge in Grade 8 to be prepared to achieve CCR status in high school. The assessments are summative because they include all the content students are expected to learn at their grade level and are given within the last 30 days of the school year.

Alignment

Based on the HumRRO evaluations, the 2017 SC READY assessments demonstrated very good alignment between the content standards, test blueprints and test items for ELA and good to acceptable alignment for Mathematics. For Mathematics, the items were strongly aligned to the standards and matched the test blueprint, but the panelists disagreed with the weighting of items by reporting category. However, because blueprint weights are based on subjective judgment and enough items are needed to provide sufficient subscore reliability for reporting, one can conclude that there is satisfactory evidence of the alignment between the SC READY assessments and the state content standards. That alignment, in turn, supports the assertion that the SC READY assessments are standards-based.

[31] Chapter 5 (Task 5), Technical Manual, *supra* note 13, p. 36-41; program output from Rasch software.

Table 12. Ratings of SC READY Consistency with Identified Test Scaling & Equating Standards

Standard	Description	Rating	Comments*
5.1	Clear explanations of the intended interpretations of scale scores	4	SRUG includes interpretive information for the reported scores and precision; Reading PLDs not referenced as a link
5.2	Clear description of and rationale for procedures used to construct scales	4	Tests post-equated except for Grade 3 ELA with early reports for Reading; insufficient detail for replication
5.5	Rationale for interpreting criterion-referenced classification categories	5	SC READY tests are criterion-referenced; classification consistency estimates in TM
5.6	Stability checks of scales used across multiple years	4	2017 first year of the vertical scale; stability checks TBD; 2016 & 2017 performance levels are comparable but not scale scores
5.8	Norms based on clearly-described populations of interest	4	PRs for SC & other states, and lexiles® & quantiles® from contractors' user norms
5.12	Rationale & supporting evidence for inter-changeability of scores from alternate forms	4	Consistent test development and quality control; annual post-equating
5.13	Description of methods and accuracy of equating procedures	4	Described in separate technical reports; error metrics not mentioned; Grade 3 ELA?
5.15	Description of selection, content representativeness and characteristics of anchor items used in equating	4	15-18 anchor items from lower grade; number deleted by grade not reported in TM; selected to be content representative‡
5.17	Evidence of score comparability for scores derived from linking studies	5	Lexile®/quantile® study by MetaMetrics; no written documentation; NWEA Study
5.18	Clear description of limitations of linking tests that are not closely parallel	5	Some for NWEA; none reported for ACT Aspire® used in standard setting or for lexiles®/quantiles®

* SRUG=Score Report User's Guide, PLD=performance level descriptors, TM=Technical Manual; TBD=to be determined; PR=percentile rank; ‡ Selection criteria for anchor items, p. 37 of Technical Manual; *Source:* Chapter 5 (Task 5), Technical Reports and special studies.

Benchmarking to User Groups of States with Limited Comparability Data

As indicated in the section on Legislative Criterion 1, the SC READY assessments are directly benchmarked to performance by students in relatively large and small user norm groups from two contractors. This reported information includes percentile ranks for three unidentified states plus South Carolina that are described as having similar content standards, and lexile® and quantile® percentile ranks presumably derived in part from partner states of MetaMetrics.® Indirect benchmarking is also provided by ACT Aspire® impact data tied to ACT CCR benchmarks and NAEP data used in the vertical moderation activities during standard setting. However, in the available documentation there is insufficient direct evidence of alignment of the content standards from any of these states or tests or demographic data to adequately evaluate the quality of the benchmarking. Nonetheless, as stated earlier, these data appear to provide the best available CCR benchmarking currently possible for the SC READY tests.

SC READY Benchmarking to College and Career Readiness

As an alternative, the state might consider using South Carolina data to validate a chain of performance linking each grade level to preparedness for the following grade level with a culminating prediction of sufficient content knowledge in Grade 8 to be prepared to achieve CCR status by high school graduation if current effort is maintained and passing grades are achieved in appropriate high school CCR courses. Current information indicates that South Carolina is planning to move its end-of-course testing from English I and Algebra I to English II and Algebra II. When that happens, it would be desirable to complete the relevant alignment studies and empirical research studies linking ACT or SAT benchmarks directly to the EOCEP English II and Algebra II test scores followed by linking the end-of-course test scores to SC READY Grade 8 scores via English I and Algebra I. With the Grade 8 CCR prediction target established, Grade 7 on track performance could be linked to predictions of achievement of the Grade 8 target, Grade 6 to Grade 7, and so on.

Nonetheless, making statements about CCR for students in Grades 3-5 is not recommended because such predictions contain unacceptably large errors and may cause undue stress and anxiety for parents and educators. There are simply too many unaccounted for factors influencing student achievement across multiple years to reliably predict on track performance for high school CCR status from assessments administered in the elementary grades. It would be better to label such scores as *on track for the next grade level* leading to an *on track for CCR* designation for those students who achieve the ACT-linked, SAT-linked or other appropriately linked targets in Grade 8. Studies linking end-of-course performance with grades in nonremedial, credit-bearing, entry-level college courses would also be useful to support the validity of CCR benchmarks tied to SC READY test scores.

Although the lexile® and quantile® trajectories to Grade 12 CCR ranges provide useful evidence for claims of *on track performance for CCR*, particularly for students who *meet expectations*, but the accuracy of such predictions for South Carolina students has not yet been documented. The long term accuracy of such predictions is also not known and use of only the Reading subtest ignores relevant additional information provided by the ELA test scores.

Reliability

Reliability estimates for SC READY were generally high and met the Assessment TAC recommendation of .85 for all subjects, grade levels and groups except students with disabilities in Grades 7 and 8 Mathematics. The lower values may have occurred because the disabilities and accommodations represented by these students are very diverse and their achievement tends to have greater variability as grade level increases. It may be useful to seek input from special education administrators to ascertain possible reasons for the lower reliabilities in middle school Mathematics for these students. Overall, though, the SC READY reliabilities are judged to be acceptable to very good for the purposes for which the scores are being used.

In addition to decision consistency estimates for SC READY total scores, decision consistency estimates should also be reported for ELA Reading, especially in Grade 3. For Grade 3 students, the state is currently providing preliminary (early) Reading indicators of performance at or above a minimum cut score for the purpose of satisfying the statutory requirement that reading scores be considered when promotion and retention decisions are made. Also, rather than the KR21 reliabilities estimated from 2016 data, Reading subscore reliabilities should be calculated using the same methodology used for the SC READY total test scores.

Currently, there are no reliability estimates for the reporting category scores. Reporting category scores are classified as low, middle or high based on ability metrics from the underlying Rasch model. The ability estimate from the total test that is equivalent to the *meets expectations* cut score is located on the scale for each reporting category and an interval of plus and minus one standard error around this value forms the middle interval. Low reporting category scores fall below that interval and high reporting category scores above it. For example, the raw score ranges for the Grade 3 ELA and Grade 8 Mathematics primary reporting categories are shown in Table 13.

Table 13. Primary Reporting Category Raw Score Ranges for 2017 SC READY Grade 3 ELA and Grade 8 Mathematics Tests

Grade 3 ELA	Read Literary Text	Read Informational Text	Writing	Inquiry
Low	0 – 9	0 – 7	0 - 14	0 – 4
Middle	10 – 13	8 – 12	15 – 19	5 – 7
High	14 – 19	13 - 19	20 - 35	8 - 10

Grade 8 MATH	Number System	Functions	Algebra‡	Geometry Measurement	Data Analysis Stat & Prob
Low	0 – 4	0 - 6	0 - 7	0 - 6	0 - 5
Middle	5 – 7	7 - 9	8 - 10	7 - 9	6 - 7
High	8 – 9	10 - 14	11 - 16	10 - 14	8 - 9

‡ Algebra includes expressions, equations and inequalities;
Source: 2017 SC READY Vertical ELA (Math) Raw Score to Scale Score Tables

According to the contractor, the Assessment TAC advised the SCDE not to provide reporting category raw scores because there are too few items in each category to provide sufficient reliability and stability. Yet educators in the field requested “diagnostic” scores to provide an indication of students’ relative strengths and weaknesses. The SCDE compromised by providing the low, middle, and high indicators for the reporting categories. On the advice of the Assessment TAC, the calculation of reliabilities for these indicator scores has been delayed for a few years until the scores are more stable.[32] However, the 2014 *Test Standards* quoted below indicate that it is not psychometrically appropriate to report any scores for which reliability estimates are unavailable.

> **Standard 2.3**
> For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.
>
> ***Comment:*** *It is not sufficient to report estimates of reliabilities and standard errors of measurement only for total scores when subscores are also interpreted. The form-to-form and day-to-day consistency of total scores on a test may be acceptably high, yet subscores may have unacceptably low reliability, depending on how they are defined and used. Users should be supplied with reliability data for all scores to be interpreted, …*

[32] DRC Response to Questions, *supra* note 24.

The SCDE is urged to calculate preliminary reliability estimates for the indicator scores now, and then later reconfirm and revise them if warranted when scores are more stable.

Vertical Scaling

The methodology and properties of the SC READY 2016 within-grade-level and 2017 vertical score scales are described above. Recall that the 2017 vertical score scale was developed from 2017 data for which lower grade items were administered in adjacent upper grades. This is a common psychometric method for developing vertical scales, although some programs strengthen the observed relationships by also administering carefully selected upper grade items to students in the adjacent lower grade where feasible.

A major issue with the current SC READY vertical scale is the potential for confusion or distress when students with equivalent scale scores are compared or negative growth is reported. Vertical scale scores are designed to be on a common scale across grades to support comparisons. But the large overlap in adjacent grade level SC READY vertical scale score distributions may lead parents and educators to misunderstand their interpretation. Students at adjacent or non-adjacent grade levels may achieve identical vertical scale scores suggesting to the uninformed observer that their performance is similar. For example, for the SC READY Mathematics assessments, a vertical scale score of 545 *exceeds expectations* in Grade 3, *meets expectations* in Grades 4, 5, and 6 and *approaches expectations* in Grades 7 and 8.

Perhaps of more concern, the 2017 vertical scale scores may exhibit the undesirable and unsupportable property of negative growth. That is, a student who is classified as proficient in two adjacent grades may have a smaller vertical scale score in the upper grade than in the lower grade. As a result, when one evaluates the “vertical scale score progression” on the SC READY score report, the “gain” from one year to the next may be negative. The potential for negative growth scores with the 2017 vertical scale is discussed more fully in the section for Legislative Criterion 3 below.

Alternatively, if one assumed purely for illustration purposes that the within-grade-level scale scores reported for SC READY in 2016 were actually on a vertical scale, the potential for misinterpretation would be greatly reduced. The distributions in that scaling model generally overlapped the *exceeds expectations* category for a lower grade with the *does not meet expectations* category for the adjacent upper grade level. Although comparative conclusions are inappropriate because the scale scores are not actually on a vertical scale, more meaningful interpretations would be facilitated by the properties exhibited by the within-grade-level scale scores. Less harm is likely to occur when an *exceeds expectations* scale score in a lower grade level corresponds to a *does not meet expectations* scale score in the adjacent upper grade level because a casual observer will probably conclude correctly that the lower grade student has not yet learned the content taught in the upper grade.

Misinterpretations in the reverse direction (upper grade L1 = lower grade L4 score) are also not likely because unless the student was new to the state, parents and educators would have the previous SC READY lower grade level scale score available on the score report for comparison with the current SC READY upper grade scale score. Finally, a student scoring *exceeds expectations* in a lower grade would have to actually score *does not meet expectations* the next year to receive a scale score in the upper grade level that is lower than the one earned in the lower grade level. Such an event is highly unlikely, suggesting that nearly all students would show positive scale score gains. Positive growth is desirable because it recognizes that an additional year of schooling should result at minimum in some positive achievement gain.

3. Creation of SC READY Scores for Achievement of State Standards, Preparation for the Next Grade Level, and Student Growth in ELA (reading, writing) and Mathematics

Evidence

Evidence relevant to Legislative Criterion 3 includes a description of SC READY reported scores that indicate achievement of state standards, interpretations of SC READY scores as indicators of preparedness for the next grade level, and SC READY results related to student growth in ELA and Mathematics.

SC READY Scores for Achievement of State Standards

Individual student score reports for the SC READY ELA and Mathematics tests include several different types of scores designed to provide evidence of student achievement of state standards. For the ELA total score, the ELA reading subscore, and the Mathematics total score, the student receives a performance level designation of *exceeds expectations, meets expectations, approaches expectations, or does not meet expectations* as defined by the South Carolina grade level content standards and standard setting activities (*see* the *Evidence* section of Legislative Criterion 5 for descriptions of the four performance levels). The cut scores delimiting the performance levels for each grade level and subject were recommended by panels of South Carolina educators who carefully considered the content demands of the test questions and the requirements of the corresponding state content standards that they measured.

Panel members were aided in their task by sets of performance level descriptors (PLDs) that identified essential skills from the state content standards for students in each performance level by subject and grade level. For example, third grade students who *meet expectations* in ELA Reading are expected to *explain the differences between first and third person points of view* (*see* the *Evidence* section of Legislative Criterion 5). A small, representative subset of the educator panel participants then considered the reasonableness, consistency, and external validity relative to ACT Aspire® and NAEP state results for the total set of cut scores for the four performance levels across Grades 3-8. This vertical moderation panel recommended adjustments where appropriate to create a more coherent system of performance expectations.

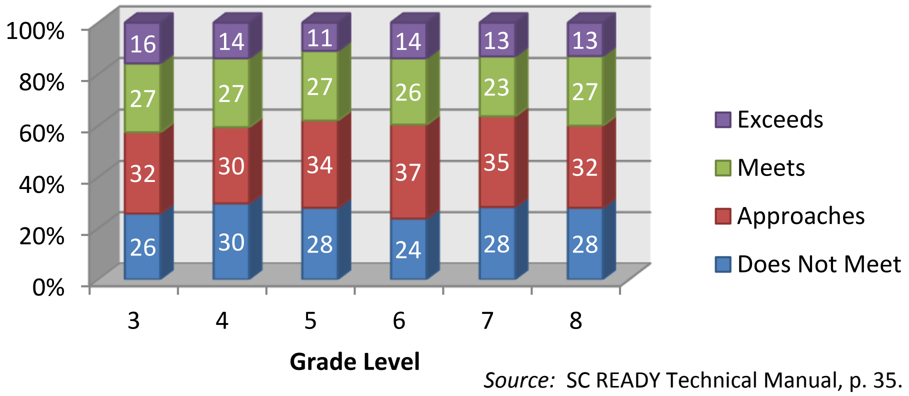
In addition to reporting performance levels for each total score, the SC READY test score reports also report each student's vertical scale score and an interval in which the student's scale score would likely fall if the student were to test again under similar circumstances. These intervals are based on the standard error of measurement of the test and can be used to evaluate the likelihood the student would remain in the same performance category if tested again. For example, consider Sam, a hypothetical sixth grade student who *meets expectations* and earned a scale score of 545 in Mathematics with a corresponding interval of 535-555. The *meets expectation* performance category for Grade 6 Mathematics begins at a scale score of 543, so if Sam were to test again, there is a reasonable likelihood that Sam's retest score could fall in the *approaches expectations* performance category.

A sample 2-page SC READY test score report from the Score Report User's Guide is presented in Exhibit C at the end of this report. The SC READY performance level scores and scale scores are presented on page one of the score report. Interval estimates are on page two.

2017 State Results. The percent of students scoring in each of the four performance levels for the 2017 SC READY ELA and Mathematics assessments by grade level is presented in Charts

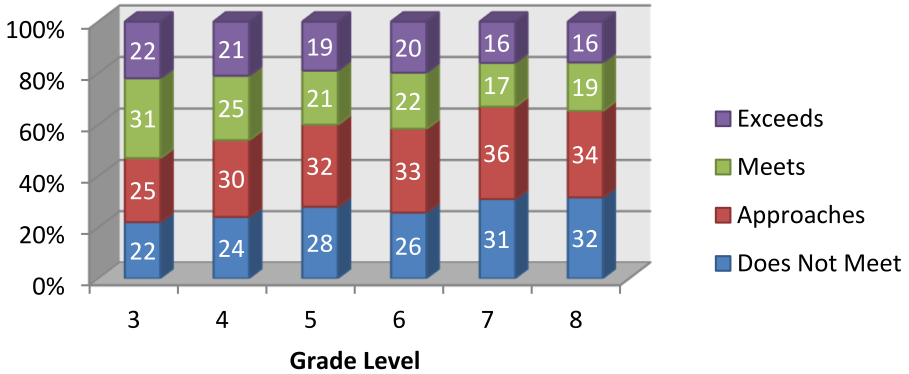
3 and 4. As indicated in the charts, performance for *meets expectations* and *exceeds expectations* was similar across grade levels for ELA but declined across grade levels for Mathematics. At all grade levels, slightly more students scored in the *exceeds expectations* category for Mathematics than for ELA. Overall, ¼ to ⅓ of students *did not meet expectations*.

CHART 3
2017 SC READY *ELA*
Percent of Students by Performance Level



Source: SC READY Technical Manual, p. 35.

CHART 4
2017 SC READY *Mathematics*
Percent of Students by Performance Level



Source: SC READY Technical Manual, p. 35.

Charts 5a and 6a compare the percent of students estimated to score in each performance level based on 2016 data with the actual 2017 impact data. The 2016 estimates were used by panelists during the standard setting activities, and because the two estimates were close, the 2016 data provided reasonable guidance for the panelists. Charts 5b and 6b present impact data by group. The 80% White statistic is a rule of thumb often borrowed from Title VII employment law that creates a rebuttable presumption of disparate impact when minority group performance falls below it as shown in the charts. But the state can successfully counter these data by demonstrating that the assessments are consistent with the *Test Standards,* follow psychometric best practices, and produce achievement scores that are valid and reliable. Provision of appropriate remediation is also helpful (e.g., Grade 3 Reading). Sometimes statistical significance tests are used instead of the 80% rule, but those analyses are inappropriate here because the data reflect subpopulations, not samples.

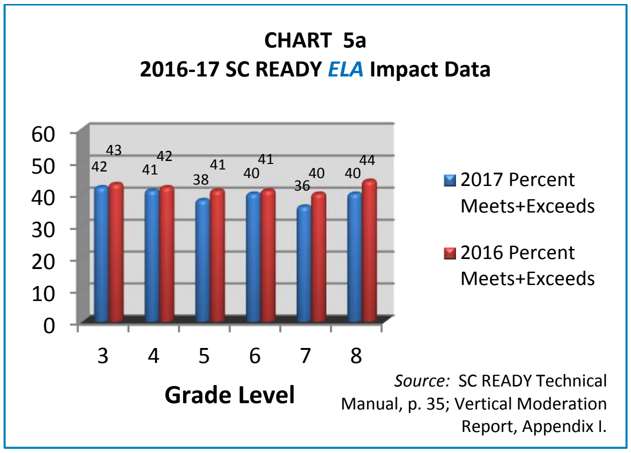


CHART 5a
2016-17 SC READY *ELA* Impact Data

Source: SC READY Technical Manual, p. 35; Vertical Moderation Report, Appendix I.

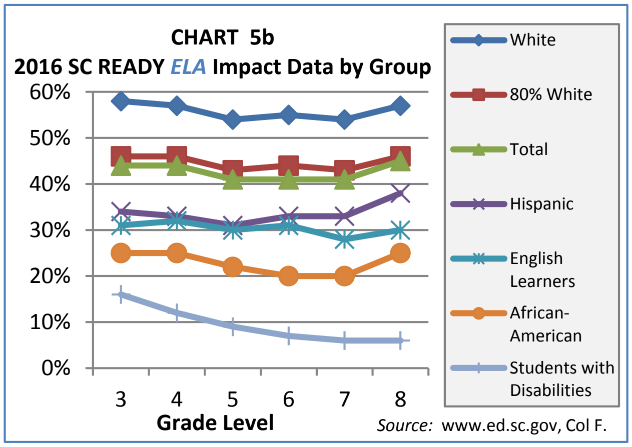


CHART 5b
2016 SC READY *ELA* Impact Data by Group

Source: www.ed.sc.gov, Col F.

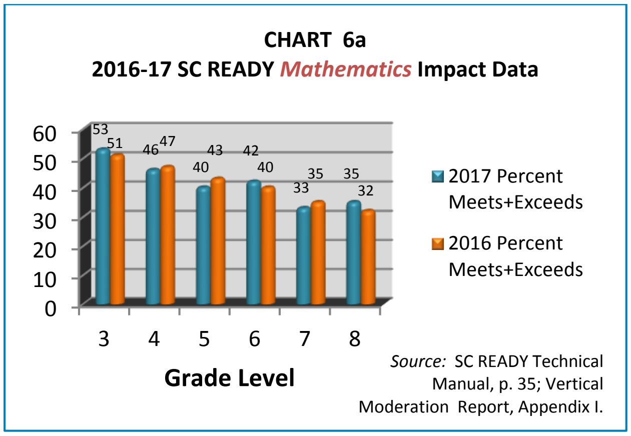


CHART 6a
2016-17 SC READY *Mathematics* Impact Data

Source: SC READY Technical Manual, p. 35; Vertical Moderation Report, Appendix I.

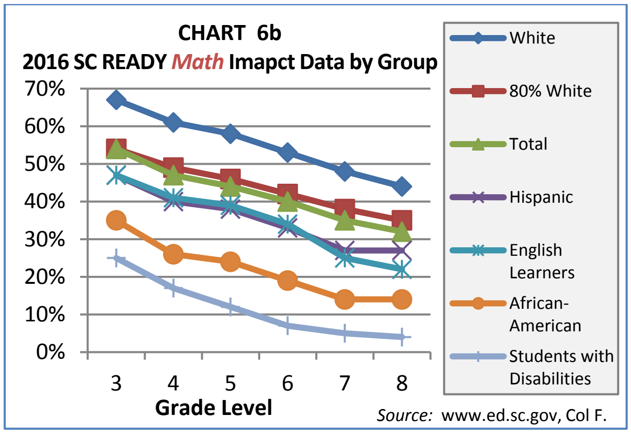


CHART 6b
2016 SC READY *Math* Imapct Data by Group

Source: www.ed.sc.gov, Col F.

Preparation for the Next Grade Level

Consider the following four reasonable inferences from the documentation for the SC READY assessment program.

1. The common core CCR standards on which the South Carolina content standards are based were designed to spiral increasingly complex strands of content knowledge and skills beginning at Grade 3 and progressing through Grade 8;

2. There is substantial overlap between South Carolina state content standards and the common core CCR standards;

3. The test items for the SC READY assessments were selected from the contractor's ELA and mathematics CCR item banks (developed to assess the

Common Core standards at each grade level) to measure important content from the South Carolina ELA and Mathematics content standards; and

4. When determining the level of test performance that reflects student knowledge and skills that just barely meet the minimum expectations for achievement of grade level content standards, the South Carolina educators who participated in the standard setting panels considered what students needed to know and be able to do to be prepared for instruction at the next grade level.

Given these inferences, one might logically conclude that students who score at or above the *meets expectations* performance level cut score on their grade level SC READY ELA or Mathematics tests have sufficient prerequisite knowledge and skills to be adequately prepared for the material covered at the next grade level. However, continued success at the next grade level is dependent on continued maintenance of effort by the student and adequate educator review of critical, prior-grade skills when school resumes after the summer break.

Example. Returning to the hypothetical sixth grade student Sam, described above, his teacher might use information on his SC READY score report to evaluate his readiness for seventh grade mathematics classes. The teacher's initial impression might be that Sam is prepared for seventh grade mathematics because his sixth grade Mathematics total score was classified as *meets expectations.* However, based on Sam's interval of likely performance if he were retested as described above, Sam's Mathematics performance may actually fall at the upper end of *approaches expectations.* This performance level result would indicate that Sam is not fully prepared for seventh grade mathematics work.

One piece of information Sam's teacher might use to identify areas of weakness for which Sam might not be adequately prepared for seventh grade is Sam's SC READY test performance in each of the Mathematics reporting categories. Performance for these subsets of items within the Mathematics test is rated high, middle or low. For example, suppose Sam's test results were as shown in Table 14.

Table 14. Sam's Grade 6 Mathematics Test Results

Reporting Category	Rating
The Number System	Middle
Ratios and Proportional Relationships	Middle
Expressions, Equations, and Inequalities	Low
Geometry and Measurement	High
Data Analysis and Statistics	Low

Sam's low performance on expressions/equations/inequalities and data analysis/statistics test items suggests that these are areas Sam should work on before beginning seventh grade. Ways in which Sam's teacher and parents might use his quantile® measure to assist Sam to review these mathematics skills are described in the evidence section for Legislative Criterion 8 (*see* Exhibit A for a sample Quantile® Score Report).

SC Read to Succeed Legislation. Another area in which SC READY results are used to gauge preparedness for the next grade level involves the South Carolina Read to Succeed Act.[33] Section 59-155-160 requires Districts to evaluate third grade students' reading abilities when deciding whether they should be promoted to fourth grade, attend summer school, or be retained in third grade for another year. To provide one piece of objective evidence for that decision, Districts can receive preliminary results for the SC READY Grade 3 ELA Reading subscore within three (online) to six (paper/pencil) days of test administration (online) or contractor receipt of answer documents (paper/pencil).[34]

For the prior Palmetto Assessment of State Standards (PASS) ELA test, a *Not Met 1* cut score had been established such that scores at or above it indicated sufficient reading achievement to be minimally prepared for fourth grade. The SC READY Grade 3 ELA Reading subscore judged equivalent by an educator panel to the earlier *Not Met 1* performance standard is used to classify third grade students' reading achievement as *at or above* or *below* the required performance standard.[35]

Separate ELA Reading subscores and associated performance levels are also reported for Grades 4 through 8. The ELA Reading performance levels indicate whether students are keeping up as reading demands increase across grade levels for a combination of literary and informational texts. These scores, along with reporting category indicators for reading literary text, reading informational text, writing, and inquiry, provide more specific evidence of sufficient prerequisite knowledge and skill to be prepared for instruction on the ELA content standards at the next grade level. In addition, because reading becomes increasingly important for instruction in subjects other than ELA as grade level increases, the reading subscore performance level also provides some evidence of preparedness for the reading demands of other subjects at the next grade level.

SC READY ELA Essay Score. A final piece of evidence of preparedness for ELA instruction at the next grade level is provided by the ELA text dependent analysis (TDA) essay score. The TDA essay item requires the student to read one or two passages and write an essay that addresses a content question about the passage(s) (*see* Legislative Criterion 6 for a sample item). A writer's checklist and scoring rubric are provided to guide the student while responding (*see* Exhibits E and F). The student's essay is scored by two raters. If the raters' scores differ by more than one point, the essay is scored by a third rater. Ratings are averaged and weighted by four to produce a maximum essay score of 16.

The SC READY score report includes the student's total number of points out of 16 possible on the TDA (essay) item. The scoring rubric and student responses selected as anchor papers for ratings 1 to 4 are used to score the TDA essay item. According to the scoring rubric, a student with an average TDA item score of 3-4 (total points of 12-16) "demonstrates *adequate to effective analysis* of text and *appropriate to skillful* writing" (*see* Exhibit F). Given the reasonable inferences that the stimulus material and content question for the TDA essay item are appropriate for the student's grade level and the response is scored consistent with the writing skills listed in the content standards for that grade level, one might reasonably conclude that students scoring at a 3-4 level on the TDA essay item have adequate writing skills to be prepared for the next grade level.

[33] S.C. Code Ann. § 59-155-160 (2014).

[34] *Score Report User's Guide*, *supra* note 7, p. 6; SCDE (2017c). *SC READY and SCPASS Spring 2017 Test Administration Manual* [Test Administration Manual], Columbia, SC: Author, p. 6.

[35] *See* DRC (2017a). *Reading Grade 3 Standard Setting Report*, Maple Grove, MN: Author, p.1.

SC READY Mathematics. For SC READY Mathematics, the spiral design of strands of content that increase in difficulty across grade levels demonstrates that skills taught in the next grade build on achievement of prerequisite skills from the prior grade.[36] This connection of skills across grades in the state content standards provides support for the assertion that students who meet expectations in their current grade are prepared for instruction in the next grade.

For example, the data analysis strand reported in the lower grades (Grades 3-5) adds statistics in Grade 6 and probability in Grades 7 and 8. Thus, if a student's data analysis indicator score is high in Grade 5, that result provides some evidence that the student has sufficient data analysis skills to be prepared for instruction in statistics in Grade 6. If the student then achieves a data analysis/statistics indicator of high in Grade 6, that result provides further evidence that the student has sufficient skills to tackle instruction in probability in Grade 7. Similarly, algebraic thinking and operations skills assessed in Grades 3-5 are prerequisite to success with the reporting category of expressions, equations, and inequalities in Grade 6. In addition to providing evidence of achievement of the grade level expectations for Mathematics contained in the state content standards, the performance levels for the total score and the performance indicators for the five Mathematics reporting categories also provide evidence of achievement of the prerequisite skills necessary for success at the next grade level.

Student Growth in ELA (reading, writing) and Mathematics

There are several ways student results on the SC READY tests can demonstrate growth in ELA and Mathematics. These include maintaining a *meets or exceeds expectations* performance level in the prior and current testing years, exceeding the prior year's lexile® or quantile® scores, increasing one's vertical scale score, and increasing one's federal accountability growth score.

Scoring Meets Expectations in Successive Grades. Maintaining effort and staying in the *meets expectations* performance level from one grade to the next demonstrates growth because the material students are expected to learn becomes harder at the next grade than it was for the previous grade. To illustrate this concept, once more consider the hypothetical sixth grade student Sam, whose SC READY sixth grade Mathematics scale score of 545 placed him in the *meets expectations* performance category. If Sam scores a 579 on his SC READY Mathematics test the following year in seventh grade, his scale score will have increased by 34 points and his performance level will again be *meets expectations*. To accomplish this, Sam would have to learn new and different mathematics content because the Grade 7 test is aligned to different state content standards unique to seventh grade. For example, in Grade 6, students are expected to learn about basic statistics, but in Grade 7, they are required to extend this knowledge to learn about probability. Algebraic equations provide another example. In Grade 6, the state Mathematics content standards specify that students are to solve one-step linear equations, and the following year in Grade 7 are to solve multi-step linear equations, a more difficult skill to master and one that requires already knowing how to solve one-step problems.[37]

Achieving the Same or Higher State Percentile Rank. Another way to quantify Sam's growth from Grade 6 to Grade 7 is to compare his state percentile ranks for the two years. Sam's Grade 6 PR was 59 and his seventh grade PR would be 67, indicating that Sam improved his relative position with respect to state sixth and seventh grade students. Clearly Sam would have to grow mathematically and learn more content to improve his relative position among other

[36] SCDE (2016a). *South Carolina College- and Career-Ready Standards for Mathematics,* Columbia, SC: Author.
[37] *Id.*, Grades 6 and 7.

state students who were also being taught the seventh grade state mathematics content standards.

Lexile® and Quantile® Gains. Each year, South Carolina students will receive score reports showing their lexile® and quantile® measures corresponding to their SC READY ELA Reading and Mathematics Total scores (*see* sample score reports in Exhibit A).[38] In addition to presenting their current lexile® and quantile® scores, the accompanying graph will also show students' corresponding lexile® and quantile® scores from previous administrations of SC READY grade level tests. As students encounter instruction of increasing complexity tied to content standards for higher grade levels, one would expect the student to be able to handle reading texts and mathematical instructional materials at higher lexile® and quantile® levels, respectively. Students' score reports after several grade levels of testing will graphically depict the growth in reading and mathematics ability across grade levels. For example, a student whose lexile® score is 695L in Grade 3 and 822L in Grade 4 the following year has gained 127L and moved from the Grade 3 range to the Grade 4 range in terms of the complexity of reading texts that student can comprehend (*see* Lexile® Score Report in Exhibit A).

Vertical Scale Score Progressions. Another method suggested for tracking SC READY student growth is based on the vertical scale scores. The first page of the individual student score report (*see* Exhibit C) has a box at the bottom labeled "Your Student's Scale Score Progression." Reported in the box below this heading are the student's SC READY ELA and Mathematics vertical scale scores for each grade level for which the student has been tested.[39] In parentheses next to each reported scale score is the performance level corresponding to that score. A paragraph in the SC READY Score Report User's Guide explaining these scores appears under the heading "Scale Score Progression."[40] These labels seem to suggest that student vertical scale scores are expected to increase from one grade level to the next as an indicator of student growth. In addition, a table of vertical scale score cuts prepared by the contractor states "The [highest obtainable score] was set to fall within the 99th percentile of each grade, but is designed to increase by grade for students to have the opportunity to show growth" (*see* Exhibit G).

Accountability Growth Scores. There is also an accountability growth score for SC READY used for federal reporting for schools and districts under the Every School Succeeds Act (ESSA).[41] The accountability growth score is based on the Education Value Added Assessment System (EVAAS) methodology. The state has commissioned the creation of a growth index for each school based on a composite index for all students plus an additional growth index for the lowest achieving quintile. Student level projections are also available through EVAAS but are not printed on the individual student report.[42]

Evaluation

Each of the areas covered by Legislative Criterion 3 is discussed separately below. These areas include achievement of state standards, preparation for the next grade level and growth.

[38] *See also* the section on Legislative Criterion 1 for a description of lexile® and quantile® scores.

[39] There is only one entry for 2017 because it was the first year vertical scale scores were reported. In 2018+, scale scores for all years the student tested will be reported. Score Report User's Guide, *supra* note 7, p. 10.

[40] Score Report User's Guide, *supra* note 7, p. 10-11.

[41] Every Student Succeeds Act (ESSA), 20 U.S.C. § 6301 et seq. (2015).

[42] SCDE Response to Questions, *supra* note 6, p. 8.

Achievement of State Standards

There is substantial evidence that the SC READY assessments provide appropriate scores indicating achievement of state standards and preparation for the next grade level. Reported scores include Preliminary Grade 3 ELA Reading subscores, performance levels (including a category labeled *meets expectations*) for ELA, ELA Reading and Mathematics, and indicator scores for ELA and Mathematics reporting categories. Performance levels are not reported separately for writing but students do receive indicator scores for writing, meaning/content/craft and language. In addition, a raw score is reported for the TDA essay item.

Preparation for the Next Grade Level

The assertion that proficiency in one grade (*meets expectations* or above) is an indication of adequate preparation for the next grade level and ultimately for achievement of CCR in high school is supported logically by the organization and general similarity of the state content standards to the Common Core, a set of grade level ELA and mathematics content expectations specifically targeted toward ultimate achievement of college and career readiness by high school graduation. However, other than the Achieve Report that identified some weaknesses relative to key elements of CCR (*see* Legislative Criterion 1), there appears to be no other alignment evidence documenting the comparability of the South Carolina content standards to the Common Core to support this claim. Further, as yet no data have been collected to empirically validate whether proficiency on each grade level SC READY test does indeed predict success at the next grade level for South Carolina students.

Growth

The evidence for growth measures is somewhat weaker. Growth in reading and mathematics can be tracked across grades using lexile® and quantile® measures. However, there is no measure for ELA or writing growth. Alternatively, the SC READY 2017 vertical scale scores for ELA and Mathematics are described as "progressions," suggesting that they too could be used for tracking student growth. However, interpreting SC READY vertical scale scores as growth measures may lead to confusion and/or misleading conclusions because the reported "growth" may be negative and the scale scores from different tests may not be comparable in the usual sense in which this concept is understood. These concerns are described in more detail below.

Scale Scores. If the purpose for constructing a vertical scale was to place test scores from different tests at different grade levels on a common scale to report annual student growth, the 2017 vertical scale score model developed for the SC READY assessment system appears not to have achieved its goal. The 2017 vertical scale allows negative growth for adjacent years, a contradictory message to parents and educators when students have maintained the same performance level for the two years. This and other contradictions and potential misinterpretations are explained more fully in the next several paragraphs.

Negative Growth. The vertical scale score progressions reported on the individual student score report for SC READY may show growth from one grade to another. On the other hand, consider the plausible alternative demonstrated by the following example.

Returning to the hypothetical sixth grade student Sam, suppose his SC READY ELA and Mathematics Grade 6 and Grade 7 test scores placed him in the *meets expectations* performance level for both grade levels in both subjects. Also assume that he scored near the top of the *meets expectations* performance level in Grade 6 and near the bottom of the *meets* *expectations* performance level in Grade 7. Sam's vertical scale scores could be as shown in Table 15. Other SC READY test information consistent with those scale scores for those grades and subjects is also shown Table 15. Note that the shaded information in Table 15 would not appear on Sam's individual score report (ISR).

Table 15. Scale Score Comparisons for Hypothetical Student Sam

	Performance Level	Scale Score	Rasch Ability	Raw Score Pct Correct	SC PR	Other States PR
ELA						
Grade 6	***Meets***	655 } **−30**	1.59	68/96 = 71%	82	88
Grade 7	***Meets***	625	1.30	59/96 = 61%	66	64
MATH						
Grade 6	***Meets***	620 } **−40**	1.37	47/60 = 78%	79	82
Grade 7	***Meets***	580	.097	35/60 = 58%	68	71

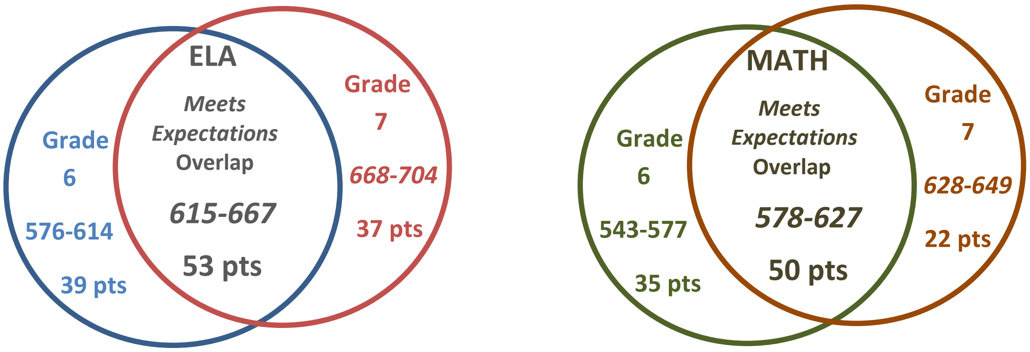
Source: 2017 Scale Score Tables; Technical Manual

Consider the contradictory message that is being sent by the information that would appear on the ISR sent to Sam's parents and teachers. Sam is considered *proficient* in ELA and Mathematics (*meets expectations*) on the content standards for the respective grade levels, and the tested content standards for Grade 7 are more difficult and complex than those for Grade 6. For example, in Mathematics, Grade 7 students are expected to learn about probability while Grade 6 students are not. Having a *meets expectations* performance level in both grades for both subjects, one could conclude that Sam has shown growth in his ELA and Mathematics skills from Grade 6 to Grade 7. Sam's quantile® scores would probably show positive growth.

But now consider his reported vertical scale scores. Sam's scale scores show regression, not progression, because these vertical scale scores (on a common scale across Grades 3-8) show that Sam has lost 30 points in ELA and lost 40 points in Mathematics. However, the information not shown on the score report does indicate correspondingly lower ability levels in Grade 7 than Grade 6 and fewer items answered correctly in Grade 7 than in Grade 6 (on different content with similar average grade level p-values). This information is consistent with the decrease in scale scores as is the decrease in relative standing (percentile ranks) from Grade 6 to Grade 7 compared with students in South Carolina and other states as shown in the last two columns of Table 15. In the sense that Sam has not maintained his relative position within these norm groups, he might be considered to have lost ground. But Sam has learned new content and met grade level expectations so one would expect his vertical scale score to increase or at the least remain the same.

What does one say to explain to parents or educators who receive Sam's score report why he *meets expectations* but his growth is negative? Has he really lost ground in seventh grade or is this the result of an artifact of the scale score model? As already stated, one common purpose for developing a vertical scale is to create comparable scores across grade levels in order to quantify student growth, so it would be distressing if the SC READY assessments appeared to show negative growth for students labeled *proficient* in adjacent grade levels. This contradiction occurs because the ranges of vertical scale scores corresponding to *meets expectations* for adjacent grades overlap substantially. For example, in Exhibit D, note that the respective *meets expectations* scale score intervals for Grades 6 and 7 are (576-667) and (615-704) for ELA and (543-627) and (578-649) for Mathematics. These overlapping scale score relationships are shown graphically in Figure 4.

Figure 4. Vertical Scale Score Overlap for SC READY
Meets Expectations Performance Levels



Identical Minimum Scores. Consider yet another contradiction for the SC READY 2017 vertical scale. Two students with minimum ELA vertical scale scores of 100, one in Grade 8 and one in Grade 3, each have not achieved the content standards for their respective grade levels, but this provides no evidence suggesting their achievement is similar. On the contrary, even though the Grade 8 student cannot read and comprehend Grade 8 texts, a scale score of 100 on the test provides little information about the student's actual reading level. In addition, the student has several more years of experience and education than the Grade 3 student so the Grade 8 student's achievement is probably unlike the Grade 3 student, even though the two students have the same reported scale score. Yet, because vertical scale scores are supposed to be comparable, these two students would appear to be starting in the same place with respect to their ELA knowledge and skills.

Cut Score Confidence Intervals. Another unusual property of the 2017 vertical scale is the few points needed to remain at the cut score for *meets expectations* from one grade to the next relative to the conditional standard errors of measurement reported for those cut scores. These data are shown Table 16 for ELA and Mathematics.

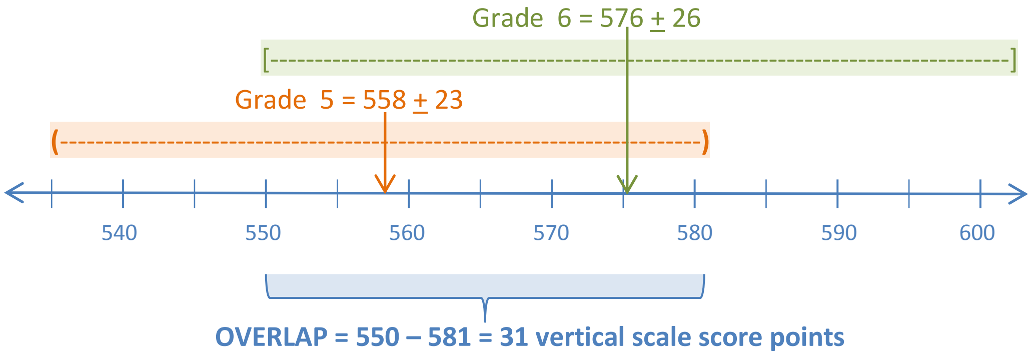
Table 16. Scale Score Differences and Conditional SEMs for Meets Expectations Cut Scores

			Scale Score Difference Between *Meets Expectations* Cut Scores	Conditional Standard Error of Measurement (cSEM) at the Cut Scores
ELA	Grades	3-4	57	24+27 = 51
		4-5	49	26+27 = 53
		5-6	***18***	23+26 = 49
		6-7	39	23+23 = 46
		7-8	28	23+23 = 46
MATH	Grades	3-4	44	28+32 = 60
		4-5	54	28+29 = 57
		5-6	***7***	28+29 = 57
		6-7	35	28+28 = 56
		7-8	37	28+28 = 56

Source: Technical Manual, p. 34, 44.

For every grade combination shown above (except Grade 3 ELA where it is almost true), the distance between the *meets expectations* cut scores for adjacent grades is less than the combined conditional standard errors for the two cut scores. This result is particularly noticeable for Grades 5 and 6 where the differences between cut scores are very small (18 points for ELA and seven points for Mathematics) and the distributions overlap more than for other adjacent grades (*see* Figure 2). This means that if confidence intervals of + 1 cSEM were placed around each of the Grade 5 and 6 *meets expectations* cut scores, the resulting intervals would overlap as shown in Figure 5, indicating that those cut scores are not reliable indicators of differential achievement. Yet, based on the state content standards and PLDs, the knowledge and skills necessary to score *meet expectations* for ELA in Grades 5 and 6 are clearly different.

Figure 5. ELA *Meets Expectations* Confidence Intervals (cut score + 1 cSEM)



Within-Grade-Level Score Scale. Alternatively, the 2016 within-grade-level scale scores for SC READY, though not a true vertical scale, do not demonstrate such contradictions because the ranges of scale scores that correspond to the meets expectations performance levels for adjacent grades do not overlap. Thus, a student who meets expectations for two adjacent grade levels cannot have a reported scale score for the upper grade level that is lower than the scale score for the lower grade level.[43] Similarly, minimum scale scores increase substantially from one grade level to the next so students in different grades taking different grade level tests will not earn equivalent minimum scores.

Summary. It is unfortunate that the 2017 vertical scale score model does not provide traditional growth scores with reasonable interpretations. Its contradictory properties for scores that are supposed to be comparable may make its scale scores distressing and confusing for important audiences such as parents, educators and the public. In addition, the user samples from which the “other states” normative percentile ranks were derived is composed of only three states plus South Carolina with content standards claimed to be similar to the Common Core but for which this alignment has not been clearly documented. This leaves only the lexile® and quantile® scores as reasonable measures of growth over time. However, these scores are incomplete growth measures for ELA because they include reading but not writing. Moreoever, the samples used to link the SC READY scores to the lexile® and quantile® scales were quite small relative to the student population, and student motivation for the separate linking tests may have been diminished because students likely knew it was a research study with no reporting of individual student scores. As a result, lexile® and quantile® measures may not be completely satisfactory replacements for the 2017 vertical scale scores for measuring student growth. Although the accountability growth scores might serve as substitute growth measures for individual students, an evaluation of this alternative is beyond the scope of this report.

[43] *See* Vertical Moderation Report, *supra* note 28, p. 2-3; Description of the vertical scale in the section on Legislative Criterion 2; Figures 2 and 3.

4. Measurement of Student Progress Toward National College- and Career-Ready Benchmarks Derived from Empirical Research and State Standards

Evidence

Evidence relevant to Legislative Criterion 4 includes direct evidence based on lexile® and quantile® growth trajectories and indirect evidence based on adjustment of performance standards using ACT Aspire® impact data.

Direct Evidence: Lexile® and Quantile® Growth Paths

In addition to using a student's lexile® and quantile® scores to select appropriate reading texts and mathematics instructional materials (described in the section for Legislative Criterion 1), these measures are also used to predict whether the student is likely by high school graduation to achieve lexile® and quantile® scores within the estimated ranges for postsecondary education and the workplace (i.e., CCR). These predictions are based on typical growth patterns for students in North Carolina who were followed longitudinally from Grade 3 through Grade 11.[44] Validity evidence of the overlap of the SC READY *meets expectations* performance levels and lexile® *stretched* grade level ranges or quantile® next grade level ranges needed to reach CCR by Grade 12 is presented after the sample reports in Exhibit A.[45] These graphs indicate that students who score *meets expectations* on SC READY ELA and Mathematics tests have sufficient achievement at each grade level to be on track for CCR by high school graduation.

MetaMetrics® conducted empirical research to develop the lexile® (1200L-1380L) and quantile® (1220Q-1440Q) CCR ranges by analyzing typical reading texts and mathematical materials used in postsecondary education. Workplace estimates were based on the typical requirement of a high school diploma, which was represented by measures of the typical instructional materials used in required terminal high school ELA and mathematics courses. In addition, typical materials encountered in selected entry-level occupational jobs have been analyzed and placed on the lexile® and quantile® scales. Lexiles® have been the most extensively studied creating separate estimates for university (1395L), community college (1295L), workplace (1260L), citizenship (1230L) and military (1180L) settings. For mathematics, as yet only the single CCR interval is being reported.[46]

The lexile® and quantile® estimated growth paths are based on scores from the SC READY ELA Reading and Mathematics assessments that measure the achievement of state CCR content standards. Estimated growth paths are reported in graphical form as shown in the SC READY lexile® and quantile® sample reports in Exhibit A. As shown in the sample reports, the student's current lexile® or quantile® point estimate is plotted on the graph (in blue) along with a dotted line (in blue) representing the predicted growth trajectory. The predicted growth trajectory is selected from among a set of typical student growth curves from a North Carolina norm group that best fits the current (and earlier grade level, if available) point estimate(s). If the estimated growth trajectory ends within the CCR interval at the end of Grade 12 (dark yellow shading), the student is predicted to achieve CCR by the end of Grade 12 (*see* Lexile® Sample Report in Exhibit A).

[44] MetaMetrics® (June 2017). *Aggregate Growth Curves for Lexile Growth Planner® & Quantile Growth Planner™: Technical Report* [MetaMetrics® Growth Report], Durham, NC: Author.
[45] *Lexile® Linking Study*, *supra* note 9, p. 53; *Quantile® Linking Study, supra* note 10, p. 62.
[46] MetaMetrics® *Growth Report*, *supra* note 42; Score Report User's Guide, *supra* note 7, p. 14-16.

If the end of the student's estimated growth path falls below the CCR interval, the graph will also plot a recommended growth path (in solid blue) that reflects the proportional accelerated improvement across the remaining grades that will be needed to reach the CCR interval by the end of Grade 12 (*see* Quantile® Sample Report in Exhibit A). The recommended growth path provides a target level of reading texts or mathematics lessons for the student to work toward in the next grade level and beyond. When the student is administered the SC READY assessments at the end of the next grade level, the actual lexile® and quantile® reported scores will verify whether the target level was achieved and a new estimated growth path will be plotted to re-evaluate whether the student is now on track for CCR at the end of Grade 12. If not, a new recommended growth path will also be plotted. The recommended growth paths are based on MetaMetrics research that developed a methodology for closing the gap between the typical lexile® and quantile® demands of high school reading texts and mathematics instructional materials and CCR requirements reported on the same scales.[47]

Indirect Evidence: Adjustment of Performance Standards Using ACT Aspire® Data

The SC READY individual student score reports state that "SC READY measures South Carolina's College- and Career-Ready Standards" (*see* Exhibit C). The ACT Aspire® test series measures the achievement of ELA and mathematics skills in Grades 3-8 linked to the corresponding score scales for the ACT Assessment college admissions test. The National Assessment of Educational Progress (NAEP) administers ELA and Mathematics tests biennially to a sample of students in Grades 4, 8 and 12 to track national progress in these subjects. Data from both these nationally-administered assessments was considered during the vertical moderation of the SC READY performance standards. Exhibited alongside the educator-recommended cut scores, this information provided a comparison of proficiency for students in South Carolina and that of students nationally. Using these comparisons, panelists had anchors for judging where proficiency should be set on the SC READY assessments linked to the South Carolina CCR content standards. The vertical moderation procedure used in standard setting for the SC READY assessments provided an indirect link to national CCR standards.

Evaluation

There is not a single, agreed-upon definition of college and career readiness (CCR) nationally. Consequently, it is difficult to identify a single, appropriate, national benchmark for college and career readiness (CCR). Different groups have attempted to define CCR based on considerations such as the ability to enroll in a credit bearing college course without needing remediation, mastery of ELA and mathematics skills that are prerequisite for commonly required freshman courses, or achievement of a sufficient score on a college admissions test to meet its CCR benchmarks or to be accepted to particular colleges or universities. In each case, empirical research is typically conducted to create linkages between test performance and postsecondary outcomes. The "career ready" part of CCR generally has received less attention because it is even more difficult to define than "college ready."

Many states have used college admissions test benchmarks, but they apply only to high school students and are problematic because they assess content that does not align very well with most state content standards. Related tests such as ACT Aspire®, designed for the elementary

[47] Sanford-Moore, E.E. & Williamson, G.L. (Oct. 2012). *Bending the Text Complexity Curve to Close the Gap,* Durham, NC: MetaMetrics®; Stenner, J., Sanford-Moore, E. & Williamson, G.L. (Oct. 2012). *The Lexile® Framework for Reading Quantifies the Reading Ability Needed for "College & Career Readiness,"* Durham, NC: MetaMetrics®; Williamson, G.L., Sanford-Moore, E. & Bickel, L. (July 2016). *The Quantile® Framework for Mathematics Quantifies the Mathematics Ability Needed for College and Career Readiness,* Durham, NC: MetaMetrics®.

and middle school grades, are linked to the corresponding college admissions test (ACT) but again may not be measuring the same knowledge and skills as the state content standards.

MetaMetrics® has taken a different approach by quantifying the complexity of reading text or mathematical materials typically encountered in entry-level college courses or jobs requiring a high school diploma. The SC READY assessments have been linked to the lexile® and quantile® scales that include a definition of CCR based on the complexity of reading texts and mathematics materials typically encountered in postsecondary education and workplace settings. In addition, national ACT Aspire® and NAEP data were consulted when the performance level standards were set for the SC READY assessments.

Currently, the empirical research available for lexiles® and quantiles® is generalized validity evidence from MetaMetrics® selected texts and instructional materials, a user norm group, the North Carolina longitudinal study, and the *meets expectations* comparisons in Exhibit A. However, there is no targeted validity data yet for South Carolina students. In the future, after several years of SC READY data have accumulated, it will be possible to evaluate the accuracy of the growth path predictions for South Carolina students. Until then, the state may want to consider reporting some indication of the possible error in these predictions, perhaps by surrounding the predicted growth path with error bands estimated from the North Carolina or source data used to develop the typical growth expectations on which they are based.

The validity data in Exhibit A linking SC READY *meets expectations* scores to the lexile® and quantile® on track for CCR target ranges provide persuasive evidence that longitudinal data collected for South Carolina will support current CCR predictions. In addition, the CCR ranges are based on several empirical research reports and the growth modeling is based on longitudinal data from a large sample of students from the neighboring state of North Carolina. The reasonableness of using North Carolina data to predict CCR outcomes for South Carolina students depends on the degree to which the content standards, assessments, and educational challenges in the two states are similar. Short of joining a testing consortium such as Smarter Balanced or PARCC, or conducting empirical research studies for South Carolina students that will require several years to complete, the lexile®/quantile® CCR trajectories derived from SC READY assessment scores are probably the best estimates currently available for evaluating whether South Carolina students are "on track for CCR."

5. Establishment of at Least Four Student Achievement Levels

Evidence

Evidence relevant to Legislative Criterion 5 includes the policy definitions and performance level descriptors for four student achievement (performance) levels and the standard setting activities that delimited the test score intervals corresponding to each of the four performance levels for the SC READY ELA and Mathematics assessments in Grades 3-8.

SC READY Student Achievement Levels

There are four student achievement levels for the SC READY assessment system. The four levels are labeled *exceeds expectations, meets expectations, approaches expectations,* and *does not meet expectations* and are defined based on the content standards and aligned items for ELA and Mathematics at each grade level. These four *achievement levels* are also known as *performance levels* because they are recommended by educators asked to identify the minimum level of test performance that would just barely be enough to approach, meet or exceed grade level expectations. The four, color-coded performance levels for the SC READY assessments are described by the following policy definitions.[48]

- ***Exceeds expectations* –** The student *exceeds expectations* as defined by the grade-level content standards. The student is *well prepared* for the next grade level and is *well prepared* for college and career readiness.
- ***Meets expectations* –** The student *meets expectations* as defined by the grade-level content standards. The student is *prepared* for the next grade level and is *on track* for college and career readiness.
- ***Approaches expectations*** – The student *approaches expectations* as defined by the grade-level content standards. The student *needs additional academic support* to ensure success in the next grade level and to be on track for college and career readiness.
- ***Does not meet expectations* –** The student *does not meet expectations* as defined by the grade-level content standards. The student *needs substantial academic support* to be prepared for the next grade level and to be on track for college and career readiness.

One might interpret the color coding from top to bottom as smooth sailing, good to go, cautious optimism, and stop, look and listen. Or if you like trains, fast-track, on track, decelerating, and derailed. But word play aside, more than just a global indicator is required to understand test performance and decide what to do next.

Performance Level Descriptors

To provide more specific information about the content knowledge and skills students are expected to master at each performance level, specific grade level ELA and Mathematics performance level descriptors (PLDs) were created. The PLDs are derived from the corresponding state content standards using the statements and indicators from the standards to create more detailed descriptions of the knowledge and skills that best characterize the expectations for the typical student scoring at each performance level.

Note that students are also expected to know and be able to do the PLD content at all the levels below the specified performance level. For example, students who score *meets expectations* should have mastered all the PLD content listed for the *meets, approaches* and *does not meet expectations* levels.

Using Grade 3 ELA Reading as an example, one can follow the progression of achievement across performance levels as described by the PLDs. Table 17 presents elaborated descriptions of the performance levels for Grade 3 ELA Reading.[49]

[48] SCDE (2016b). *Performance Level Descriptors (PLDs) Policy Statements,* Columbia, SC: Author.
[49] *Id.*, Grade 3 ELA PLDs.

Table 17. Progression of Grade 3 ELA Reading Achievement Across Performance Levels

Does Not Meet	Approaches	Meets	Exceeds
A student who performs at the **Does Not Meet Expectations** level tends to read and comprehend informational texts and literature that do not meet the demands of grade level texts that would signal this student is on track for CCR and requires substantial instructional support to improve reading skills.	A student who performs at the **Approaches Expectations** level tends to read and comprehend informational texts and literature of low-to-moderate complexity and sometimes struggles to meet the demands of grade level texts that would signal this student is on track for CCR and requires some instructional support to enhance reading skills.	A student who performs at the **Meets Expectations** level reads and comprehends informational texts and literature of moderate-to-high complexity and is meeting the demands of grade level texts that signal this student is on track for college and career readiness.	A student who performs at the **Exceeds Expectations** level reads and comprehends informational texts and literature of high complexity and is meeting and often exceeding the demands of grade level texts that clearly signal this student is on track for college and career readiness.

Source: SCDE, Grade 3 ELA Reading PLDs, www.scde.gov.

Drilling down to more specific Grade 3 ELA Reading skills, the PLDs also describe detailed progressions linked to specific state standards for student achievement across the four performance levels. Table 18 provides two examples.[50]

Table 18.Progression of Skills for Two Grade 3 ELA Reading Content Standards

Reading Standard	Does Not Meet	Approaches	Meets	Exceeds
RL.MC.6.1	*Identifies explicitly stated* themes by recalling details.	*Determines simple* themes by recalling *supporting* details.	*Determines* themes by recalling *supporting* details.	*Determines implicit* themes by recalling and analyzing *key supporting* details.
RL.LCS.11.1	*Identifies clearly stated* first or third person points of view.	*Explains* the differences between *clearly stated* first or third person points of view.	*Explains* the differences between first and third person points of view.	*Explains* the differences between *implied* first and third person points of view.

Source: SCDE, Grade 3 ELA Reading PLDs, www.scde.gov, emphasis added.

The skills listed in the Table 18 are only a sample of the 27 reading skill progressions contained in the PLD document for Grade 3 ELA. Similar progressions are also provided for the SC READY PLDs for the other grade/subject combinations.

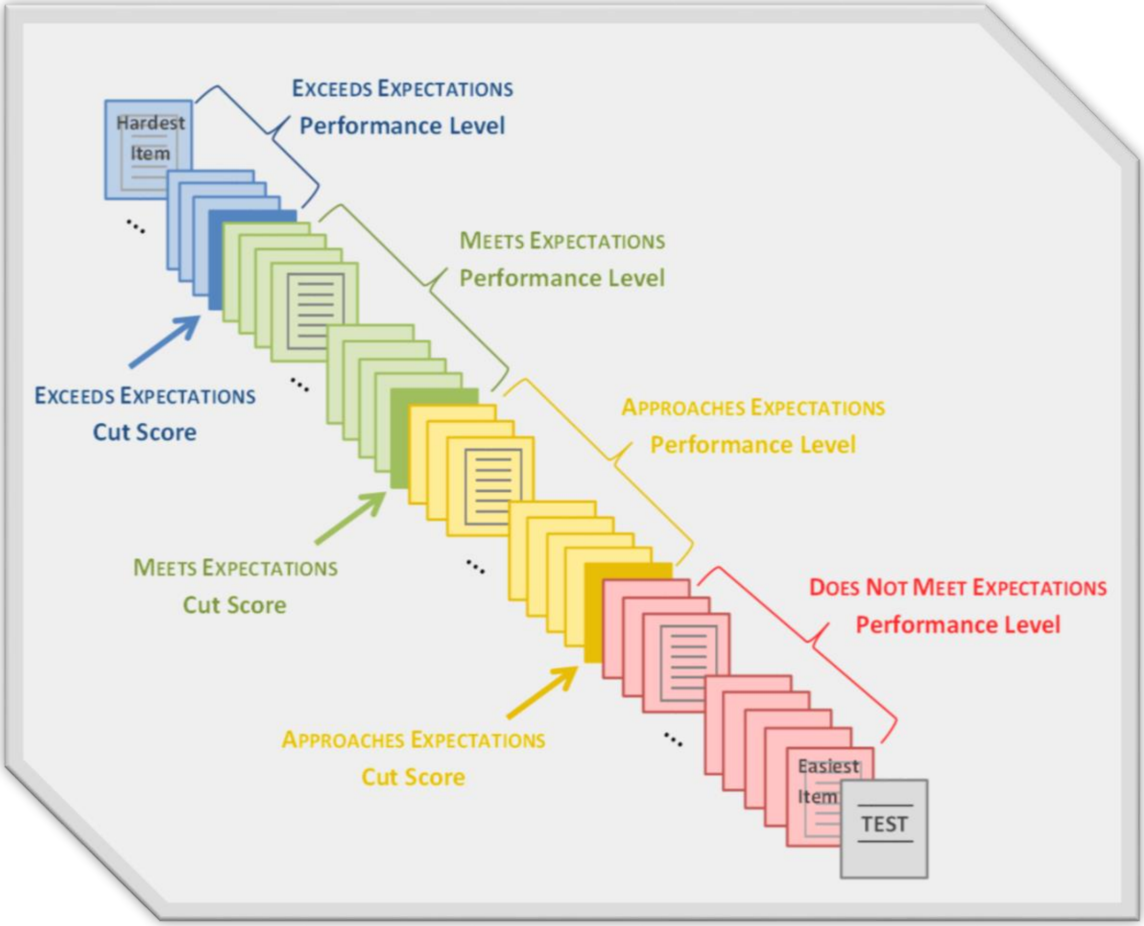
Standard Setting

The color-coded performance levels described above were created using a psychometric process called standard setting with the bookmark method. In June of 2016, after the spring field test was completed and the data analyzed, panels of South Carolina educators met to recommend cut scores that divided the test scores into intervals corresponding to the four performance levels. The items from a test booklet were ordered from easiest to hardest and the

[50] *Id.*, emphasis added.

educator panels were asked to recommend three points (cut scores) that marked the boundaries between *does not meet* and *approaches*, *approaches* and *meets*, and *meets* and *exceeds expectations* as shown in Figure 6. The next sections provide additional details about the composition, training and activities of these educator panels.

Figure 6. Setting Performance Standards with the Bookmark Method



Composition. Panel members were recruited by SCDE and the standard setting activities were conducted by the contractor. The composition of the ELA and Mathematics educator panels that recommended performance level standards is shown in Table 19 along with the average student composition for Grades 3-8 in the state.[51] Comparative gender and ethnic data are depicted graphically in Chart 7.

Compared to ELA, the Mathematics panel was more representative of the composition of the Grades 3-8 students in the state with 27% African-American and 65% White panelists for the Mathematics panel compared to 33% African-American and 51% White students for the state. Although Grades 3-8 students in South Carolina are approximately 9% Hispanic, there were no reported Hispanic panel members for either the ELA or Mathematics educator panels.

[51] DRC (2016b). *South Carolina SC READY Standard Setting Report,* Maple Grove, MN: Author, p. 2; Technical Manual, *supra* note 13, p. 21, mathematics test data.

*Table 19. Composition of Educator Panels and South Carolina Grades 3-8 Students**

Panel	N	Female	Ethnicity		Classroom Teacher	Mdn Yrs Teaching	Students		
			AA	W			AA	H	W
ELA	34	97%	15%	82%	15%	25	33%	9%	51%
							SD	EL	F
Math	37	89%	27%	65%	57%	20	15%	4%	49%

* N=number, AA=African-American, W=White, H=Hispanic, SD=students with disabilities, EL= English learners, F=female; *Source:* Standard Setting Report, p. 2; SC READY Technical Manual, p. 21, Mathematics test data.

The majority of Mathematics panel members were current classroom teachers (57%) but only a small percentage of ELA panel members were teachers (15%). However, an additional 49% of the ELA panelists were educators. The gender composition of both panels was predominately female, probably reflective of the teaching pool in the state but much greater than the near even gender distribution of Grades 3-8 students in the state. The number of panel members with experience teaching students with disabilities or English learners was not reported.

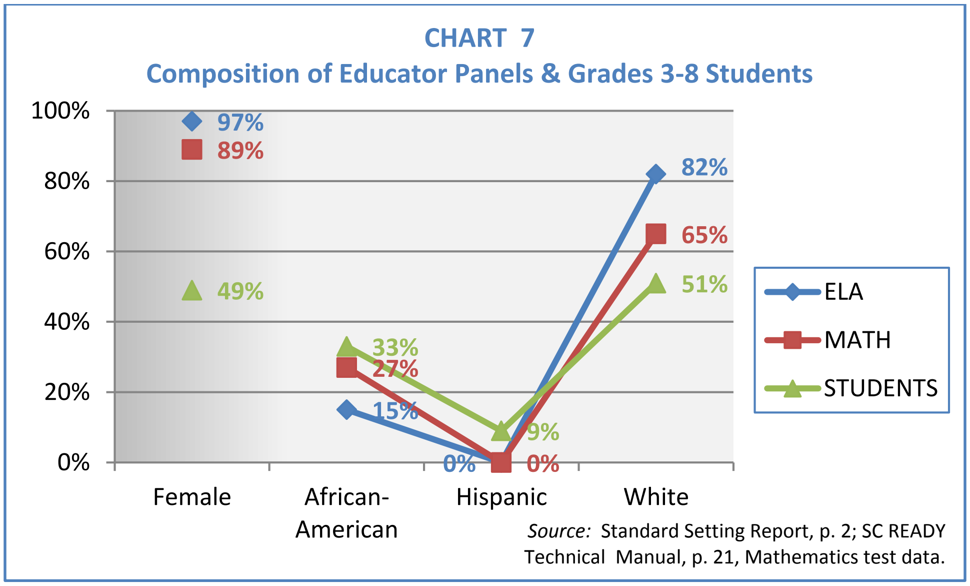


CHART 7

Composition of Educator Panels & Grades 3-8 Students

Source: Standard Setting Report, p. 2; SC READY Technical Manual, p. 21, Mathematics test data.

Training. The meeting to recommend performance standards lasted four days. ELA and Mathematics panel members met separately and were each split into three groups responsible for Grades 3-4, Grades 5-6, and Grades 7-8, respectively. Before beginning their work, the educator panel members experienced the Spring 2016 operational test for their subject/grade to learn about the types and difficulty of the items. After taking the test, panel members received training to familiarize them with the state content standards and performance level descriptors for their subject/grade, and then they discussed the characteristics of students just barely approaching, meeting or exceeding the grade level expectations. Finally, they were trained on the procedures they would use to provide their recommendations and completed a practice exercise.

Activities. Panel members provided their recommendations using a psychometric method known as the Bookmark procedure.[52] In the Bookmark procedure, the items from the 2016 operational forms were reordered from easiest to hardest and placed in a booklet with one test item per page. Panelists were then asked to begin with page one and bookmark the last page where 67% of students just barely passing into the meets expectations performance level would answer the item correctly. The placement of this bookmark defined the cut score for the *meets expectations* performance level. This process was then repeated to identify the cut scores for the *approaches* and *exceeds expectations* performance levels. This procedure is depicted graphically in Figure 6 where the dark, solid-colored pages identify the three decision points.

Multiple Rounds. Panelists completed three rounds of bookmark placements. After each round, the workshop leader shared the distribution and median bookmark placements of panel members and facilitated a discussion of the correspondence between the skills required to answer the items correctly and the requirements described in the PLDs for each performance level. After the first round, the leader also shared the state percent of students correctly answering each item in the booklet (p-values) for panelists to use when reconsidering their bookmarks. For example, an item answered correctly by only 40% of all students would be unlikely to be answered correctly by 67% of students just barely meeting expectations. After panelists completed the second round of recommendations, the leader again facilitated discussion and shared impact data to assist panelists in evaluating whether their bookmarks were realistic for the student population. Impact data quantify the percent of students classified in each performance level given the median bookmark placements of the panelists.

Workshop Evaluation. After completing the final (third) round of bookmark placements, panelists were asked to evaluate the quality of the training provided and their confidence level in placing their bookmarks. Regarding training, 75% or more of the panelists responded positively and felt that the amount of time allotted to the various activities was about right. Overall for ELA, panelists were 75% or more confident of their bookmark placements with slightly less confidence for the *approaches* cut score and more confidence for the *exceeds* cut score. For Mathematics in the lower grades, 80% or more were confident of their bookmark placements while 63-75% were in the upper grades. When asked whether the procedures used will produce appropriate results, 97% of ELA and Mathematics panelists said "yes." When asked if their bookmark placements accurately represented the PLDs, all of the ELA and 97% of the Mathematics panelists said "yes."

Vertical Moderation. After the standard setting workshops concluded, a vertical moderation meeting was held to smooth and adjust the uneven results from the educator panels. The goal was to create continuity and consistency across grades in line with policy goals. Participants recruited for the vertical moderation workshops included five panelists from the ELA and eight panelists from the Mathematics standard setting workshops. Demographic information for these individuals was not reported.

The vertical moderation workshop was conducted in one day and included an introductory presentation of the goals of the activity and two rounds of recommended adjustments. The criteria for adjustments included:

- Consistency with policy goals
- Legal defensibility
- Stakeholder acceptance, and
- Efficient use of remediation resources.

[52] *See* Lewis, Mitzel, Green, & Patz (2000). *The Bookmark Standard Setting Procedure,* Monterey, CA: CTB/McGraw Hill.

The emphasis for the standard setting meetings was on content while the emphasis for the vertical moderation meeting was on impact. Participants were asked to consider common policy definitions, statistical measures of uncertainty, external sources of performance data and a preference for relatively monotonic trend lines across grades. The external performance data available to participants included ACT Aspire® and NAEP results.

Three quarters (6 of 8) of the Mathematics participants agreed with the final recommendations and the other two agreed with written exceptions they provided on their recommendation forms. For ELA, one participant agreed with the final recommendations and the other four agreed with exceptions listed on their forms. The results of the vertical moderation meetings were communicated to SCDE, reviewed and finalized. The contractor then used the Rasch model to determine the final scale score cut scores for each performance level shown in Table 20.

Table 20. Final Cut Scores and 2016 Impact Data

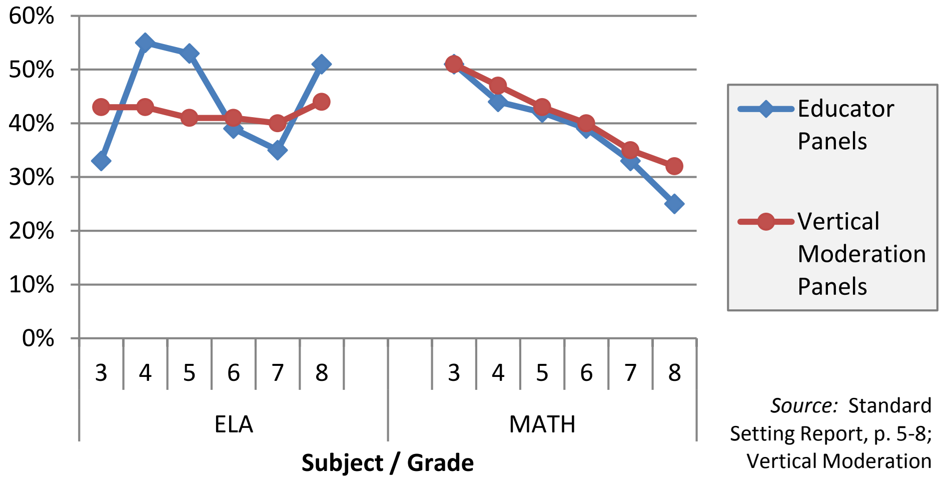
	OIB*	CUT SCORES ITEM NUMBER / PERCENT			STUDENT IMPACT				MEETS + EXCEEDS	
	ITEMS	Approaches	Meets	Exceeds	Not Meet	Appr	Meets	Exceeds	Educ Panel	Vert Mod
ELA Grade 3	72	9 / 13%	28 / 39%	53 / 74%	23%	34%	29%	14%	**33%**	**43%**
4	72	11 / 15%	29 / 40%	55 / 76%	25%	32%	29%	14%	**55%**	**43%**
5	72	15 / 21%	36 / 50%	61 / 85%	24%	35%	28%	13%	**53%**	**41%**
6	83	14 / 17%	40 / 48%	64 / 77%	22%	38%	27%	13%	**39%**	**41%**
7	84	13 / 15%	40 / 48%	71 / 85%	23%	37%	27%	13%	**35%**	**40%**
8	84	13 / 15%	41 / 49%	65 / 77%	23%	33%	30%	13%	**51%**	**44%**
Math Grade 3	50	12 / 24%	28 / 56%	39 / 78%	21%	28%	31%	20%	**51%**	**51%**
4	55	8 / 15%	24 / 44%	41 / 75%	23%	31%	26%	21%	**44%**	**47%**
5	55	10 / 18%	22 / 40%	41 / 75%	23%	34%	25%	19%	**42%**	**43%**
6	61	8 / 13%	23 / 38%	40 / 66%	25%	36%	23%	17%	**39%**	**40%**
7	61	7 / 11%	22 / 36%	44 / 72%	27%	39%	19%	16%	**33%**	**35%**
8	65	7 / 11%	22 / 34%	44 / 68%	29%	39%	18%	14%	**25%**	**32%**

* OIB=Ordered Item Booklet (*see* Figure 6) with one item per page; the ELA TDA essay item had four pages, one for each score point; *Source:* Standard Setting Report, p. 5-8; Vertical Moderation Report, Appendix I.

Results. Recall that the items in each Ordered Item Booklet (OIB) used with the educator panels contained essentially Spring 2016 operational test forms. The items in the OIB were ordered from easiest to hardest with one item per page. For the ELA TDA essay item, there were four pages representing the difficulty of achieving each of the four possible score points.

In the parlance of accountability, the performance levels can be grouped into two categories, proficient = *meets* + *exceeds expectations*, and not proficient = *does not meet* + *approaches expectations*. Using these categories and the full four performance levels, Table 20 summarizes the results of the standard setting process, including both final cut scores and student impact. The final two columns contrast the impact of the panel recommendations with the final vertical moderation recommendations in terms of the estimated percent of students labeled proficient. These latter statistics are compared graphically in Chart 8.[53] The data in Chart 8 indicate that larger adjustments were made for ELA than for Mathematics.

CHART 8
Impact on Percent of Proficient Students for
Educator and Vertical Moderation Recommendations



Source: Standard Setting Report, p. 5-8; Vertical Moderation Report, Appendix I.

Consistency with Test Standards. HumRRO evaluated the standard setting activities for SC READY against a selected set of relevant Standards from the 2014 *Test Standards.* The Standards, a short description, the staff ratings and comments are presented in Table 21. All three Standards received high ratings due to the good documentation, use of impact data, reporting of conditional standard errors of measurement at the cut scores, training of panelists, three rounds of ratings, workshop evaluations, vertical moderation, and use of external data and policy goals for informing adjustments.

***ELA Reading*.** Beginning in the 2017-18 school year, the Act 284 Read to Succeed legislation requires students to be retained in third grade if they fail to demonstrate minimal grade level reading proficiency. On the state's prior reading test, SC PASS, the lowest level of proficiency, Not Met 1, had been designated as the criterion for identifying students with insufficient reading

[53] Standard Setting Report, *supra* note 49, p. 5-8; Vertical Moderation Report, *supra* note 28, Appendix I.

proficiency. These students were required to attend a reading summer camp and were then retested to determine whether they had met the standard and could proceed to fourth grade.[54]

Table 21. Ratings of SC READY Consistency with Identified Standards for Setting Cut Scores

Standard	Description	Rating	Comments
5.21	Clearly documented procedures for establishing cut scores	5	Standard Setting, Vertical Moderation and TM Reports; bookmark method with impact data, confidence intervals; classification consistency good; cSEM at cut scores
5.22	Methods for recommending cut scores that permit educators to apply their knowledge and experience reasonably	5	Panelists experienced with grade level/subject; took operational test, trained, practiced, discussed three rounds of ratings/PLDs; considered impact data; positive evaluations at end of workshops
5.23	Substantive interpretations of performance levels informed by empirical data	4	Vertical moderation increased consistency across grades; SC impact data, 2015 ACT Aspire® and NAEP considered

* cSEM=conditional standard error of measurement; TM=Technical Manual; SRUG=Score Report User's Guide
Source: Chapter 5 (Task 5), Standard Setting Report, Vertical Moderation Report, Technical Manual.

In Feb. 2017, a two-day standard setting meeting was held in Columbia, South Carolina to set a corresponding Read to Succeed minimum reading performance standard for the new SC READY Grade 3 Reading assessment. Educators were asked to judge the level of performance equivalent to the *Not Met 1* cut score from the prior Grade 3 Reading test.

The meeting included 25 panelists recruited by SCDE to represent a variety of South Carolina educators. The group was 96% female, 56% White, 40% African-American, 68% classroom teachers and 60% from rural schools. Similar to the other Bookmark standard setting meetings, performance level descriptors (PLDs) were developed to describe the skills of the student with just barely enough reading skill to meet the minimum standard for promotion to fourth grade.[55]

Also similar to the previous standard setting workshops, the panelists were trained and completed three rounds of the Bookmark procedure using an ordered item booklet (OIB) with 49 Grade 3 Reading items ordered from easiest to hardest. These items consisted of the 2016 SC READY Grade 3 operational reading items plus a few aligned items from the prior SC PASS Grade 3 Reading test. Panelists were asked to start at the beginning of the OIB and identify the last item for which 50% of the students described by the PLDs as barely achieving the reading proficiency represented by the *Not Met 1* cut score would correctly answer the item. After each round of individual bookmark placements, tables of panelists received group feedback and discussed their choices. After Round 2, 2016 impact data were also shared with the panelists.[56]

The results suggested that the impact data were given serious consideration. The median of the table medians of OIB bookmarked page numbers was 12 for Round 1, 10 for Round 2 and after viewing impact data, six for Round 3. At the recommended Round 3 cut score, 5.3% of 2016 Grade 3 students were estimated to be identified for retention. In their evaluations, all panelists were confident or very confident of their bookmark choices, and more than 70% strongly agreed

[54] Reading Grade 3 Standard Setting Report, *supra* note 33, p. 1; SCDE Responses to Questions, *supra* note 6, p. 8-9.
[55] Reading Grade 3 Standard Setting Report, *supra* note 33, p. 2-3.
[56] *Id.*, p. 3-4.

that the process produced appropriate results and accurately represented the PLDs.[57] The cut scores for the other performance levels of the Reading subtest were determined using the ability estimates corresponding to the cut scores established for the total ELA assessment.[58]

*Table 22. Decision Consistency Estimates for SC READY Performance Levels**

	Four Performance Levels								Two Performance Levels							
	Total	Gender		Ethnic					Total	Gender		Ethnic				
		F	M	AA	H	W	EL	SD		F	M	AA	H	W	EL	SD
ELA Grade 3	.73	.72	.73	.74	73	.72	.73	.81	.88	.88	.89	.89	89	.88	.89	.93
4	.70	.68	.71	.73	.70	.68	.70	82	.88	.87	.88	.90	.88	87	.88	.94
5	.74	.73	.75	.76	.74	.73	.75	.83	.89	.88	.89	.91	.89	.88	.89	.96
6	.77	.76	.77	.78	.77	.76	.77	.83	.90	.89	.91	.92	.90	.89	.91	.96
7	.77	.76	.78	.79	.77	.75	.78	.85	.90	.90	.91	.93	.91	.89	.91	.97
8	.79	.76	.78	.79	.77	.76	.78	.86	.90	.90	.91	.92	.91	.90	.91	.97
Mean	.75	.74	.75	.77	.75	.73	.75	.83	.89	.89	.90	.91	.90	.89	.90	.96
MATH Grade 3	.70	.69	.71	.70	.69	.71	.69	.75	.88	.87	.88	.87	.87	.88	.87	.90
4	.71	.70	.72	.70	.70	.71	.70	.75	.88	.87	.88	.87	.87	.88	.87	.91
5	.71	.70	.72	.72	.70	.71	.71	.78	.88	.88	.89	.89	.88	.88	.88	.93
6	.73	.72	.74	.73	.72	.73	.73	.79	.89	.88	.89	.90	.88	.89	.89	.95
7	.72	.71	.73	.74	.71	.71	.72	.78	.88	.88	.89	.91	.89	.88	.89	.96
8	.72	.71	.73	.73	.71	.70	.72	.80	.88	.88	.89	.90	.88	.88	.89	.96
Mean	.72	.71	.73	.72	.71	.71	.71	.78	.88	.88	.89	.89	.88	.88	.88	.94

* AA=African-American; H=Hispanic; W=White; EL=English learners; SD=students with disabilities
Source: Technical Manual, p. 45-48.

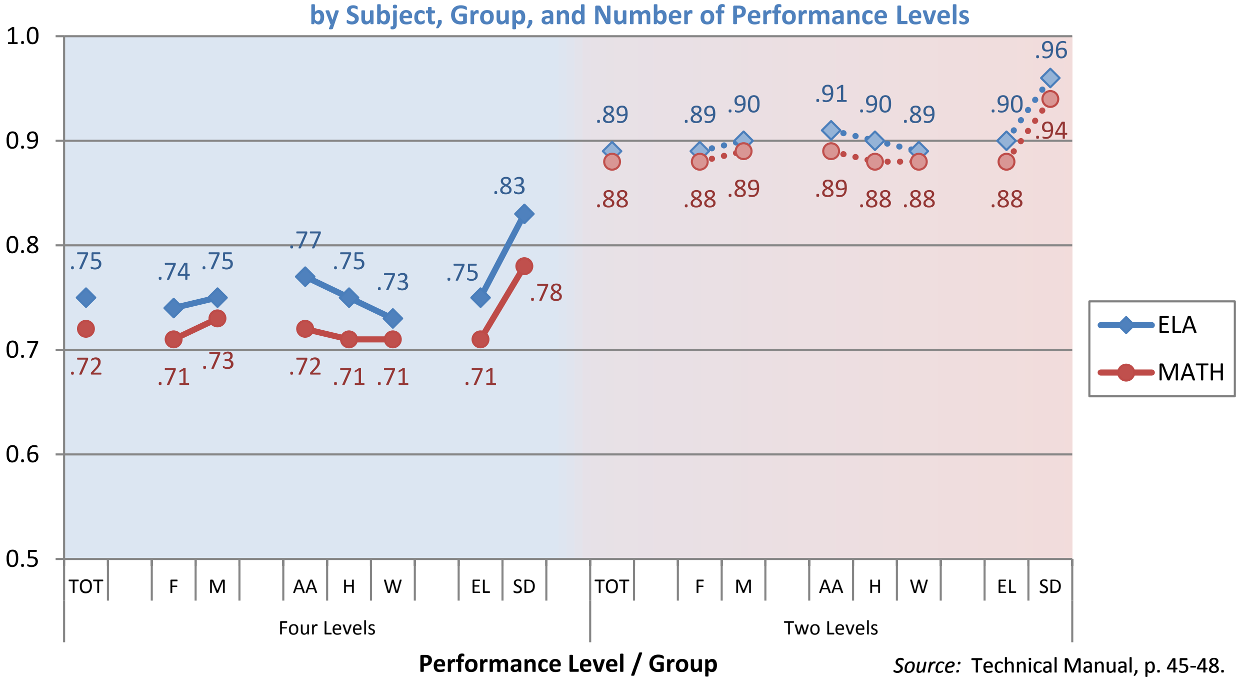
[57] *Id.*, p. 4-5, 17-18.
[58] DRC Response to Questions, *supra* note 6, p. 8.

Decision Consistency of Performance Level Classifications

The consistency with which the SC READY assessments are predicted to classify students in the same performance level if they were to retest under similar conditions is quantified by statistical estimates of *decision consistency*. Table 22 presents decision consistency estimates for the SC READY ELA and Mathematics tests by group (total, gender, ethnic, English learners & students with disabilities. The mean (average) decision consistency estimates are also presented graphically in Chart 9.

The statistics reported in Table 22 are the proportion of agreement for all four achievement levels (blue) or the proportion of agreement for the dichotomous proficient / not proficient classification used for federal ESSA accountability (purple). The proportions of agreement were calculated using a statistical model developed by Huynh[59] that provides consistency estimates based on a single administration of a test.

CHART 9
SC READY Average Decision Consistency Estimates for Grades 3-8
by Subject, Group, and Number of Performance Levels



Source: Technical Manual, p. 45-48.

For example, in Table 22, 73% of all third grade students who were administered the SC READY ELA test in Spring 2017 would be expected to be classified in exactly the same performance level (*exceeds, meets, approaches, does not meet expectations*) if retested under similar circumstances. If only two performance levels (proficient=*exceeds*+*meets* and not proficient=*approaches*+*does not meet*) are reported, 88% are estimated to be classified in exactly the same level. For Mathematics, the corresponding estimates are 70% and 88%. As the data in Table 22 and Chart 9 indicate, these estimates are similar across groups.

[59] Huynh, H. (1979). Computational and Statistical Inference for Two Reliability Indices Based on the Beta-Binomial Model. *Journal of Educational Statistics*, 4:231-46.

As indicated in the rows labeled “Mean” in Table 22, the average decision consistency estimates across grades with four performance categories ranged from .73 to .83 for ELA and .71 to .78 for Mathematics. The corresponding values for two performance levels are .71-.78 for ELA and .88-.94 for Mathematics. Again, the mean values are similar across groups with students with disabilities scoring slightly higher than the other groups.

Evaluation

The SC READY assessments include four performance levels, two that signify proficiency and two that do not. Each of the performance levels is described by general policy statements related to the subject matter and by more specific performance level descriptors related to the state content standards. There is good documentation of the standard setting activities that recommended cut scores to delimit the four performance levels on the test score scales, and decision consistency estimates were high, especially for two performance categories.

The standard setting activities included educator panels that recommended cut scores based on the state content standards and a vertical moderation panel that adjusted the cut scores based on policy goals. The educator panels were composed primarily of teachers or former teachers who were best qualified to judge the content demands of the test items. The SBE, with the advice and consent of the EOC per Section 59-18-320(D), should officially adopt the cut scores.

Because vertical moderation panels are asked to judge the recommended cut scores based on policy goals, their members usually include representatives of external stakeholder groups with policy expertise such as legislative staff, advocacy groups, parents, teachers’ unions, business leaders, and administrators. Also, representatives from the educator panels typically are invited to the vertical moderation meeting to share their educational perspectives. For the SC READY assessments, the vertical moderation panels were subsets of the educator panels and may not have had sufficient policy expertise and diversity to adequately represent the variety of South Carolina stakeholder perspectives and policy concerns of groups whose support is beneficial.

6. Inclusion of a Variety of Question Types that Test Student Understanding of the Content

Evidence

Evidence relevant to Legislative Criterion 6 includes item types and scoring, confirmatory studies, timing, item quality, item alignment, forms construction and field testing of replacement items.

Item Types and Scoring

There are six different question types utilized in the SC READY assessments. Each is designed to address a different type of student understanding of the content and is matched to the subject matter, ELA or Mathematics, and the grade levels for which it is most appropriate. Table 23 provides an overview of the item types, subject matter, grade levels and content understanding for which each item type is used in the SC READY assessments.

The next sections contain descriptions of each of the item types and their scoring. For purposes of illustration, the examples are drawn from sample questions provided online for the two tests that were reviewed for this chapter of the report, ELA Grade 3 and Mathematics Grade 8. ELA Grade 3 was chosen because its results are used to satisfy the Read to Succeed statutory requirements for student promotion and retention. Grade 8 Mathematics was chosen because its content is closest to CCR requirements and because it contains the most diverse and complex item content and item types.[60]

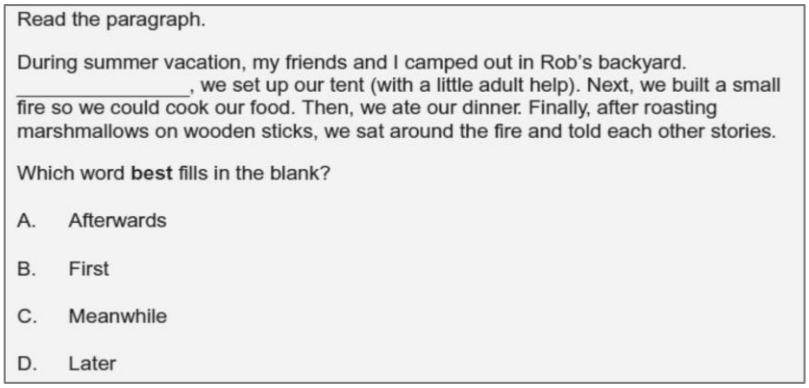
Table 23. Summary of SC READY Item Types

ITEM TYPE	SUBJECT MATTER	GRADE LEVELS	CONTENT UNDERSTANDING
Selected Response: Multiple-Choice	ELA & Mathematics	3 - 8	*Recognize* a correct answer
Selected Response: Multi-Select	ELA & Mathematics	3 - 8	*Distinguish* multiple correct and incorrect answers
Evidence-Based Selected Response	ELA	3 - 8	*Use* evidence from a text to justify and support an answer
Short Answer or Gridded Response	Mathematics	6 - 8	*Supply* a correct answer by typing (online) or filling out a number grid (paper)
Technology Enhanced	ELA & Mathematics	6 - 8	ONLINE ONLY: *Create* a correct answer by drag & drop options, clicking on a spot, or graphing
Text-Dependent Analysis	ELA Session 1	3 - 8	*Write* an extended response supported by evidence from a text passage

Source: Technical Manual, p. 10-11.

Selected Response: Multiple-Choice. Selected response items include traditional multiple-choice, multi-select, and evidenced-based items. A traditional multiple choice item, shown in Example 1, consists of a question with four possible answer choices. The student is instructed to choose the single best answer. Multiple choice items are each worth one point.

Example 1: Grade 3 ELA Multiple-Choice Item



Read the paragraph.

During summer vacation, my friends and I camped out in Rob's backyard. ________________, we set up our tent (with a little adult help). Next, we built a small fire so we could cook our food. Then, we ate our dinner. Finally, after roasting marshmallows on wooden sticks, we sat around the fire and told each other stories.

Which word **best** fills in the blank?

A. Afterwards

B. First

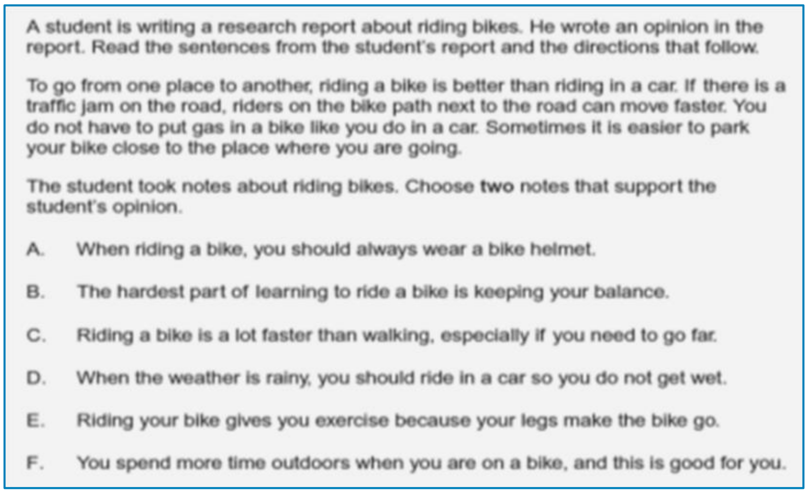
C. Meanwhile

D. Later

[60] The source for the Grade 3 ELA sample items was the SCDE website, www.ed.sc.gov, and for the Grade 8 Mathematics sample items was the online test tutorial (OTT). Note: the short answer and technology enhanced examples from the OTT were labeled Grades 6-8 and may actually represent content from any of the three grades.

Selected Response: Multi-Select. Multi-select items consist of a question with more than four answer choices. The student is instructed to choose a specific number of correct answers or to select ALL correct answer choices. Multi-select items are also worth one point but credit is awarded only if the student selects all the correct responses and no incorrect responses. Example 2 is a multi-select item that measures content from the *inquiry* ELA standards.

Example 2: Grade 3 ELA Multi-Select Item (Inquiry)



A student is writing a research report about riding bikes. He wrote an opinion in the report. Read the sentences from the student's report and the directions that follow.

To go from one place to another, riding a bike is better than riding in a car. If there is a traffic jam on the road, riders on the bike path next to the road can move faster. You do not have to put gas in a bike like you do in a car. Sometimes it is easier to park your bike close to the place where you are going.

The student took notes about riding bikes. Choose **two** notes that support the student's opinion.

A. When riding a bike, you should always wear a bike helmet.

B. The hardest part of learning to ride a bike is keeping your balance.

C. Riding a bike is a lot faster than walking, especially if you need to go far.

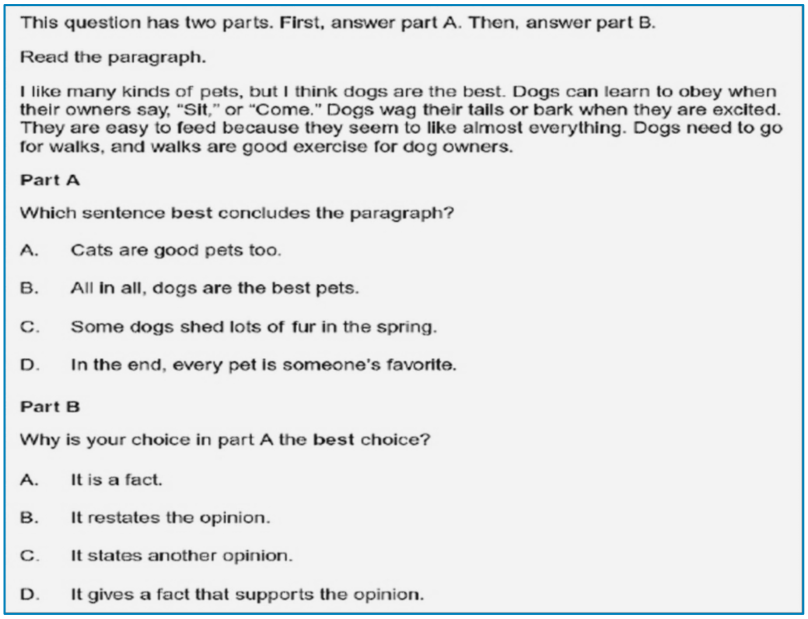
D. When the weather is rainy, you should ride in a car so you do not get wet.

E. Riding your bike gives you exercise because your legs make the bike go.

F. You spend more time outdoors when you are on a bike, and this is good for you.

Evidence-Based Selected Response. Evidence-based selected response questions are two-part, multiple-choice questions that appear only on the ELA assessments. Part A of the question asks students to respond to a multiple-choice question about a text passage. Part B is another multiple-choice question that asks the student to select the evidence from the text that best supports the answer chosen for Part A. Evidence-based questions are worth one point and students are required to select a correct answer to both parts to receive credit. Example 3 is an evidence-based, selected response item.

Example 3: Grade 3 ELA Evidence-Based Selected Response Item



This question has two parts. First, answer part A. Then, answer part B.

Read the paragraph.

I like many kinds of pets, but I think dogs are the best. Dogs can learn to obey when their owners say, "Sit," or "Come." Dogs wag their tails or bark when they are excited. They are easy to feed because they seem to like almost everything. Dogs need to go for walks, and walks are good exercise for dog owners.

Part A

Which sentence best concludes the paragraph?

A. Cats are good pets too.

B. All in all, dogs are the best pets.

C. Some dogs shed lots of fur in the spring.

D. In the end, every pet is someone's favorite.

Part B

Why is your choice in part A the best choice?

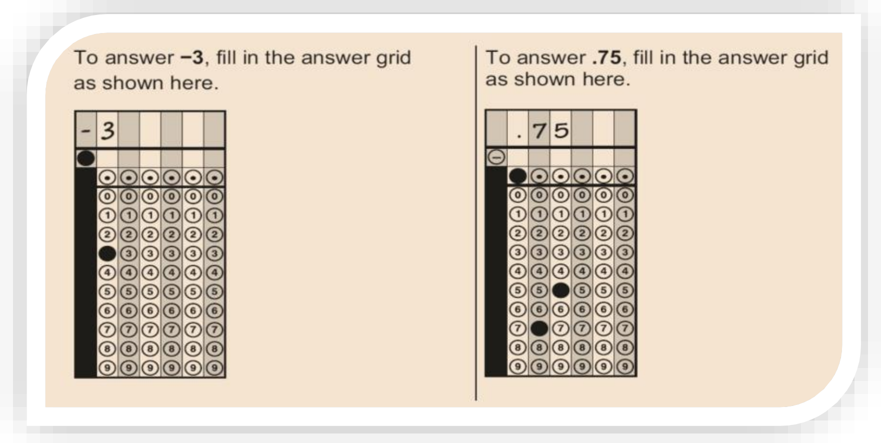
A. It is a fact.

B. It restates the opinion.

C. It states another opinion.

D. It gives a fact that supports the opinion.

Example 4: Grades 6-8 Mathematics Response Grid (Paper/Pencil)

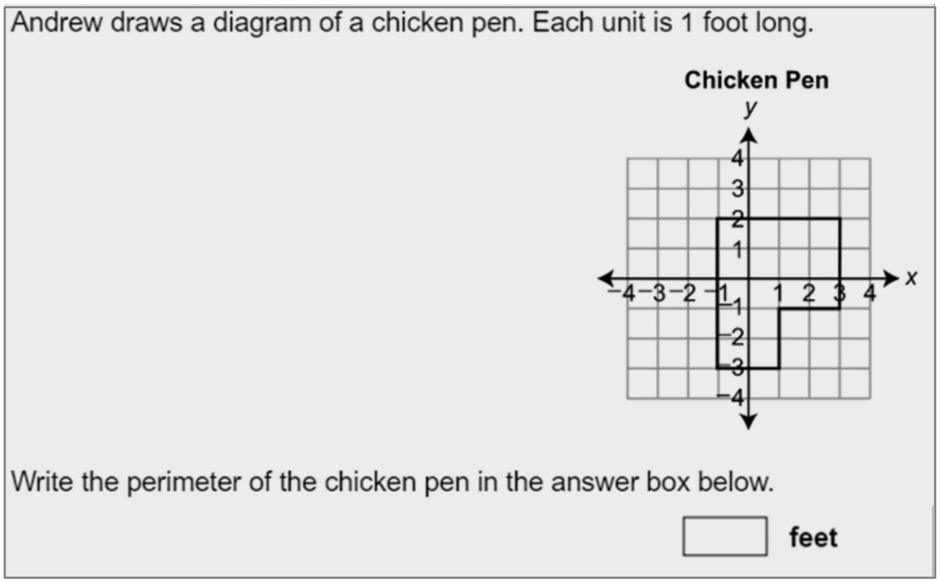


To answer **−3**, fill in the answer grid as shown here.

To answer **.75**, fill in the answer grid as shown here.

Short Answer (SA) or Gridded Response (GR). These question types require students to solve a mathematics problem and supply the correct answer. For online tests, students type in their responses from the keyboard. For paper/pencil tests, students use a numerical grid to record their answers. Unlike multiple-choice questions where students can look at the answer choices and guess which one is correct, short answer and gridded response questions require students to construct their own answers without any help from a list of choices. Example 4 explains how to fill out a paper/pencil numerical grid and Example 5 illustrates an online item where students are required to use the computer keyboard to type in their numerical answers.

Example 5: Grades 6-8 Mathematics Key Entry Item (Online)



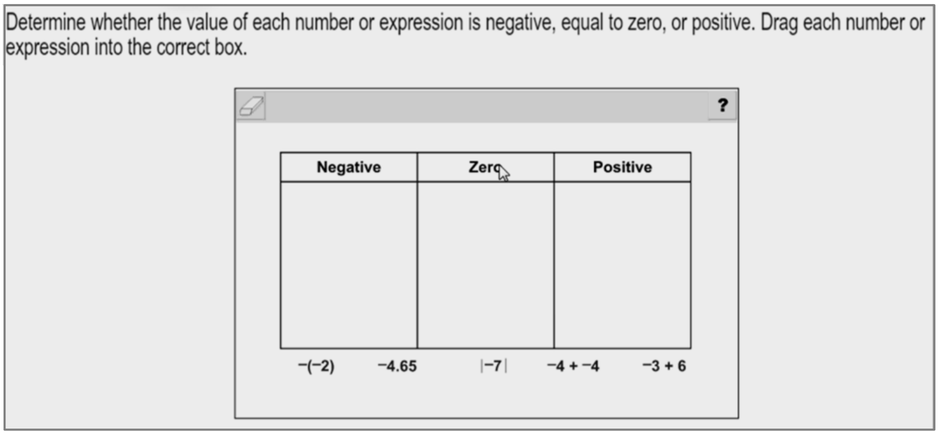
Andrew draws a diagram of a chicken pen. Each unit is 1 foot long.

Write the perimeter of the chicken pen in the answer box below.

□ feet

Technology Enhanced. This item type appears only on online forms of the test. Students interact with the question through the contractor’s testing platform by clicking on a particular spot in an illustration, arranging response options in order, matching responses to descriptions or moving responses that satisfy certain conditions to a box below the question. The latter interaction requires students to click on a response or object and drag it to the appropriate place on the screen according to the specific directions provided in the question. Selected response or multi-select questions replace technology enhanced question types on the paper/pencil test forms. Example 6 illustrates an online technology enhanced item.

Example 6: Grades 6-8 Mathematics Technology Enhanced Item (No Calculator)



Determine whether the value of each number or expression is negative, equal to zero, or positive. Drag each number or expression into the correct box.

Negative	Zero	Positive

−(−2) −4.65 |−7| −4 + −4 −3 + 6

Text-Dependent Analysis (TDA). Text-dependent analysis questions are essay questions that require two types of skills, writing skill and connection of the response to specific information contained in an associated text passage. There is exactly one TDA question in Part I of the ELA test for each grade. Example 7 is a TDA item alongside a sample student response. The writer's checklist that appears with this item is reproduced in Exhibit E.

Holistic Scoring. TDA questions are scored by two raters. Scoring guidelines for the TDA question present a holistic rubric with the four possible score points shown below. Additional detail is contained in the rubric which is available for students to review while they are writing. The scoring rubric for the TDA items is reproduced in Exhibit F.

4 – Demonstrates ***effective*** analysis of text and ***skillful*** writing

3 – Demonstrates ***adequate*** analysis of text and ***appropriate*** writing

2 – Demonstrates ***limited*** analysis of text and ***inconsistent*** writing

1 – Demonstrates ***minimal*** analysis of text and ***inadequate*** writing.[61]

A zero score is given if the response is unscorable. There are separate codes for blank, off topic, in a language other than English, unreadable, insufficient, copied, and refused to answer responses so the reason for the zero score can be printed on the student's score report. To provide increased weight to the ELA score for the authentic writing represented by the TDA question, the average rating is multiplied by four to produce a total scale of 0-16 points possible for the TDA question.

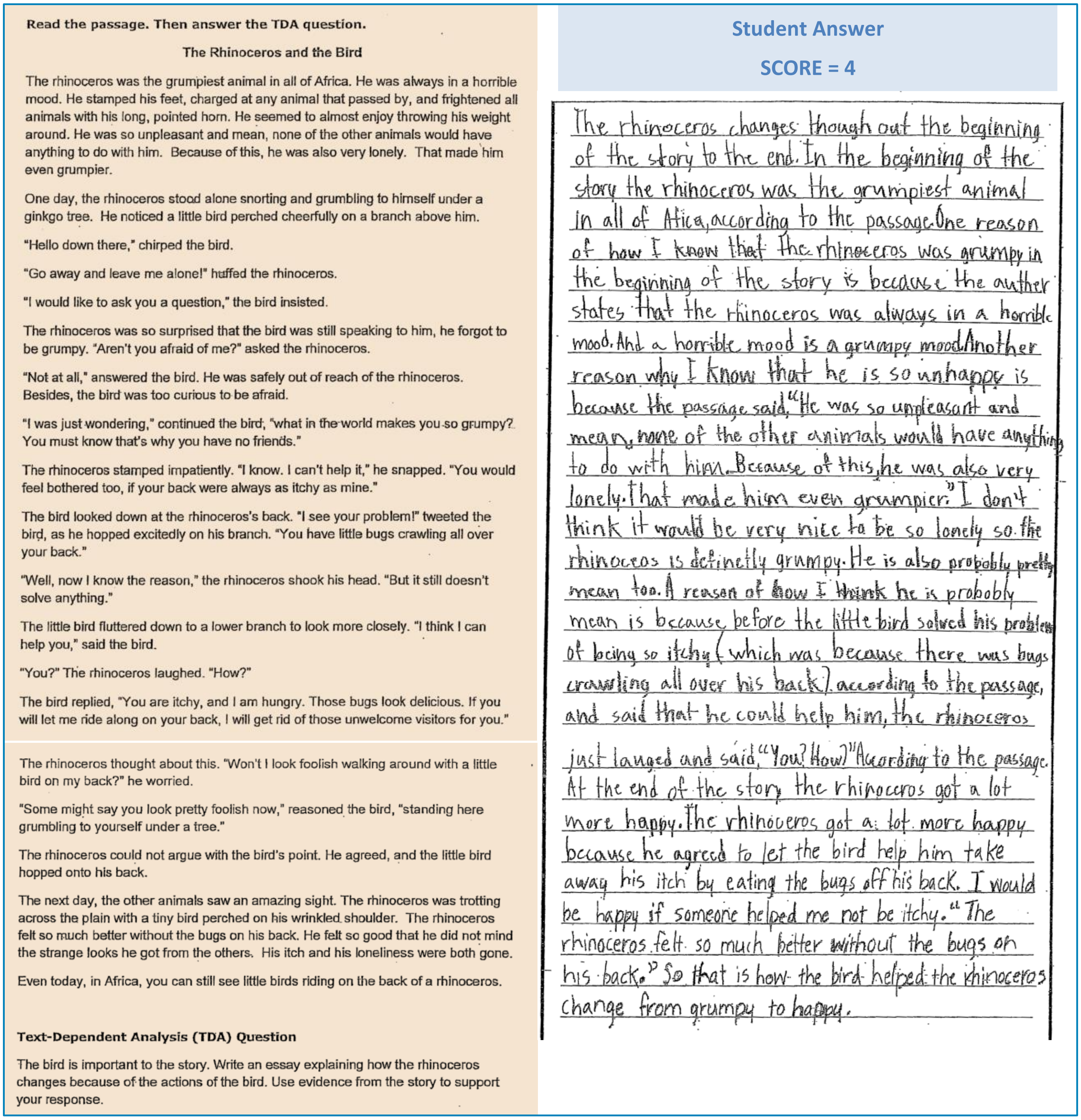
Two raters score each TDA response on the 0-4 scale described above. Raters receive extensive training, including application of the bullet points for each score provided in the scoring guide, anchor papers of actual student responses judged to be barely, at, and the top of each score, guided practice applying the rubric, and a qualifying round requiring 70% exact

[61] Score Report User's Guide, *supra* note 7, Appendix C, emphasis added.

agreement to begin rating responses. During scoring, team leaders monitor rater agreement and periodically assign disguised validity packets of papers with known scores to check rater accuracy. Additional review and retraining are provided as needed. If the two ratings for a response differ by more than one point, a third rating is obtained.

The contractor is required to maintain at least 70% exact agreement throughout the scoring process. Rater agreement statistics were presented in the *Reliability* section of Legislative Criterion 2 and indicate the 70% exact agreement target was exceeded for all grades.

Example 7: Grade 3 ELA Text-Dependent Analysis (TDA) Essay Item



Read the passage. Then answer the TDA question.

The Rhinoceros and the Bird

The rhinoceros was the grumpiest animal in all of Africa. He was always in a horrible mood. He stamped his feet, charged at any animal that passed by, and frightened all animals with his long, pointed horn. He seemed to almost enjoy throwing his weight around. He was so unpleasant and mean, none of the other animals would have anything to do with him. Because of this, he was also very lonely. That made him even grumpier.

One day, the rhinoceros stood alone snorting and grumbling to himself under a ginkgo tree. He noticed a little bird perched cheerfully on a branch above him.

"Hello down there," chirped the bird.

"Go away and leave me alone!" huffed the rhinoceros.

"I would like to ask you a question," the bird insisted.

The rhinoceros was so surprised that the bird was still speaking to him, he forgot to be grumpy. "Aren't you afraid of me?" asked the rhinoceros.

"Not at all," answered the bird. He was safely out of reach of the rhinoceros. Besides, the bird was too curious to be afraid.

"I was just wondering," continued the bird, "what in the world makes you so grumpy? You must know that's why you have no friends."

The rhinoceros stamped impatiently. "I know. I can't help it," he snapped. "You would feel bothered too, if your back were always as itchy as mine."

The bird looked down at the rhinoceros's back. "I see your problem!" tweeted the bird, as he hopped excitedly on his branch. "You have little bugs crawling all over your back."

"Well, now I know the reason," the rhinoceros shook his head. "But it still doesn't solve anything."

The little bird fluttered down to a lower branch to look more closely. "I think I can help you," said the bird.

"You?" The rhinoceros laughed. "How?"

The bird replied, "You are itchy, and I am hungry. Those bugs look delicious. If you will let me ride along on your back, I will get rid of those unwelcome visitors for you."

The rhinoceros thought about this. "Won't I look foolish walking around with a little bird on my back?" he worried.

"Some might say you look pretty foolish now," reasoned the bird, "standing here grumbling to yourself under a tree."

The rhinoceros could not argue with the bird's point. He agreed, and the little bird hopped onto his back.

The next day, the other animals saw an amazing sight. The rhinoceros was trotting across the plain with a tiny bird perched on his wrinkled shoulder. The rhinoceros felt so much better without the bugs on his back. He felt so good that he did not mind the strange looks he got from the others. His itch and his loneliness were both gone.

Even today, in Africa, you can still see little birds riding on the back of a rhinoceros.

Text-Dependent Analysis (TDA) Question

The bird is important to the story. Write an essay explaining how the rhinoceros changes because of the actions of the bird. Use evidence from the story to support your response.

Student Answer

SCORE = 4

The rhinoceros changes though out the beginning of the story to the end. In the beginning of the story the rhinoceros was the grumpiest animal in all of Afica, according to the passage.One reason of how I know that the rhinoceros was grumpy in the beginning of the story is because the auther states that the rhinoceros was always in a horrible mood. And a horrible mood is a grumpy mood.Another reason why I know that he is so unhappy is because the passage said, "He was so unpleasant and mean, none of the other animals would have anything to do with him. Because of this, he was also very lonely. That made him even grumpier." I don't think it would be very nice to be so lonely so the rhinoceros is definetly grumpy. He is also probobly pretty mean too. A reason of how I think he is probobly mean is because before the little bird solved his problem of being so itchy (which was because there was bugs crawling all over his back). according to the passage, and said that he could help him, the rhinoceros just lauged and said, "You? How?" According to the passage. At the end of the story the rhinoceros got a lot more happy. The rhinoceros got a lot more happy because he agreed to let the bird help him take away his itch by eating the bugs off his back. I would be happy if someone helped me not be itchy. "The rhinoceros felt so much better without the bugs on his back." So that is how the bird helped the rhinoceros change from grumpy to happy.

Confirmatory Studies

Several studies conducted by HumRRO for this evaluation support the quality of the SC READY items. They include consistency with the 2014 *Test Standards* for scoring, review of item statistics, and replication of psychometric processing,

***Consistency with Test Standards* for Scoring.** As part of the analysis of the SC READY scaling, equating, and scoring processes (Task 5), HumRRO staff rated the consistency of the item scoring procedures for the TDA items against three scoring Standards from the 2014 *Test Standards* identified as relevant. These ratings are presented in Table 24. All the ratings were good.

Table 24. Ratings of SC READY Consistency with Identified Scoring Standards

Standard	Description	Rating	Comments
6.8	Documentation of rubrics, procedures & criteria for scoring with human judges	4	Two trained scorers used rubrics and anchor papers to rate TDA essays with SCDE required ≥ 70% exact agreement
6.9	Documentation of quality control processes, criteria, training & monitoring	4	Accuracy monitored by back reading, validity checks and consistency checks; retraining and 3rd readers if necessary
6.10	Written interpretations appropriate for the audience to accompany released scores	4	SRUG interpretive information for all SC READY score reports; SS confidence intervals & TM cautions re ordinal subscores

* TM=Technical Manual; SRUG=Score Report User's Guide; SS=scale score
Source: Chapter 5 (Task 5), Score Report User's Guide, Technical Manual.

Review of Item Statistics. Content representation, discussed in the section for Legislative Criterion 2, is the primary factor used to select items for each SC READY test. However, the statistical properties of items are also important indicators of the quality of the items for measuring students' knowledge and skills at the appropriate level of difficulty, for distinguishing between ability levels, and for measuring all students fairly. The statistical properties of the SC READY items were evaluated by HumRRO and the results are summarized in Table 25.

The data in Table 25 are organized by subject, grade and item type. The blue section reports classical item statistics. These include the mean difficulty (percent of correct answers) and mean discrimination (correlation between item answers and total scores) for each item type. Larger difficulties are easier items. Larger discriminations indicate that students with higher test scores are more likely to answer the item correctly.

The purple section of Table 25 presents difficulty and misfit statistics from the Rasch model used by the contractor to analyze items and test forms. Rasch difficulties range from about -3 to +3 with larger, positive values indicating more difficult items and smaller, negative values indicating easier items. The reported misfit is a psychometric procedure specific to the Rasch model that indicates when the item data are significantly inconsistent with the model.

The brown section of Table 25 reports the number of SC READY items found to exhibit differential performance (DIF) for three group comparisons: females as compared to males (gender DIF), African-Americans as compared to Whites (ethnic DIF), and online test takers as compared to paper/pencil test takers (mode DIF). DIF was evaluated using the MantelHaenszel

*Table 25. Summary of Item Statistics for 2017 SC READY Operational Tests**

			Classical Statistics‡		Rasch Model‡		DIF Flags**		
	Item Type*	Number of Items	Mean Difficulty	Mean Discrim	Mean b	Misfit	F/M	AA/W	Mode
ELA Grade 3	MC EB MS TE TDA	59 4 4 3 1	55 29 31 51 30	.38 .49 .42 .28 .66	-0.982 0.400 0.302 -0.818	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0
4	MC	64	62	.40	-0.814	0	0	***1***	0
	EB	1	48	.21	-0.099	0	0	0	0
	MS	3	36	.48	0.527	0	0	0	0
	TDA	1	26	.52	1.372	0	0	0	0
5	MC	63	59	.40	-0.270	0	0	***1***	0
	EB	2	51	.62	0.146	0	0	0	0
	MS	3	37	.51	0.934	0	0	0	0
	TDA	1	29	.53	1.646	0	0	0	0
6	MC	67	58	.41	0.049	***2***	0	***3***	0
	EB	8	42	.50	0.887	0	0	0	0
	MS	5	36	.47	1.204	0	0	0	0
	TDA	1	25	.55	2.199	0	0	0	0
7	MC	69	57	.40	0.471	0	0	0	0
	EB	5	46	.53	1.016	0	0	0	0
	MS	6	36	.41	1.590	0	0	0	0
	TDA	1	32	.66	1.898	0	0	0	0
8	MC	69	61	.41	0.582	***2***	0	***1***	***1***
	EB	7	41	.47	1.715	0	0	***1***	0
	MS	4	36	.53	1.982	0	0	0	0
	TDA	1	44	.69	1.562	0	0	0	0
TOTAL		452				4	0	7	1
MATH Grade 3	MC	50	61	.43	-1.185	***3***	0	***1***	***1***
4	MC	56	55	.41	-0.529	***1***	0	***4***	0
5	MC	54	54	.42	-0.020	***1***	0	***2***	0
	MS	2	23	.47	2.075	0	0	0	0
6	MC	54	58	.41	-0.090	0	0	***1***	***1***
	SR	3	64	.53	-0.441	0	0	0	0
	MS	3	46	.61	0.551	0	0	0	0
	TE	1	35	.45	1.172	0	0	0	0
7	MC	54	51	.39	0.476	***1***	0	***2***	***1***
	SR	3	49	.60	0.541	0	0	0	0
	MS	3	28	.45	1.744	0	0	***1***	0
	TE	1	15	.41	2.736	0	0	0	0
8	MC	55	53	.39	0.759	0	0	***1***	0
	SR	3	23	.54	2.480	0	0	0	0
	MS	4	19	.43	2.826	0	0	0	0
	TE	3	37	.47	1.643	0	0	0	0
TOTAL		349				6	0	11	3

* MC=multiple choice, EB=evidence-based, MS=multi-select, TE=technology enhanced, TDA=text-dependent analysis (essay), SR=short answer/gridded response;

‡ Difficulty (p-value)=proportion of students correctly answering an item; Discrim=item-total correlation (point biserial)=higher test scores associated with correct answers; b=Rasch difficulty (larger number, harder item);

** F=female; M=male; AA=African-American; W=White; Mode=online, paper/pencil;

Source: Chapter 6 (Task 6), Technical Manual, p. 52-55.

statistic with the ETS decision rules that classify the amount of DIF as A=none to minimal, B=moderate, and C=significant. Items classified as *C DIF* indicate that students of *equal ability* in the focal group (e.g., African-Americans) correctly answered the item significantly less often than students in the reference group (e.g., Whites). Items classified as exhibiting *C DIF* should not be used unless needed to meet the test blueprint. The number of SC READY items identified as exhibiting *C DIF* is listed in the brown section of Table 25.[62]

ELA Item Statistics. Several trends are apparent in Table 25. For the 452 ELA objective items reviewed, the multiple choice items were easiest with an average correct answer rate (difficulty or p-value) of 55-62%. Except for Grade 3, the multi-select items were the most difficult with average correct answer rates of 36-37%. In Grade 3, the multi-select and evidence-based items were similar in difficulty with average correct answer rates of 31% and 29%, respectively. All item types across Grades 3-8 had good average item discriminations. The TDA item tended to be hard and to differentiate high and low performing students relatively better on average than the other item types. Rasch difficulties (b values) demonstrated similar trends. No ELA items were flagged for p-values outside the acceptable range of 10-95 and only 1-4 multiple-choice items at each grade level were identified as having distractors that were chosen more often than the correct answer. Only one Grade 6 item was flagged for a discrimination value that was too low (<.10) and only 1-2 multiple-choice items per grade were flagged for having a distractor that correlated more highly with the total score than the correct answer. Such items may be ambiguous or unnecessarily tricky.

Misfitting items demonstrate unusual student response patterns given the item difficulty. There were only four misfitting items identified for ELA, two in Grade 6 and two in Grade 8. Differential item functioning (DIF) statistics identify items that perform differently for students of equal ability from two separate groups of interest (e.g., females and males). Only the most extreme *C DIF* items were flagged. No ELA items were flagged for gender DIF and only one was flagged for mode DIF between online and paper/pencil. Seven items had flags for ethnic DIF between African-Americans and Whites and were spread across multiple grades.

DIF statistics do not by themselves indicate unfairness but instead are an indicator that the item should receive additional scrutiny. Sometimes fairness/sensitivity review panels can identify the probable source of DIF and the item can be revised and re-field tested. If not, and if the content reviewers believe that the item appropriately measures an important skill from the content standards, the item may be retained and used if needed to satisfy the test blueprint. The very small number of ELA operational items identified for DIF indicates that the fairness procedures employed by the contractor were successful.

Mathematics Item Statistics. The statistics for the 349 SC READY Mathematics items reviewed were similar. In the lower grades, nearly all the items were multiple-choice and had average difficulties (p-values) of 54-61%. In the upper grades, multiple choice items tended to be relatively easier on average but there was no consistent pattern among the other objective item types. These results may have been a function of the particular content measured by the very few items of each of these other item types. Multi-select and technology enhanced items were particularly difficult in Grades 7-8.

No Mathematics items were flagged for p-values above 95% (very easy) and two multi-select items were flagged for p-values less than 10% (very hard). Five items from three grades were

[62] *See* Chapter 6 (Task 6) and the Technical Manual, *supra* note 13, p. 52-55, for more detail about these comparisons.

identified as having misfitting student response patterns. Compared to ELA, a greater number (23) of Mathematics items had flags for a distractor being chosen more often than the correct answer, perhaps because these attractive distractors represented common mathematical mistakes. Average discrimination statistics were generally good with two items flagged for values less than 0.10. Nine items had distractors that correlated more highly with total scores than the correct answer.

There were six misfitting items, all multiple choice and mostly in the lower grades. Again, there were no items flagged for gender DIF and only two for mode DIF. Eleven items were flagged for ethnic DIF, with a total of four in Grade 4. The small number of items flagged for additional scrutiny due to DIF is an indication that there were no systematic fairness issues. See Chapter 6 (Task 6) for additional details regarding the review of SC READY item statistics.

Timing

All of the SC READY assessments are untimed. The ELA assessments are divided into two sections. Part I measures writing and inquiry (research) skills and includes the TDA essay item. Part II consists of multiple literary and informational passages with sets of associated questions. For Grades 6-8, the Mathematics assessments include a section for which use of a calculator is permitted and a section where it is not. Each ELA and Mathematics test session also includes a small set of non-scored field test items spiraled within classrooms or randomly assigned to online test sessions to collect item evaluation data. Items that survive field testing by exhibiting acceptable statistics are used as replacement items on the next year's SC READY test forms.

The estimated times for scheduling test sessions provided in the *Test Administration Manual* (TAM) and the median testing times for students reported in the *Technical Manual* (TM) are presented in Table 26.[63] For ELA Session 1, the reported median times are about 12% to 25% less than estimated, but for ELA Session 2 they are about the same. For Mathematics, the median times are about 25% to 50% less than estimated for the lower grades and about 15% to 33% less than estimated for the upper grades.

The differences between the actual and estimated times may reflect the additional time needed for directions at the beginning and collection of materials at the end of the test. In addition, the 75th percentile student times are greater than the estimated times for ELA Session 1 in the lower grades and ELA Session 2 in all grades. For Mathematics, the 75th percentile student times are close to the estimated times for all grades. In sum, the time estimates are probably reasonable for most students but there are likely some students who may finish early or need significant additional time.

Item and Test Form Quality

An important consideration when using a variety of item types for an assessment is the consistency of item quality across item types. Evidence related to content validity, item construction, fairness/sensitivity, item alignment, and field testing of replacement items contributes to a collective judgment of the overall quality of the SC READY test items.

[63] Test Administration Manual, *supra* note 32, p.26; Technical Manual, *supra* note 13, p. 28.

Table 26. SC READY Estimated and Actual Administration Times by Test Section

	TAM Estimate	Median Actual Time
ELA Session 1	2 hours	Grades 3-5 = 1¾ hours Grades 6-8 = 1½ hours
ELA Session 2	1½ hours	Grades 3-5 = 1½ hours Grades 6-8 = 1½ hours
Mathematics	2 hours	Grades 3-5 = 1½ hours
Calculator		Grades 6-8 = 1¼ hours
No calculator		Grades 6-8 = ½ hour
		1¾ hours

Source: Test Administration Manual (TAM), p. 26; Technical Manual, p. 28.

Content Validity. One of the most important psychometric characteristics of a test is the validity of the intended test score interpretations. Specifically, for standards-based assessments, evidence of content validity is most relevant. Content validity refers to the collective congruence between the substance of the items that constitute a test form and the corresponding content standards intended to be assessed. That is, do the items on a test form measure what they are supposed to measure?

Standard 1.11 from the 2014 *Test Standards* states:

> When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and … the [content] domain it is intended to represent. …
>
> *Comment:* The match of test content to the targeted [content] in terms of cognitive complexity and the accessibility of the test content to all members of the intended population are also important considerations.

Content validity evidence for a standards-based test is typically obtained via the judgments of educators with experience teaching the subject matter at the grade level(s) of interest. Standard 1.9 from the 2014 *Test Standards* provides the following guidance for content validity evidence based on educator judgments:

> When a validation rests in part on the opinions or decisions of expert judges, … procedures for selecting such experts and for eliciting judgments … should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. …
>
> *Comment:* Systematic collection of judgments or opinions may occur at many points in test construction (e.g., eliciting expert judgments of content appropriateness or adequate content representation) ….

Several aspects of test development utilize educator and expert judgments that contribute to the evidence for content validity and item quality. These aspects include item development, content reviews, fairness/sensitivity reviews, alignment studies, forms construction, quality control procedures and field testing. The following paragraphs describe the test development activities for the SC READY assessments that support content validity and item quality.

Item Development. The items in the contractor's CCR item bank were written by professional item writers and edited by contractor staff. These items were matched to the CCR Common Core content standards and included a variety of item types. The items underwent tryouts with small groups of students and were reviewed by a committee of content experts with subject matter knowledge and experience with students at the targeted grade level. Efforts were made to follow the principles of *universal design* which are intended to render items accessible to the widest possible range of students, including students with disabilities and English learners.[64]

Separate content review committees were constituted for ELA and Mathematics. There were 12 ELA and 10 Mathematics content reviewers. The composition and experience of the educators selected to serve on these content review committees is critical to establishing item quality. Specific demographics for these educator panels were not reported but they were described as experienced educators from a variety of fields, levels and special populations.[65]

Prior to evaluating items, content reviewers received training to familiarize them with the relevant content standards, principles of *universal design*, and common item flaws. Content reviewers evaluated each item individually followed by a group discussion. The goal of the discussion was to obtain a consensus on whether each item should be retained, revised or rejected for inclusion in the item bank.

Item Reviews for Fairness and Sensitivity. The items in the bank were also reviewed for fairness and sensitivity by a ten-member committee of educators familiar with such concerns and representative of relevant demographic groups such as gender, ethnicity, and special populations. Fairness reviews seek to identify and revise any content that might disadvantage a subgroup of students based on vocabulary, reading level, unfamiliar content, or other irrelevant factors. For example, urban students may not be familiar with farming techniques and students from southern states may lack experience with snow and ice.

Sensitivity reviews seek to identify any content that may evoke an unintended emotional reaction or distraction for certain subgroups of students. In particular, fairness reviewers scrutinize items for stereotyping, gender imbalance, regionalism, ethnic/cultural issues, socioeconomic/class issues, religious content, age discrimination, appropriate presentation of persons with disabilities, accessibility, and the potential for computer and pencil/paper accommodations. In terms of sensitivity, particularly in reading passages, topics such as controversial matters (e.g., abortion, gun control, immigration), inappropriate behaviors (e.g., stealing, cheating, murder), family problems (divorce, job loss, death) and politics are avoided.

The contractor has written guidelines for fairness/sensitivity reviews and accessible assessments that provide interesting context and examples for the work of this committee. A checklist for content reviewers and fairness reviewers has also been developed.[66]

[64] *See* Technical Manual, *supra* note 13, p. 12-13.
[65] *See id.*, p. 13-14.
[66] DRC (2016). Fairness in Testing: Guidelines for Training on Bias, Fairness, and Sensitivity Issues, Maple Grove, MN: Author; DRC (2015). Accessible Assessments: Making Assessments Accessible and Inclusive, Maple Grove, MN: Author; DRC (no date). *Item Writer Manual Supplement: Content and Fairness Checklists,* Maple Grove, MN: Author.

Following the content and fairness reviews, items were field tested on samples of students at the targeted grade level. From the field test data, item statistics were calculated to quantify the difficulty of the item (p-value=percent answering correctly; Rasch difficulty), the degree to which knowledgeable students tended to get the item right more often than non-knowledgeable students (referred to as item discrimination), and the frequency with which the alternative answer choices were selected (*see* the review of item statistics earlier in this section).

In addition, when sample sizes permitted, a statistic was calculated to quantify the degree to which students in a focal group (e.g., Female, African-American) of equal ability to students in a reference group (e.g., Male, White) correctly answered an item less frequently. This statistic is referred to as differential performance (DIF) and is used to identify items that potentially could disadvantage members of the focal group. DIF statistics are commonly classified into three groups: A=none to minimal; B=moderate; C=significant. Typically, the majority of items on a test are classified as category A and only a small fraction as category C. *C DIF* items are generally considered outliers and are to be revised or avoided if possible. *B DIF* items may be reviewed to determine if any characteristic of the item can be identified as causative and are used when no category A items are available to satisfy a particular cell in the test blueprint. If a causative characteristic for *C DIF* and *B DIF* items can be identified by review committees, these items are generally revised and re-field tested. *See* Table 25 above and its associated text for information for additional data about DIF statistics for the SC READY assessments.

Alignment. Evidence of the alignment between the content standards, test blueprint and test items was discussed earlier in the section on Legislative Criterion 2. The same HumRRO alignment studies also provided evidence of item quality. The educator panels provided additional ratings of the following item quality indicators:

1. **Alignment between the depth of knowledge (DOK) of the test items and the DOK of the content standards –** qualitative judgments of whether the complexity of cognitive processing required by a test item, across the four levels of recall, skills/concepts, strategic thinking, and extended thinking, matches that specified by its corresponding content standard;

2. **Evaluation of item quality –** ratings of item clarity, accuracy, grade-level appropriateness, support of research-based instruction, and fair/non-offensive content. The latter provides a check on the work of the Fairness/Sensitivity Review Committee.

3. **Overall holistic rating of the alignment of the items and the content standards**.

This section highlights the results from these item alignment ratings relevant to Legislative Criterion 6. Summaries of the item quality ratings from the alignment reviews for the SC READY Grades 3-8 ELA and Mathematics tests are presented in Table 27 (ELA) and Table 28 (Mathematics).

Table 27. Item Quality Results for SC READY Grades 3-8 ELA

ELA		POSITIVE RATINGS OF ITEM QUALITY*					
	DOK Items At or Above Standards	Clarity	Accuracy	Grade Level Appropriate	Supports Research-based Instruction	Fairness	Holistic Rating (Number of Panelists)
Grade 3	45%	99%	99%	99%	100%	100%	(5) Good
4	54%	99%	99%	100%	100%	100%	(5) Good
5	26%	100%	100%	100%	100%	100%	(5) Good
6	52%	98%	99%	99+%	99+%	100%	(4) Good (2) Needs Improvement
7	32%	98%	99%	99%	100%	99%	(4) Good (2) Needs Improvement
8	31%	99+%	99+%	100%	100%	100%	(6) Good

Source: Chapter 2 (Task 2).

For the ELA tests, the depth of knowledge (DOK) levels of the items were uneven with respect to the standards they assessed. For example, if an item was rated DOK 1 Recall but the standard it assessed was rated DOK 2 Skills/Concepts, there was a mismatch. The percent of items with DOK levels at or above their corresponding standards varied from a high of 54% for Grade 4 to a low of 26% for Grade 5. Webb, the originator of this indicator, recommends that the DOK matching percent should be at least 50%. The test items for only two grades (4 and 6) barely met this recommended value. The Technical Manual states that among other skills, item writers were trained on Webb's four levels of cognitive complexity used for the DOK ratings.[67] However, DOK values were not part of the SC READY test blueprint so were not considered directly when ELA items were matched to the test blueprint by contractor staff and content reviewers.

The inquiry items for all grades had particularly low levels of DOK match. The percent of inquiry items with DOK levels below the DOK levels for their corresponding standards were 74%, 100%, 100%, 56%, 89%, and 76% for Grades 3-8, respectively. See the sample multi-select item earlier in this section for an example of an inquiry item for Grade 3 ELA.

All of the Grade 5 items also had particularly low levels of DOK match. The percent of DOK matching items for the subscores of reading literary text, reading informational text, writing and inquiry were 24%, 6%, 47% and 0%, respectively.

Positive ratings of item quality for ELA were near perfect for all grades and criteria. These data strongly support the quality control procedures employed by the contractor during item writing

[67] Technical Manual, *supra* note 13, p. 12.

and item review. In particular, the fairness/sensitivity review panels appear to have identified all content that might disadvantage or be offensive to minority subgroups.

Holistically, more than 80% of panelists at every grade level rated the quality of the ELA items as *good*.

Table 28. Item Quality Results for SC READY Grades 3-8 Mathematics

MATH		POSITIVE RATINGS OF ITEM QUALITY*					
	DOK Items At or Above Standards	Clarity	Accuracy	Grade Level Appropriate	Supports Research -based Instruction	Fairness	Holistic Rating (Number of Panelists)
Grade 3	56%	97%	99%	98%	99%	100%	(5) Good
4	72%	99%	100%	100%	100%	99+%	(5) Good
5	75%	99+%	100%	100%	100%	100%	(5) Good
6	73%	97%	99%	98%	95%	99%	(5) Good (1) Fair
7	75%	95%	99%	99%	99%	100%	(5) Good (1) Fair
8	74%	99%	99%	99%	98%	100%	(5) Good (1) Fair

* *Source:* Chapter 2 (Task 2).

For the Mathematics tests, the depth of knowledge (DOK) levels of items with respect to the standards they assessed were much higher than for ELA across grade levels. The percent of items with DOK levels at or above their corresponding standards varied from a high of 75% in Grades 5 and 7 to a low of 56% in Grade 3. Webb's 50% recommendation was exceeded for all grade levels in Mathematics. Similar to ELA, the Technical Manual states that among other skills, mathematics item writers were trained on Webb's four levels of cognitive complexity used for the DOK ratings.[68] However, DOK values were not part of the SC READY Mathematics test blueprints so were not considered directly when Mathematics items were matched to the test blueprint by contractor staff and content reviewers.

Positive ratings of item quality for the Mathematics tests were near perfect for all grades and criteria. These data strongly support the quality control procedures employed by the contractor during item writing and item review. In particular, the fairness/sensitivity review panels appeared to have identified all content that might disadvantage or be offensive to minority subgroups.

Holistically, more than 80% of panelists at every grade level rated the quality of the Mathematics items as *good*.

[68] *Id.*, p. 12.

Forms Construction. The contractor and SCDE collaborated to select items for the SC READY operational forms from the contractor's CCR item bank that met the following criteria:

- Item types and content representation matched the test blueprint, and
- Items were fully aligned to the South Carolina CCR Content Standards.[69]

HumRRO reviewed the documentation for the SC READY forms construction process and conducted a site visit to observe the process in action. The forms construction process was evaluated by comparing it to the following eight 2014 *Test Standards* identified as most directly relevant to this task.

- **Standard 4.1 –** the test blueprint should describe the purpose, content domain, student population, and interpretations for intended uses of the test;
- **Standard 4.2 –** the test blueprint should also describe the test content; length; item formats; psychometric item/form properties; item ordering; administration timing, directions, security procedures and accommodations; required materials, scoring; reporting; and hardware/software requirements for computer-based tests;
- **Standard 4.4 –** document content, psychometric specifications, validity, reliability, comparability of different versions of the test (e.g., computer, paper/pencil);
- **Standard 4.5 –** identify, describe and provide a rationale for any test administration variations, the qualifying students and any requirements for use;
- **Standard 4.7 –** document item tryouts, reviews and selection criteria;
- **Standard 4.9 –** describe the selection procedures and characteristics of item tryout samples that should be as representative of the test taking population as possible;
- **Standard 4.10 –** document the model (e.g., classical, Rasch), sample of adequate size and diversity, screening data and criteria (e.g., difficulty, discrimination, differential functioning for major student groups) and model fit, if appropriate, for analyzing the psychometric properties of test items;
- **Standard 4.13 –** investigate and eliminate potential sources of irrelevant test score variance when indicated by credible evidence and to the extent feasible.

Two HumRRO staff rated the documentation for fidelity to each identified Standard on a five-point coverage scale of 1=no evidence, 2=little evidence, 3=some evidence, 4=substantial evidence including key aspects, and 5=full coverage, and the observational site visits on a five-point adherence to documented procedures scale of 1=not followed, 2=rarely followed, 3=inconsistently followed, 4=mostly followed, and 5=always followed. The staff members then met to discuss their ratings, arrive at a consensus rating, and consolidate comments on incomplete or missing coverage. Additional details about the methods for these evaluations are provided in Chapter 3 (Task 3).

The results of the HumRRO forms construction evaluations are summarized in Table 29. One set of ratings is provided for both ELA and Mathematics because both utilized the same procedures and those common procedures are described in a single set of documents.

The results of the forms construction evaluation indicates that all relevant Standards were substantially or fully met based on the available documentation. Suggestions for improvement are included in Table 29. The average rating for documentation consistency with the Standards for forms construction was 4.6.

[69] *Id.*, p. 15.

Table 29. SC READY Forms Construction Evaluation: Ratings and Comments

Test Standard	Documen-tation Rating	Comments	Documented Procedure	Site Visit Rating	Comments
4.1	5	The *TAM* cites SC statutory accountability as the primary use of test scores; intended population of students is inferred from grade level test blueprints aligned to content standards	1. Content specialist selects items	3	Judgmental identification of ≈25% replacement items does not explicitly consider prior exposure; reused items positioned similarly to previous year; Step 8 is combined with Step 1
4.2	5	Document *016_Guidelines for Item Analysis and Form Construction_R.pdf* and the *TAM* provide a detailed description of forms assembly, including factors considered	2. Psychometric review by senior psychometrician	5	Content specialists shared their item selections via an Excel spreadsheet
4.4	4	Online forms constructed first, then paper/pencil with necessary item substitutions of TE items with MC items possessing similar statistics; item DIF rare in mode comparability study; separate scales for some grades	3. Compare proposed form with desired psychometric characteristics	3	Comparisons included item difficulty, discrimination, correct answer distribution, DIF, reuse, and sequencing with no written rules to be followed
4.5	5	Some variations available to all students online; accommodations allowed for students with IEP or 504 plans; districts with waivers can test paper/pencil	4. Feedback sent back to content specialist	5	Psychometrician suggested deviations from ideal that the content specialist should try to correct with acceptable item replacements
4.7	5	Item exposure not tracked; annually ≈25% of items replaced by undocumented judgmental criteria	5. Adjustments by content specialist based on feedback	4	Content validity was paramount; acceptable item replacements were usually available
4.9	4	Embedded field testing of new items representative of SC but replacement items from other states may not be; Grade 3 reading pre-equated	6. Revised form sent back for psychometric review	4	5 ELA forms were approved on the first submission and the 6th with one revision; math adjustments usually involved one or two items
4.10	4	Forms construction based on classical model; Rasch model calibrations used to equate test forms; model fit is not addressed	7. Repeat steps 3-6 until agreement is reached	5	The goal was the best possible content representation using items with the most favorable statistics
4.13	5	Fairness reviews, DIF statistics, mode comparability statistics, universal design principles and content rechecks are used to remove potential sources of irrelevant variance	8. List of operational items sent to SCDE for review and feedback	5	Step 8 actually occurs during Step 1 and considers factors such as alignment to standards, the balance of passage and item types, content similarity and clueing, gender roles, DIFF statistics and sequencing*

* Items also reviewed for appropriateness by a visually/hearing impaired specialist; *Source:* Appendix G & Technical Manual.

When test forms are constructed, both content and psychometric requirements must be satisfied. The documented steps shown in Table 29 indicate that this is a recursive process between the content specialist and the psychometrician who must both agree that all important requirements have been met to the extent feasible given the constraints of the available items in the item bank. The ratings of staff that observed this process for SC READY indicate that although the initial selections and simultaneous adherence to content and psychometric requirements could be improved with greater automation, the processes currently in place were able to fully meet the requirements with repeated rounds of adjustments consistent with the documented procedures for forms construction. The overall average rating for adherence to documented procedures as observed during the site visit was 4.3. Average ratings for ELA and Mathematics were 4.5 and 4.11, respectively.

Quality Control. Alignment data provide one type of evidence supporting the quality of the SC READY items. Other data collected by HumRRO in other studies also support SC READY item quality. Items from the contractor's CCR bank that had survived the processes described above were selected to match the SC READY test blueprints and to align with the South Carolina content standards to be assessed. Several quality checks were performed on the selected items to verify content alignment, rigor, distractor plausibility, fairness, accessibility, answer keys, and stimuli.[70]

The quality of the final sets of items that comprised the SC READY test forms is supported by the information presented above and the data presented in Chapters 1-3 (Tasks 1-3) and Chapters 5-6 (Tasks 5-6) of this report.

Item Bank Development. The contractor developed field test items specifically for SC READY using the same procedures described earlier for the items selected from the contractor's item bank. These new items were field tested on the Spring 2017 SC READY test forms and those that survived statistical review will be available for use on future SC READY test forms. A total of 40-47 field test items per grade were written for ELA and 27-37 per grade for Mathematics.

For the 2018 SC READY test form, up to about 60% of the ELA and half of the Mathematics items could theoretically be replaced if all the field tested items survived. Survival rates are often about 50% for field tested items so SCDE's plan to replace about 25% of the items on the 2017 forms to create new forms for 2018 is likely achievable.

The quality of the SC READY item bank improved from 2016 to 2017 as shown in Table 30. For 2016, the average difficulty (p-value) for items in the bank was .55 (55%) for ELA and .40 (40%). In 2017, those values had risen to 59% and 54%, respectively. The ideal range for targeting achievement test items is 60-80% to minimize random guessing. Random guessing increases substantially when the items are too hard for most students and contributes additional error variance to student scores. The small, 4-percentage point decrease in ELA average item difficulty placed most grades in or near this range. For Mathematics, the change was more dramatic at an overall average decrease of 14 percentage points, but the bank items still remained relatively difficult on average for the student population.

The differences between ELA and Mathematics average item bank difficulties were reflected in the 2017 test forms. For ELA, the average test form difficulty was near the bank value for most grades, but for Mathematics, the average test form difficulty was at or slightly above that for the

[70] DRC (no date). *Item Writer Manual Supplement: Content and Fairness Checklists,* Maple Grove, MN: Author; DRC (Dec. 2016). *Power Point Presentation for Content Review Meeting,* Columbia, SC.

item bank, suggesting that the test developers tried to make the test forms easier but were limited by the substantial number of hard items in the bank. This may reflect the fact that CCR mathematics content is hard for most students, especially if they have not received CCR-targeted instruction in the past. However, as teachers focus on the state content standards and tested students obtain more years of instruction on the CCR standards, the test items may become somewhat easier.

Median point biserial item statistics (correlations between item scores and total test scores) were generally good in 2016 and in 2017, generally held steady or improved slightly for ELA and improved moderately for Mathematics in most grades. The major exception was Grade 3 Mathematics where the median was very low in 2016 and improved substantially in 2017. Substantial numbers of 2016 items must have been replaced in 2017 to achieve such a large improvement.

Table 30. Comparison of 2016 and 2017 SC READY Item Banks and Operational Tests

	2016		2017	
	Mean p-value Bank	*Median Pt Biserial* Bank	*Mean p-value* Bank (Test)	*Median Pt Biserial* Bank (Test)
ELA Grade 3	.50	.37	.54 (.52)	.37 (.38)
4	.57	.38	.62 (.60)	.39 (.42)
5	.61	.41	.62 (.58)	.39 (.42)
6	.50	.41	.59 (.54)	.43 (.42)
7	.55	.42	.58 (.54)	.42 (.42)
8	.58	.41	.61 (.58)	.43 (.42)
MATH Grade 3	.36	***.14***	.59 (.61)	.41 (.44)
4	.44	.33	.56 (.55)	.40 (.42)
5	.44	.38	.53 (.53)	.43 (.44)
6	.45	.38	.58 (.57)	.42 (.43)
7	.38	.34	.49 (.49)	.37 (.42)
8	.35	.30	.49 (.49)	.41 (.42)

Source: HumRRO Report #1, p. 43-44; Technical Manual, p. 32.

Item Maps. SC READY Rasch item maps for Grade 3 ELA and Grade 8 Mathematics are reproduced in Exhibit H. Item maps place the difficulty of the items and the abilities of the students on the same graph so they can be compared. In both cases, the item distributions are centered on the student distributions, but not surprisingly, the student abilities are more variable, especially in the higher grade. To provide additional validity evidence for the standards-based score interpretations for the SC READY assessment system, the cut scores could be superimposed on the item map and the items within each performance level identified. The content of those items could then be compared to the PLDs to further strengthen and refine the descriptions of the skills students are mastering at each performance level.

Evaluation

The SC READY assessments are composed of a variety of item types that measure student understanding of the content in different ways. For some items, students select a correct answer and for others, the student must produce the answer. Some items require distinguishing multiple correct and incorrect answers and some require identification of evidence that best supports an answer. For students testing online, a few items utilize some of the unique features of the technology. There is also an extended essay item that requires students to combine text analysis, writing skill and use of evidence to support an answer.

Item Quality. The 2017 SC READY Grade 3 ELA and Grade 8 Mathematics operational forms were reviewed in both online and paper/pencil formats. The items were found to be clearly stated, free from common item flaws, well-written to elicit application and reasoning skills, and populated with plausible distractors.

SC READY items are leased from the contractor. They were written by professional item writers to align with the CCR Common Core content standards, reviewed by educational experts for content and fairness/sensitivity, and field tested to check item statistics. Item statistics for the 2017 SC READY tests were generally very good with only a small fraction of items exhibiting less desirable properties. Nontraditional item types had more flags for difficulty at the middle school grade levels suggesting that further review of these items might be warranted.

Nearly all ELA and Mathematics items were rated by HumRRO alignment panelists as clear, accurate, grade-level appropriate, supportive of research-based instruction and fair for all students. These ratings strongly support the effectiveness of the content and fairness review panels. Though not specifically stated, one might reasonably assume that the 2017 TDA item anchor papers will continue to be used in future years for training raters to score new TDA items with accuracy (validity) and consistency (reliability) and to avoid scale drift.

Choice of Item Statistics. From a psychometric perspective, biserial and point biserial statistics can be used to assess the extent to which correct answers to items distinguish between students of high and low ability (item discrimination). Point biserial calculations use dichotomous (0=incorrect, 1=correct) item responses while the biserial statistic assumes an underlying, continuous distribution of the ability to correctly answer an item. However, from a legal perspective, the value of this item statistic is to choose comparatively better items and to avoid possible miskeyed or ambiguous items. Both of these goals can be achieved with either statistic. Moreover, when deliberating between two possible items with acceptable item discrimination statistics, content validity (alignment to the test blueprint and content standards) is the most important consideration for the decision.

Blueprint Weighting. ELA panelists overall rated the blueprint as adequately representing the content standards but expressed reservations about the assessment of inquiry skills, believing that local performance-based testing would be more effective. The inquiry items also tended to exhibit the most unique variance, suggesting they are measuring skills that are somewhat different than those measured by the rest of the ELA assessment. For Mathematics, the panelists rated the blueprint in Grade 4 as adequately covering the content standards but thought the alignment for the other grades could be improved by adjusting the reporting category weights. As suggested in Chapter 2, SCDE may want to convene an experienced panel of South Carolina educators to reconsider the blueprint reporting category weights for Mathematics. On the other hand, the content emphases must be balanced against the need for sufficient numbers of items to provide satisfactory reliability for reporting category scores. SCDE could reasonably conclude that this balancing of goals is best achieved by retaining the current weights. If convened, the experienced educator panel could also address the desirability of retaining the inquiry items in the ELA assessments.

Test Form Construction. The evidence for the content validity, alignment, differential functioning, reliability and quality control all supports the appropriateness and quality of the SC READY items and test forms. The HumRRO alignment studies – for both ELA and Mathematics – verified a close link between the test items and the corresponding content standards they assessed and between the actual number of items per reporting category and the ranges specified in the test blueprints. The SC READY assessments are untimed and the time allotments suggested in the Test Administration Manual appear to be adequate for most students.

HumRRO observers noted that the forms construction meetings were very well organized, consistent with best practices within the industry, and in most respects, faithfully followed the documented procedures. Nonetheless, the observers provided several recommendations for improving and streamlining forms construction that are worth considering (*see* Chapter 3 (Task 3)). In particular, explanations for item rejections should be documented and the frequency of item usage across years should be tracked so items can be targeted for retirement based on exposure rather than chance when approximately 25% of the items are replaced each year.

No indicators of text complexity, such as readability indices or passage/form word lengths, are reported for the SC READY assessments to assist in judging the progression of ELA text complexity and Mathematics reading load across grades. DIF statistics are within normal limits for a standards-based achievement test but ethnic DIF is reported only for African-Americans. There appear to be enough Hispanic students to also calculate DIF statistics for that group.

Review of Ethnic B DIF and C DIF Items. The vast majority of SC READY items exhibited no ethnic DIF. To determine if any patterns were evident for those that did, the SC READY ELA and Mathematics tests with the greatest number of items identified as exhibiting *B DIF* (moderate) or *C DIF* (substantial) for African-Americans, Grade 8 ELA and Grade 7 Mathematics, were examined. There was no clear pattern of item difficulty identified. P-values for items with *C DIF* ranged from .42 to .66 for Grade 8 ELA and .25 to .68 for Grade 7 Mathematics. The p-value ranges for *B DIF* items were .11 to .80 and .26 to .78, respectively.

However, there were some content similarities. For Grade 8 ELA, items exhibiting *B DIF* or *C DIF* tested less-common writing conventions such as ellipses and dashes, vocabulary, interpreting text/drawing conclusions, and opinion/point of view. For Grade 7 Mathematics, items exhibiting *C DIF* tested fractions, decimals, areas (circle, cross-section, surface), and ratios. The HumRRO alignment studies indicated that these topics matched the content standards, but they may have received insufficient instruction, emphasis or practice.

Psychometric best practice is to ask the fairness/sensitivity committee to re-evaluate items exhibiting DIF to determine if the committee members can identify anything about the items likely to have caused the DIF. If yes, the items can be revised and re-field tested. If not, these items may be examples of false positives, that is, they may have been identified purely by chance when in fact there was no actual DIF. For example, for Grade 7 Mathematics, one of the *B DIF* items involved solving arithmetic expressions without a calculator. The item had few words so reading load was unlikely to have been a problem, and the numbers were single digit. For Grade 8 ELA, one of the *C DIF* items involved pronoun/antecedent agreement, a skill that is not likely to be difficult due to text complexity or vocabulary, and is likely to have been taught and practiced.

Whenever a statistical procedure is used, significant results may be obtained by chance. That is why it is important to ask educator committees to re-examine items identified as potentially disadvantaging a focal group to determine if there are any plausible arguments for irrelevant factors to have caused the observed differential performance. If not, and if the item content is verified to align with grade-level content standards, the item can be retained in the item bank for future if needed to satisfy test blueprint requirements. The likelihood that a small percentage of items will be identified for DIF purely by chance is one reason that only the outlier *C DIF* items are typically avoided or revised.

Committee Demographics. More complete documentation of the demographic characteristics of educators serving on content and fairness/sensitivity review committees and qualifying as scorers for the TDA essay items is necessary for evaluating the quality of these activities and following the *Test Standards*. Information similar to that provided for standard setting committee members would be useful.

Field Testing Replacement Items. As discussed in Chapter 3, some replacement SC READY items from the contractor's item bank were field tested in other states where performance may not be representative of South Carolina students in terms of ability or exposure to South Carolina content standards. In addition, initial item tryouts in 2014 used a volunteer, convenience sample that may have been affected by lack of motivation so their item statistics may be less accurate or stable as a result. In the future, it would be preferable to use imbedded field testing to obtain South Carolina item statistics before using these items operationally. This is especially important for Grade 3 where preliminary ELA Reading scores are reported based on pre-equating data. Post equating checks are performed before final score reports are issued but Read to Succeed promotion/remediation decisions may already have been made by then.

Mode Comparability Equating. It is important to conduct mode comparability equating as long as significant numbers of students continue to test paper/pencil. Even though very little item mode DIF has been observed, there could still be forms mode DIF due to scrolling, page turning, reference to diagrams or formulas in separate pop-up windows, use of the online calculator, or the 3-point raw score difference for Grade 6 ELA. A useful methodology for doing so annually is to create matched groups by selecting demographically representative samples from the larger group that match the smaller group to create reference and focal groups of approximately equal size and ability. Judging by the progress to online testing between 2016 and 2017, it may be possible to discontinue mode equating within another few years.

DOK Levels. Although the Mathematics items demonstrated adequate DOK match to the DOK levels for their tested content standards, the results for ELA were uneven and generally below recommended values. As advised in Chapter 2, SCDE might want to consider including DOK levels in the test blueprints to improve the consistency between the DOK levels of ELA items and their corresponding content standards.

Documentation and Verification. In previous chapters, HumRRO evaluators have recommended that the program documentation needs to be expanded to support increased quality control verification by contractor staff and the SCDE, and possibly a third party independent replication of the equating, scaling, and production of scoring tables. Also, scattered program documents or summaries of that information need to be consolidated and incorporated into a single Technical Manual with relevant appendices and references.

7. Test Administration in Paper-Based and Computer-Based Formats

Evidence

Evidence relevant to Legislative Criterion 7 includes mode administration data, the district waiver policy, test forms, a mode comparability study, separate scale score tables, test accommodations policies, and test security policies.

2016 and 2017 Mode Administration Data

In 2017, all districts and schools were required to administer SC READY assessments online unless they had received a waiver from SCDE or were administering an individual test with an accommodation that could not be provided online.[71] The percent of students testing online and with paper/pencil for 2016 and 2017 SC READY by grade level are presented in Table 31. Mathematics counts were chosen because some English learners are not included in the ELA counts.[72] Comparison data for 2016 and 2017, Grades 3 and 8 are shown in Chart 10.

As indicated in Table 31, overall in 2016 about 35% of students tested online and 65% tested on paper. The percent of students testing online increased from about ¼ in Grade 3 to approaching ½ in Grade 8. In 2017, the percent of students testing online improved substantially, ranging from nearly 60% in Grade 3 to almost 85% in Grade 8 and posting gains of 31 percentage points in Grade 3 and 40 percentage points in Grade 8. Although the legislative goal of all students testing online by 2017 (except for accommodations)[73] was not met, there has been substantial progress made toward that goal. Nonetheless, substantial change still will be required in the elementary grades to achieve total online testing statewide.

District Waiver Policy. Waivers of the requirement to test all students online are granted by the State Board of Education (SBE). A special proviso authorized district requests for waivers in 2017 (proviso 1.88) and 2018 (proviso 1.77). In 2017, the SBE granted 55 waivers, primarily for lack of sufficient infrastructure and testing devices.[74] For 2018, the SBE has received requests from a number of Districts for paper/pencil testing and these requests will be acted upon at the December meeting.[75]

Test Forms. In 2017, there was one unique SC READY ELA and one unique Mathematics operational test form per grade administered online. Two scrambled forms of the operational online test were also created with consideration given to item position effects.[76] There was also one unique operational test form per subject and grade for paper/pencil administrations that consisted of the online form items with a handful of substitutions for technology enhanced items that could not be reproduced on paper. No paper/pencil scrambled forms were created because the 2017 paper/pencil testing population was expected to be small due to the statutory requirement that all assessments be administered online.[77]

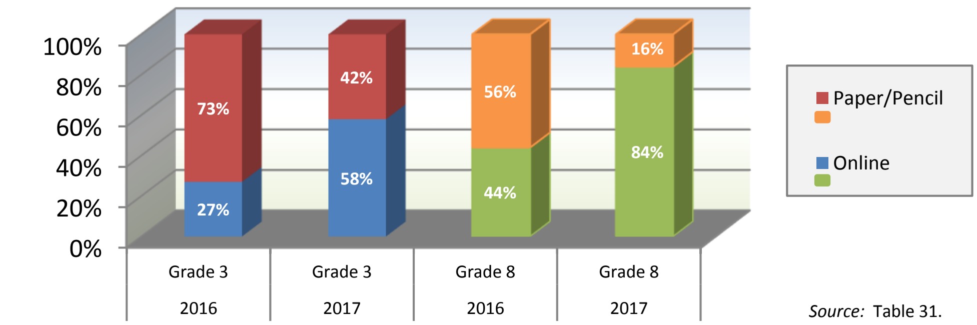
[71] Test Administration Manual, *supra* note 32, p. 5.
[72] DRC (2016d). *SC READY and SCPASS Comparability Study: Paper and Pencil vs. Online Administration* [Mode Comparability Study], Maple Grove, MN: Author, p. 4; DRC (Dec. 22, 2017). Further Responses to Questions.
[73] Section 59-18-325.
[74] SCDE Responses to Questions, *supra* note 6, p. 10; SCDE (2017d). *List of Waivers Granted by SBE for 2017*.
[75] SCDE Responses to Questions, *supra* note 6, p. 10.
[76] DRC Responses to Questions, *supra* note 6, p. 11.
[77] *Id*.

*Table 31. SC READY Percent of Students Testing Online and Paper/Pencil**

	N (Math)	2016		N	2017	
		Online	Paper Pencil		Online	Paper Pencil
Grade 3	59,652	27%	73%	59,740 (ELA)	58%	42%
Grade 4	57,107	28%	72%		NA	NA
Grade 5	55,624	30%	70%		NA	NA
Grade 6	55,127	39%	61%		NA	NA
Grade 7	54,999	41%	59%		NA	NA
Grade 8	54,957	44%	56%	55,203 (Math)	84%	16%

* NA=data not available; *Source:* Mode Comparability Study, p. 4
DRC, Personal Communication, Dec. 22, 2017.

CHART 10
SC READY Online and Paper/Pencil Administrations



Source: Table 31.

The window for Spring 2017 test administration was about two months (April 7-June 12) for both online and paper/pencil administrations. Districts were required to administer the SC READY tests within the last 30 days of their school calendars.[78] The 2017 operational forms also included a small number of embedded, nonscored field test items randomly assigned to students. A total of 40-48 new ELA and 27-37 new Mathematics items were field tested to augment the item bank.

Online Test Engine. To gather information about the online testing platform, the OTT was completed and the Grade 3 ELA and Grade 8 Mathematics tests were experienced online. The OTT was found to be helpful and complete, the navigation tools were easy to use, and the displays were clear and intuitive.

Mode Comparability Study

At the request of SCDE, the contractor completed a mode comparability study for the online and paper/pencil forms using the Spring 2016 field test data. Differential item functioning (DIF)

[78] Test Administration Manual, *supra* note 32, p. 5; Technical Manual, *supra* note 13, p. 17.

statistics were computed for items administered both online and paper/pencil. The larger paper/pencil test takers served as the reference group and the online test takers the focal group. The purpose of the study was to determine if any items favored either group.[79]

The same three item classifications used to evaluate gender and ethnic item DIF were used in this study. Recall that the results of most concern are outlier items with DIF statistics in the C category. Only two of 449 (about ½%) of the SC READY ELA operational items had *C DIF* statistics (one each in Grades 5 and 8). For SC READY Mathematics, no *C DIF* items were identified. Eleven (about 2½%) of the ELA items and five (about 1½%) of the Mathematics items were classified as moderate *B DIF* items. The *B DIF* and *C DIF* items were distributed across multiple grades for each subject.[80]

The mode comparability study also examined p-value differences for online and paper/pencil tests. Summed across all the items, the study found an advantage for paper/pencil of about 1½ to 3⅓ raw score points for ELA and .03 to .62 raw score points for Mathematics.

Separate Online and Paper Scale Score Tables in Some Grades

The overlap of identical items for the online and paper/pencil forms was approximately 90%. Differences between online and paper/pencil forms for some grade/subject combinations involved a few technology enhanced online items that were replaced with selected response items on the paper/pencil forms. When the items for the online and paper/pencil operational forms were not identical, separate scale score tables were created. Separate 2017 raw score to scale score conversion tables were constructed for Grade 3 ELA and Grade 3 Mathematics.[81]

Future Plans. Separate scale score tables will continue to be created for SC READY paper/pencil test administrations in 2018 and beyond when the online test contains technology enhanced items that cannot be administered on paper and must be replaced with a companion item testing the same content in a selected response format. The need for paper/pencil administrations and separate scales will depend on how quickly the South Carolina districts receiving waivers are able to transition to total online administrations, except for a small number of accommodations for students with disabilities. Mode comparability equating should remain a priority as long as a considerable number of students continue with paper/pencil testing. The contractor and SCDE are planning discussions with the Assessment TAC at 2018 meetings to consider how the vertical scale scores should be equated and when to revalidate them.[82]

Online and Paper/Pencil Testing Accommodations Policies

The comparability of online and paper/pencil test administrations is partly a function of their respective testing accommodations policies. Online administrations have an advantage because some features can easily be made available to all students, potentially decreasing the number of students requiring special accommodations. Moreover, even for special accommodations, it may be easier to provide them with technology than with human intervention. Nonetheless, some students with disabilities may be unable to test online so there probably always will be a need for paper/pencil forms for a small number of students. However, once nearly all districts and students are testing online, comparability and equating studies for online and paper/pencil forms may be discontinued.

[79] Mode Comparability Study, *supra* note 70, p. 1.
[80] *Id.*, p. 5-7.
[81] DRC Responses to Questions, *supra* note 6, p. 10-11; Vertical ELA and Mathematics Raw to Scale Score files.
[82] DRC Responses to Questions, *supra* note 6, p. 6.

The ADA and IDEA and their corresponding Regulations require that students with disabilities be tested with accommodations specified by their IEPs or with an alternate assessment.[83] However, the state has appropriately classified accommodations as standard and nonstandard. Nonstandard accommodations alter the tested construct and/or produce noncomparable scores and are counted as not proficient for federal accountability under the ESSA.[84]

South Carolina has a detailed and strict policy for testing accommodations.[85] The Individualized educational program (IEP) team for each student with a disability determines the appropriate accommodations for each SC READY test. A very small proportion of students with severe cognitive disabilities are administered an alternate assessment. A unique feature of this policy is that both failure to provide a needed accommodation and providing the wrong accommodation are considered test security violations.[86] Many timing, format and setting accommodations are standard and are listed in the TAM. Accommodations classified as nonstandard include oral administrations for ELA Grade 3, use of a calculator on the no calculator sections of the Mathematics tests in Grades 6-8 and use of a dictionary or a thesaurus for the ELA TDA essay item.[87]

Online and Paper/Pencil Test Security Policies

Test security policies include all the codified and written rules and regulations for secure test administration. These rules are contained in statutes, regulations, test administration manuals, confidentiality agreements and other official test directives. These documents specify the responsibilities of test administrators and conduct that constitutes punishable violations. The following sections provide evidence of SC READY test security policies based on an evaluation of the consistency of test administration procedures with relevant Standards from the 2014 *Test Standards* and legal prescriptions from South Carolina laws and regulations.

Test Administration. Test administration procedures are critical for ensuring the validity of the resulting test scores by maintaining standardized testing conditions and security procedures. All students should be afforded the same opportunity to demonstrate what they know and are able to do with no favoritism or disadvantage. According to the 2014 *Test Standards*, test users have a responsibility to create fair testing conditions for all students by minimizing the potential for adverse effects on the validity of test scores from improper test administration or inadequate test security.[88]

HumRRO evaluated the test administration procedures for the SC READY assessments based on available documentation. Fourteen Standards directly relevant to test administration were identified for evaluation. The SC READY documentation was rated for consistency with the best psychometric practices described in these Standards. The same 5-point scale, ranging from no evidence to full coverage, used to rate consistency with the forms construction *Test Standards* was also used here. The results are summarized in Table 32 and described in more detail in Chapter 4 (Task 4).

[83] Americans with Disabilities Act (ADA), 42 U.S.C. § 12101 et seq. (1990); Individuals with Disabilities Education Act (IDEA), 20 U.S.C. §1400 et seq. (1991).
[84] ESSA, *supra* note 39.
[85] *See* Test Administration Manual, *supra* note 32, Appendix C.
[86] *Id.,* p. 13, Appendix C.
[87] *Id.*, Appendix C; Technical Manual, *supra* note 13, p. 27.
[88] *Test Standards, supra* note 4, Standard 6.6.

Table 32. Ratings of SC READY Consistency with Identified Test Administration Standards

Standard	Description	Rating	Comments
3.10	Document standard provisions for using and monitoring appropriate implementation of test accommodations	4	Training, Appendices C and D of the TAM for paper/pencil and the eDIRECT User Guide for online testing contain standard provisions for accommodating students with documented disabilities and English learners; monitoring appropriate implementation is not covered
4.5‡	Document and provide a rationale for permissible variations in test administration conditions	5	Documentation is in the TAM, administrator training materials and the eDIRECT User Guide; a study conducted by SCDE and the contractor concluded that oral administration of the ELA test does not adversely affect test score validity in Grades 4-8
4.15	Administration directions with sufficient clarity for replication of validity and reliability data; document the process for reviewing requests for additional testing variations	5	Detailed directions and scripts are provided in the TAM and ADM for test coordinators, administrators and monitors for online and paper/pencil administrations; also covers written requests to SCDE for variations
4.16	Provide practice questions, scoring criteria and instructions with sufficient detail to respond as intended prior to testing	4	Online Tools Training (OTT) and practice tutorials with sample item types and scoring rubrics are available to students online in advance of testing; no separate practice materials for accommodations
6.1	Test administrators should follow documented, standardized procedures	4	Training workshops, including test security case scenarios, TAM/ADM documentation and signing a confidentiality agreement provide adequate instructions; no documentation of usability studies
6.2	Inform test takers of any formal procedures for requesting and receiving accommodations in advance of testing	4	The TAM specifies that parents must be notified of testing schedules, formats and accommodations in advance; a FAQ for accommodations is posted online; lists may be accessed/updated by test coordinators
6.3*	Document and report any disruptions in standardized test administration procedures	3	Procedures for documenting, reporting and investigating test security violations exist; unclear if similar procedures for testing irregularities (e.g., fire alarm) exist
6.4	Furnish a reasonably comfortable testing environment with minimal distractions	5	The TAM has suggestions for creating a supportive testing environment; technical documentation for the online test delivery system is also available to deal with internet connectivity and technology issues
6.5	Provide appropriate instructions, practice and support necessary to reduce irrelevant test score variance	4	TAs are responsible for students' prior review on the correct device of tutorials with instructions for navigating the online test delivery system, using the available tools, and responding to test questions
6.6	Make reasonable efforts to eliminate opportunities for students to attain scores by fraudulent or deceptive means	5	TAM requires seating charts, checklists for secure test materials, appropriate separation of students, prohibition of electronic devices and signed confidentiality agreements for educators specifying prohibited behaviors
6.7	Protect the security of test materials at all times	5	Qualifications, responsibilities and required training for test coordinators, administrators and monitors, and procedures for handling, storing and returning secure test materials are specified in the TAM
7.7*	Specify qualifications required to administer, score and interpret test scores accurately	5	Qualifications for test administrators are in the TAM and for scorers of essay questions are in the Technical Manual; Score Report User's Guide covers score interpretation by educators; information for parents is in brochures and online
7.8*	Provide detailed documentation for test administration and scoring	4	TAM, ADM, eDIRECT User Guide, training requirements provide specifics; dedicated Help Desk or user hotline for questions desirable but not in documentation
7.9*	Maintain test security by following documented steps for protecting test materials and preventing inappropriate exchange of information during test administration	5	Detailed information in TAM and training materials; supported with checklists, seating charts, prohibition of electronic devices during testing and mandatory district test security policies; monitoring permissive in regulations

‡ Also evaluated for Task 3; * Also includes scoring; TAM=Test Administration Manual; ADM=Administration Directions Manual; TA=test administrator

Source: Chapter 4, TAM and eDIRECT User Guide.

The average rating for consistency with the identified Standards listed in Table 32 is 4.4, and most Standards received high ratings. The only exception was Standard 6.3 rated 3=some supporting evidence. Standard 6.3 was rated lower because the documentation did not adequately cover how testing irregularities during test administration should be documented and reported. It appeared from the TAM that these decisions are left to the local district.[89]

SC Test Security Laws and Regulations. South Carolina has strong laws and regulations regarding test security policies. S.C. Code Ann. §§ 59-1-445 (2004) addresses violations of mandatory test security regulations. §§ 59-1-445 states that it is unlawful to knowingly and willfully violate test security procedures stated in regulations adopted by the SBE for mandatory state tests and lists the following actions as violations:

(a) Giving students access to test questions prior to testing;

(b) Copying, reproducing, or inappropriately using any portion of a secure test booklet;

(c) Coaching students during testing or altering/interfering with students' responses;

(d) Giving students access to answer keys;

(e) Failing to follow prescribed security regulations for distribution, return and accounting for secure test materials at all times;

(f) Participating, directing, aiding, counseling, assisting, encouraging or failing to report any test security violations.

The statute also specifies investigation of allegations by the South Carolina Law Enforcement Division and states that violators are guilty of a misdemeanor punishable by a fine of not more than $1,000 and/or imprisonment for not more than ninety days. Administrative and/or teaching credentials of convicted violators may also be suspended or revoked.

S.C. Code Ann. §§ 59-1-447 (2004) requires the State Board of Education (SBE) to adopt regulations detailing mandatory test security procedures. Those regulations are reprinted in the TAM and codify the following affirmative duties:

(a) District school boards must develop and adopt a district test security policy for online and paper/pencil assessments and keep written test materials in locked storage while in the possession of the district and not in use for testing.

(b) Districts and schools must annually designate in writing to the SCDE an individual responsible for all testing activities.

(c) Educators involved in testing students must follow all procedures specified in SCDE manuals for mandatory testing programs.

(d) The SBE has discretionary authority to invalidate test scores with improbable gains unexplainable by changes in the student population or instruction and any evaluative criteria based on the test scores will be deemed unmet.[90]

[89] Test Administration Manual, *supra* note 32, p. 29.
[90] 2 S.C. Code Ann. Regs. 43-100 (2015).

In addition, the SBE Regulations identify the following actions as breaches of professional ethics that may jeopardize the validity of inferences from test scores and constitute test security violations subject to criminal prosecution and/or revocation of an educator's professional license:

(1) Violations of (a) through (d) from the statute listed above;

(2) Failure to administer tests on SCDE specified dates, maintain an appropriate testing environment free from distractions, and/or proctor a test to ensure students are engaged in appropriate test-taking activities;

(3) Failure to follow all test administration directions in the manual for the test, including failure to follow directions in the manual for clearing the memory of calculators used for testing;

(4) Disclosing or discussing the content of secure test materials with students or other educators before, during or after testing;

(5) Leaving content related materials in view of students during testing;

(6) Providing students with reference materials or tools other than permitted by the manual or at prohibited times;

(7) Failure to provide test accommodations specified in a student's IEP/504 plan or providing test accommodations not specified in the plan;

(8) Excluding or exempting students who should be testing or failing to return test materials for all students;

(9) Engaging in inappropriate test preparation practices that invalidate the test scores, including activities that increase test scores without simultaneously increasing students' knowledge and skills in the content area tested;

(10)Revealing test scores to anyone not responsible for the student's education;

(11)Altering test scores in electronic records or files;

(12)Failure to report a test security breach.[91]

Finally, the SBE Test Security Regulations provide that:

(a) The SCDE has the right and responsibility to monitor adherence to test security policies by observing test administration activities without prior notice.

(b) Test security violations must be reported to the SC Law Enforcement Division (SLED).

(c) The SBE may order funds equivalent to replacement costs withheld from Districts where test security violations render test items unusable.

(d) The SBE may publicly or privately reprimand or suspend or revoke the credentials of an educator who violates test security policies.[92]

SC READY Policies. The TAM includes procedures for reporting test security violations to the SCDE. Educators and the public can also report incidents anonymously to the test security manager in the assessment office of the SCDE. The District is required to investigate and document the incident on a form available online. Directions for conducting, documenting and providing supporting evidence for an alleged violation are also provided in the TAM. The SCDE determines whether the gravity of the incident warrants reporting to SLED. Note that these procedures apply, per regulation, to any deviations from test accommodations prescribed in a

[91] *Id.*
[92] *Id.*

student's IEP/504 plan. The TAM includes additional guidelines for handling such situations for paper/pencil and online tests. In particular, the IEP/504 team must be reconvened to determine the validity of the resulting test scores.[93]

All educational personnel with access to secure test materials must sign an *Agreement to Maintain Test Security and Confidentiality* form after completing training. The forms are returned to the District Assessment Coordinator (DAC) who must store the forms for five years. Online student testing tickets with usernames and passwords are considered secure and must be collected and securely destroyed.[94] Seating charts are mandatory for all test sessions and must be submitted to the contractor. The initials of the test administrator must be coded on the student's answer document. Test start and stop times are also coded on the answer document.[95]

Site Visits. The statute permits unannounced visits to schools during testing administration to check adherence to test security procedures. For Spring 2017 SC READY testing, the SCDE selected and conducted site visits for 15 schools. Schools were selected based on erasure gains, frequency of prior test security violations, and recommendations of school and district test coordinators. Monitoring checklists developed by the SCDE are used to record observations during site visits.[96]

2017 Investigations. During the 2017 SC READY testing window, a total of 186 test security violations were reported and investigated. Of these, 176 (95%) were found to be statutory or regulations violations and 10 were judged to be lesser test irregularities. SCDE responses included reporting the violation to SLED, requiring action/improvement plans, and supporting disciplinary actions for violators that were imposed by the district.[97]

Evaluation

There are several topics related to online and paper/pencil testing that may warrant some additional attention. They are included within the broader categories of mode comparability, online test administration, test security, testing accommodations and full attainment of technology goals and capabilities.

Mode Comparability

Although the mode comparability study identified very few items with significant mode differences, the p-value analysis indicated a clear advantage for paper/pencil for the ELA tests. However, this methodology did not account for differences in ability between the two groups. For example, in Grade 4, the online group (28%) had an average Rasch ability of 0.40 and an average raw score of 43 out of 70 possible points. The paper/pencil group (72%) had an average Rasch ability of 0.58 and an average raw score of 46. Although it appears that the paper/pencil advantage may have been due in part to a 0.18 average ability advantage, from these data one cannot determine with certainty to what degree the mode of test administration contributed to the 3-raw-score-point performance advantage for the paper/pencil group.

[93] Test Administration Manual, *supra* note 32, p. 17-18; SCDE Responses to Questions, *supra* note 6, p.9.
[94] DRC (2017b). *Materials Receipt and Return Supplement,* Maple Grove, MN: Author.
[95] Test Administration Manual, *supra* note 32, Appendix A. *See also,* SCDE (2017e). *SC READY Administration Directions Manual,* Maple Grove, MN: Author.
[96] SCDE Responses to Questions, *supra* note6, p. 9.
[97] *Id*.

To evaluate whether there is a true mode advantage for paper/pencil ELA test takers, one could conduct a linking study using matched samples. A common method for doing this is to choose the closest matched student from the larger paper/pencil group for each student in the smaller online group. Matching variables may include available demographic (e.g., gender, ethnicity) and achievement variables (e.g., prior year test scores). An equating analysis is then conducted on the matched samples of approximately equal ability to determine if there is a mode difference large enough to be practically significant.

Practical significance requires a judgmental rule for determining when the tests should be equated to maintain score comparability. In other applications, decisions to conduct mode equating have been made when the average difference is more than one raw score point or when differential advantages were observed in specific segments of the test score distribution. For example, the average mode difference may be less than one raw score point, but high scoring students may have a two point advantage online while low scoring students have a one point advantage for paper/pencil.

The issue here is that if educators whose schools are being evaluated based on test scores believe there is an advantage to paper/pencil testing, particularly for low-achieving students, then they may be more reluctant to convert to online testing. One way to convince them that the process is fair is to equate the test forms from the two modes when the equated raw score differences for groups of equal ability exceed a predetermined criteria (e.g., one raw score point) on average or in substantial portions of the test score distribution.

Cumulated over many students, the unadjusted raw score point advantages or disadvantages could make a difference for a school’s accountability rating. On the other hand, if equated scores are reported for the entire distribution when the evidence indicates the mode differences are more than negligible, schools can be assured that scores for all students are comparable. As students become more familiar with testing online and increasingly fewer schools are testing paper/pencil, mode differences may disappear and the equating studies can be discontinued. Meanwhile, there will be no performance incentive for educators to prefer administering paper/pencil tests.

Based on the available evidence from the 2016 Mode Comparability Study, the Mathematics mode differences are probably too small to be practically significant. However, it might be useful to complete the equating studies for Mathematics along with those for ELA for one testing cycle to verify that the differences are small enough to be ignored.

Online Test Administration

SC READY ELA Reading items associated with text passages require more than one screen to display the passage. The contractor’s test engine uses a vertical split screen to display the item next to the passage, and single clicks can be used to page forward or backward within the passage. An alternative method for moving through multiple screens of text is scrolling. As the HumRRO evaluators observed in Chapter 4, the EOC tests use scrolling and the SC READY tests use pagination, but no usability studies have been reported to support these decisions. When experiencing the SC READY ELA online tests using the contractor’s test engine, the text pagination was intuitive, easy to use and simulated reading a paperback or digital book. But the rapid, page-turning movements were a bit distracting and uncomfortable visually. To make the pagination more comfortable for students, a slightly slower page turn or dissolving to the next page might be helpful.

Test Security

Test administration and security policies for SC READY are detailed and strict. Reporting of violations is mandatory and the statutory provisions and administrative rules provide clear guidelines for investigations and sanctions for violators. However, despite admirable test security policies, there are other important actions the state should consider to bolster test security and support the validity of the test scores.

Backup Test Forms. It is risky to have only one test form for a two-month test window. Currently, SC READY assessments include one online form and one paper/pencil form with over 90% identical items administered in a two-month testing window. If a test form were to be compromised for any reason (e.g., items posted on the internet or shared with news media as happened in cases in Georgia and Michigan[98]), the state has no options for assuring the validity of the test scores obtained during subsequent administrations within the testing window.

In addition, the ELA TDA essay items are likely to be more memorable than other items, and it is virtually impossible to prevent students from discussing their testing experiences with parents and friends outside of school. Therefore, students testing late in the test administration window may have advance knowledge of the topic of the TDA essay item or other memorable test content. If this occurs, the state may not be able to support the validity and comparability of the test score interpretations for all students. In addition, if scores for a classroom or school were invalidated due to adult malfeasance, no retesting would be possible to provide valid test results for the affected students. Consequently, at a minimum, the state needs at least one backup form per subject and grade level held in reserve in the event the operational testing form is compromised before all schools have completed testing.

Detection of Violations. For test security policies to be effective in ensuring valid test scores, active monitoring and consistent enforcement are necessary. Although not all detection activities will result in sufficient evidence to prove a violation, detection activities can indicate areas where further investigation is warranted. For example, periodic internet searches may detect secure item content, data forensics may indicate improbably large score gains, and random site visits during testing may discover improper test administration practices.

SBE Regulations give the SCDE the right and responsibility to conduct unannounced site visits during testing to monitor adherence to mandated test security policies. In 2017, SCDE conducted 15 site visits to detect possible violations of SC READY test security policies. These monitoring visits should be continued and strengthened in 2018 and beyond.

Fifteen site visits annually is probably too few to provide sufficient coverage and deterrence. Although resources for site visits may be limited, it may be possible to supplement departmental resources by leveraging connections with college and university staff researchers, graduate students taking an evaluation course or third party contractors hired to provide checks on other aspects of the testing program. Related agencies within state government may also be able to assist on a short term basis. In any case, persons conducting site visits must be well trained to employ consistent, standardized procedures that create sufficiently-detailed, credible documentation that provides useful evidence when further investigation or corrective action is warranted.

If not already completed, it would be advisable to develop a written plan for site visits and seek the advice of the Assessment TAC. The plan should include procedures for selecting sites; constructing

[98] Phillips, S.E. (2010). *Assessment Law in Education*, Phoenix, AZ: Prisma Graphic, p. 336.

standardized forms for questioning, collecting information and observing that may need to be more detailed than a checklist; and creating a scale for consistently rating and classifying violations by type and seriousness (e.g., major validity threat, minor test irregularity) along with the assignment of appropriate corrective action(s). Because South Carolina has strong test security laws and regulations, appropriate action(s) can be instituted when security violations are detected. In addition to corrective action when warranted, positive feedback can be shared with schools that are doing a good job with test security. If SCDE has reason to suspect violations in particular schools, these schools should be prioritized for monitoring visits. But other schools should also be randomly selected for visits, and all districts should receive at least one monitoring visit over a period of a few years.

Testing Accommodations

South Carolina has a clear and detailed policy for testing accommodations. Decisions are made by the student's individualized education program team and are considered security violations if not administered as prescribed. There are appropriate procedures for requesting accommodated testing forms and the online test engine has several useful features available to all students. The Test Administration Manual and the required training for testing personnel provide helpful information for implementation of the state's Testing Accommodations Policy. Nonetheless, there are a few areas for which a closer examination of the validity of test score interpretations may be beneficial.

Oral Administration of Reading Tests. The Achieve Report comparing state content standards found that South Carolina standards include reading fluency standards all the way up through the upper grade levels. If the intent is for teachers to continue to work with students on decoding, fluency and phonetic reading skills as the complexity of reading texts increase across grade levels, the decision made to classify oral ELA test administrations for Section 2 (reading literary and informational texts) as nonstandard accommodations in Grade 3 may also be appropriate for the other elementary grade levels and maybe even for some middle school grades. This position is supported by the lexile® linking study that removed about 2% of the sampled students from the calibrations because they had received oral test administrations.[99]

Alternatively, if the state intends only reading comprehension to be the focus of the Reading tests in Grades 4-8, and reading comprehension and listening comprehension are viewed as equivalent, interchangeable communications skills, then it may be appropriate to continue classifying oral ELA Section 2 test administrations as standard accommodations. The important questions to be considered for a standard accommodation, as acknowledged in the Test Administration Manual, is whether (1) the measurement of the intended construct is preserved, and (2) the resulting scores are comparable to the scores for students tested under standardized conditions. Satisfying both requirements supports valid test score interpretations and ensures that the knowledge and skills intended to be measured by the content standards are congruent with the tested knowledge and skills.

Universal Design. Universal design is a process for creating test items that are accessible to the widest possible student population. Universal design procedures can improve testing for all students by simplifying unnecessary complexity, using unambiguous and easily understood language and rendering accompanying graphics more usable and interpretable. The contractor has developed an accessibility guidelines document that provides helpful examples of how to improve item accessibility for all students.[100]

[99] *Lexile® Linking Study, supra* note 9, p. 28.
[100] DRC Accessibility Guidelines, *supra* note 64.

Given limited resources, it is tempting for a testing program to require all test items to be usable with all accommodated student populations. Yet, a word of caution is in order to remind test developers to ensure that important skills specified in the content standards are not oversimplified or eliminated when universal design principles are applied. For some accommodations, the number of students needing them is relatively small and altering or eliminating items from operational forms to address these needs may adversely affect the validity of the resulting score interpretations for other students because the knowledge and skills being assessed have changed. Just as item substitutions for complex technology enhanced items are made for paper/pencil versions of online tests, strategic item substitutions can be made for needed accommodations when item alterations would adversely affect test score validity. The substitute items could be chosen to provide the closest match possible to the tested content standard and equating procedures implemented to produce comparable scale score conversion tables. For example, such procedures have been common practice in some states when particular items cannot be adequately rendered in Braille for blind students.

Accommodated Practice Materials. The online practice tests do not yet provide options for practicing with all the available accommodations. As discussed by HumRRO evaluators, to the extent feasible, it would be helpful to allow students with disabilities to practice ahead of time with the accommodations they will be using for the operational tests. Guidelines for monitoring and assisting students during practice activities may also be beneficial for all students, but especially students with disabilities.[101]

Monitoring Accommodations. In their evaluation work, HumRRO staff noted that monitoring of the appropriate implementation of testing accommodations is not covered in the documentation. However, test administrator training and the TAM make clear that failure to administer the correct accommodations, or administering accommodations not listed in the student's IEP/504 plan, are security violations. In addition, the confidentiality forms test administrators are required to sign prior to administering tests state that the entire TAM and ADM have been read and understood.

Nonetheless, knowledge of correct procedures does not necessarily guarantee that they are always followed, and because South Carolina statutes and regulations list specific test security violations and penalties, compliance monitoring is appropriate to verify that the training and documentation are communicating effectively and being implemented consistently. One method for monitoring appropriate implementation of accommodations is to make school assessment coordinators responsible for conducting random implementation checks during testing and reporting the results to the district assessment coordinator. Another monitoring option is to use SCDE site visits to examine a sample of IEP/504 plans and compare the specified accommodations with those coded on students' answer documents and/or observed during the site visit. If any implementation problems are detected, correction may include a written bulletin to all districts with reminders and/or additional guidance.

Full Attainment of Technology Goals and Capabilities

South Carolina has made substantial progress moving schools and districts to online testing. But there are still significant numbers of students testing paper/pencil in the lower grades. Providing support and incentives for meeting the near 100% goal will likely remain a challenge.

Testing online has the potential to provide many benefits over paper/pencil. But many of its capabilities have not yet been realized. The online SC READY tests are largely paper/pencil

[101] Chapter 4 (Task 4).

tests administered by a computer. Over 90% of the items are the same, and the unique technology enhanced items for the most part provide a more active method for completing a multi-select, constructed response or matching exercise. Items that simulate an experiment, use a branching strategy or create a work environment where multiple, concurrent measurements are evaluated could more fully utilize the technological capabilities of online testing. In addition, adaptive testing with a sufficiently large item bank could shorten test lengths, provide faster score report turnaround, and enhance test security while maintaining equivalent score accuracy.

8. Information Reported That Can Assist Educators to Align Assessment, Curriculum, and Instruction

Evidence

Educators have several tools available to assist them in using SC READY assessment information to align assessment, curriculum and instruction. Evidence relevant to Legislative Criterion 8 includes the South Carolina ELA and Mathematics content standards, Performance Level Descriptors (PLDs), test blueprints and sample items, SC READY Individual Student Reports, District and School Roster Reports and labels, the eDirect Information Portal and Lexile® and Quantile® Score Reports. These alternative tools for using SC READY assessment information are discussed in more detail in the next sections.

South Carolina Content Standards for ELA and Mathematics

The starting point for curricular planning is the state content standards that describe the knowledge and skills students are expected to learn and teachers are expected to teach for Grades 3-8 ELA and Mathematics. Content standards are academic statements that describe the content knowledge, skills, and cognitive processes student must demonstrate to achieve grade-level expectations. The state content standards can be cross-referenced to instructional textbooks and other instructional materials currently in use, or being considered for adoption, to identify any important content that is not included and will need to be supplemented.

Performance Level Descriptors (PLDs)

PLDs provide additional information and detail for educators to understand the specifics of what the state content standards expect students to be able to do. PLDs also provide skill progressions for the content standards that illustrate how students' skills are expected to progress across performance levels from rudimentary skills in the *does not meet expectations* level to partial mastery in the *approaches expectations* level to full achievement in the *meets expectations* level to expanded application of the prescribed skills in the *exceeds expectations* level. Examples of PLD progressions across performance levels for two ELA Reading standards are illustrated in Table 18.

Test Blueprints and Sample Items

The test blueprints outline the skills tested on the SC READY assessments. The number of items assigned to each skill indicates the relative weight of that skill in the total test score. The HumRRO alignment studies summarized in Tables 3 and 4 have linked the test blueprints to the state content standards and the test items to the state content standards and test blueprints. Sample items on the SCDE website and sample items from the Online Training Tool (OTT) provide specific examples to guide teachers in understanding how the listed content will be assessed.

SC READY Individual Student Report

The SC READY Score Report User's Guide explains the scores reported on the 2-page Individual Student Report (ISR). These scores include performance levels, percentile rank comparisons, scale scores, performance by reporting category, and text-dependent analysis (TDA) essay score information (see the sample ISR in Exhibit C). An introductory section of the User's Guide provides an overview of the SC READY assessment program including a description of the types of test items, scoring of items, alignment to standards, test blueprints, reporting categories, performance levels, scale scores, percentile ranks, and special notations on score reports.

Referring to the ISR, educators and parents can see at a glance whether the student *meets expectations* for ELA, Reading and Mathematics. Percentile rank comparisons provide a normative indication of the student's overall performance relative to the state and other states with comparable standards. These scores provide an overall indication of the strength of the student's academic achievement and an initial identification of students who have not met grade level standards or whose performance is seriously lagging behind that of other students in the two comparison groups. These are students for whom additional remedial instruction may be prescribed. Scale scores for the current and previous SC READY tests taken by the student indicates how the student's achievement has progressed over time. This information may show that a student has been struggling academically for multiple years, has shown improvement in the last year or has done especially well or poorly in the current year.

Reporting category scores on page 2 of the ISR provide additional diagnostic information to help educators and parents understand the strengths and weaknesses of a student's overall performance in ELA and Mathematics. However, diagnostic scores from a summative assessment have more uncertainty than total scores because they are based on a much smaller number of items. The SC READY assessments provide appropriate information for the reporting category diagnostic scores by using the summary descriptors of *low, middle* or *high performance*. Reporting categories with *low* indicators identify weak content for which review or remedial instruction may be warranted. To identify specific skills for targeted instruction, educators can consult the PLDs for the content standards corresponding to the reporting categories with *low performance* indicator scores. The *Test Standards* also recommend that educators combine test scores with other information available about the student to make appropriate instructional or placement decisions.[102]

The individual student score report also underscores the importance of recognizing the uncertainty of a single score obtained at one point in time by providing confidence intervals for the ELA, Reading and Mathematics total scale scores. These confidence intervals provide a range of likely performance if the student were to retest under similar circumstances. By referring to the number line shown above the reported confidence interval, one can see if the student's performance level might change if the student were retested. For instance, students whose confidence intervals cross the borderline between *meets* and *approaches expectations* might be candidates for some additional review work while those whose confidence intervals cross the borderline between *approaches* and *does not meet expectations* may benefit from more intensive remedial work.

District and School Roster Reports and Labels

District and school roster reports provide summary test information for groups of students. The Preliminary Grade 3 Reading Rosters report, posted by District and updated continuously during the testing window, provides preliminary reading scores to address the Read to Succeed

[102] *Test Standards, supra* note 4, Standard 12.10.

legislation. Students who do not meet the minimal cut score for reading proficiency (Below the *Not Met 1 Reading Cut Score*), based on the February 2017 Standard Setting Meeting described above under Legislative Criterion 5, are identified for summer camp attendance and retesting to meet the standard before entering Grade 4 in the fall. Although these results are useful for school planning for students scoring below the minimum, districts and schools are cautioned that these results may not reflect these students' total ELA scores. A student who is below the minimum standard may score high enough on the writing and inquiry item to be classified as *approaches expectations* on the total ELA test, and students who score above the minimum standard may be classified in any of the four performance levels for the ELA test.[103]

Complete Student Rosters are created for districts and schools and available only through the eDirect online portal. School rosters contain student results listed alphabetically within grades and are produced for origin and fall assignment schools. District rosters are sorted alphabetically by student within grade within fall assignment school. Roster reports list student demographics and SC READY ELA and Mathematics test results including scale scores, performance levels, reading subscores, lexile® and quantile® ranges, South Carolina percentile ranks and other states percentile ranks. Educators can use these results to assess the performance of groups of students by grade level. Drop down menus and options within eDirect allow schools and districts to analyze test information by subgroups and reporting categories. This information allows educators to evaluate instructional weaknesses for specific groups of students or reporting category subsets of items so that appropriate curricular revisions can be considered for the following school year. The roster reports can be opened in Excel where administrators can create their own analyses and summaries of the reported student data.

Labels available for placement in student records allow quick access to test information for educational support staff such as counselors. They also provide summaries of student progress as students move from grade to grade and from one school to another.

eDirect Information Portal

The contractor's eDirect online information portal allows schools to provide the contractor with census, demographic and accommodations information for testing and the schools and districts to receive electronic reports. For example, individual student reports, preliminary Grade 3 ELA Reading roster reports and school roster reports are provided via eDirect.

Lexile® and Quantile® Reports

The Lexile® and Quantile® Score Reports for individual students described in the section on Legislative Criteria 1 and 4 may provide especially useful information for making instructional decisions for individual students. Many educational reading texts and mathematical instructional materials have been placed on the lexile® and quantile® scales, respectively, and their scores can be compared to the ranges reported for students when selecting appropriate instructional materials. Sample Lexile® Framework for Reading and Quantile® Framework for Mathematics maps for typical middle school skill levels with examples designed to assist educators in planning student instruction are reproduced at the end of Exhibit A. Additional instructional suggestions for educators are presented in the linking study reports and the contractor's websites.[104]

[103] Score Report User's Guide, *supra* note 7, p. 6.

[104] Lexile® Linking Study, *supra* note 9, p. 55-60; Quantile® Linking Study, *supra* note 10, p.57, 66-68; www.Lexile.com; www.Quantiles.com.

These reports also indicate whether a student's performance is within the typical range for the student's grade level and whether the student is on target for CCR by the end of Grade 12. If not, a target trajectory is provided indicating the improvement necessary to reach the CCR goal. This trajectory will be revised in subsequent years based on actual SC READY performance. However, one should keep in mind that the elementary grades are a long way from twelfth grade and predictions that far in the future are notoriously unreliable. In addition, lexiles® only measure reading skill, a necessary but not sufficient skill for the other ELA content standards.

Evaluation

The SC READY assessments include informative score reports and user information to aid educators in utilizing the test results to align their curriculum and instruction with the tested content from the state content standards. Appropriate interpretive cautions are also included with the reported scores on the individual student score reports.

Utilizing Test Results. Summative assessments provide an overview measurement of knowledge and skill acquisition for an entire school year. The diagnostic information provided on the individual student score reports can suggest strengths and weaknesses and give educators an idea of where to start looking for content and skills that need to be remediated. The lexile® and quantile® reports can also help teachers choose reading materials and mathematics instructional lessons that address students' weaknesses at an appropriate level of difficulty. But summative assessments are not designed to tell teachers what to teach or how to teach it. Teachers must use their experience and judgment, along with the test results and their own classroom evaluations, to determine what to do next with a particular student.

For students in a teacher's class who have not earned proficient scores (*meets* plus *exceeds expectations*), the SC READY assessment results can signal that a summer school remedial class, individual tutoring or other additional academic work should be advised. The test results can also help teachers explain to parents why a student who *does not meet expectations* is not yet ready for the next grade level or to support a recommendation that a student who *exceeds expectations* enroll in an advanced class, sign up for an extra elective, participate in a gifted program at the local college or engage in other appropriate enrichment activities.

Appropriate Interpretive Cautions. The SC READY individual student score reports include appropriate cautions to encourage valid test score interpretations. Rather than reporting scores for the diagnostic subsets of items referred to as reporting categories, an indicator score (low, middle, high) is reported that is consistent with the lower reliability of scores composed of relatively fewer items. In addition, consistent with professional standards and best practices, a confidence interval is reported along with students' ELA and Mathematics total scores with the explanation that "If your student were to test again under similar circumstances, his/her score would likely remain in the following range: [student's scale score confidence interval]."[105] The Lexile®/Quantile® Score Reports also contain an appropriate score interpretation caution consistent with professional standards that states "The information in this chart is based on a single test score. These data should be considered, along with other information such as school grades, teacher reports, and other test scores, when making instructional decisions about the student."[106]

However, there appears to be some inconsistency between the confidence intervals (ranges) on the sample score reports and the information in the Technical Manual. Typically, confidence intervals for test scores are computed by adding and subtracting one standard error of

[105] Score Report User's Guide, *supra* note 7, p. 13; *see* Exhibit C.

[106] *Id.*, p. 15; *see* Exhibit C and *Test Standards, supra* note 4, Standards 3.18 and 12.10.

measurement (SEM) from the obtained score (score ± 1 SEM). A confidence interval of this size indicates that a student's retest score obtained under similar circumstances would fall in that interval about 68% of the time. The SEM can be an average estimate for the total test or a conditional estimate (cSEM) for the specific score.

Table 33 presents data for a hypothetical sixth grade student from the sample individual score report in the Score Report User's Guide (*see* Exhibit C) and the corresponding standard error estimates from the Technical Manual for the SC READY Grade 6 ELA and Mathematics tests. As indicated in Table 33, the hypothetical sixth grade student has an *exceeds expectations* ELA score of 680 and a *meets expectations* Mathematics score of 548. The SEM for the Grade 6 ELA test reported in the Technical Manual is 23, and the cSEM at the closest cut point (*exceeds expectations*) is 27. Using either of these measures of uncertainty produces typical confidence intervals that substantially overlap the next lower performance level. But the reported confidence interval on the sample score report is only ±10 scale score points, or approximately ± ⅓ to ½ standard error, and does not overlap the next lower performance level.[107]

Table 33. Confidence Intervals for Grade 6 Total Scores

SC READY	ELA	Mathematics
Sample Student Scale Score	680	548
Performance Level	*Exceeds Expectations*	*Meets Expectations*
Performance Level Range	*668 – 900*	*543 – 627*
Test SEM	23.6	29.3
Closest Cut Point cSEM	27.01	28.23
Score ± 1 SEM	656 – 704	519 – 577
Score ± 1 cSEM	653 – 707	520 – 576
Reported Confidence Interval	**670 – 690**	**538 – 558**

Source: Score Report User's Guide, p. 13; Technical Manual, p. 43-44.

Similarly for Mathematics, the total test SEM and cSEM for the closest cut point (*approaches expectations*) are 29 and 28, respectively, and the corresponding confidence intervals again substantially overlap the next lower performance level. Yet the reported confidence interval of ±10 scale score points is about ± ⅓ SEM and only marginally overlaps the next lower performance level.

[107] Technical Manual, *supra* note 13, p. 43-44; Score Report User's Guide, *supra* note 7, p. 13.

Two possibilities for resolving these contradictions are to revise the examples in the sample score reports or provide an expanded explanation of the interval calculations. It may be the case that the examples are not consistent with the data and methods used to calculate the actual ranges reported on the ISR. If so, the examples can be revised to be consistent. It would also be helpful to indicate in the text of the Score Report User's Guide the type and size of SEM used to calculate the ranges. If the example in the sample score report is correct and the ranges are all based on ±10 scale score points, then an explanation should be provided in the Score Report User's Guide for the atypical SEM size chosen and the reasons for this choice.

Task 7: Ratings

The Task 7 legal review examined and evaluated the available evidence to determine whether the 2017 SC Ready assessment system meets the eight minimum legislative criteria prescribed in Section 59-18-325. Based on this review, the eight legislative criteria were rated using the rating scale presented in Table 34.

Table 34. Rating Scale for Legislative Criteria

Rating	Description
Meets +	Robustly meets minimum legislative criteria; evidence is extensive for all aspects
Meets	Meets minimum legislative criteria; evidence is adequate for all aspects
Meets –	Barely meets minimum legislative criteria; evidence is limited for some aspects
Does Not Meet	Fails to meet minimum legislative criteria; evidence is missing or inadequate

The ratings of each of the legislative criteria reflect an assessment of the adequacy and strength of the evidence presented and the degree to which the evidence is consistent with professional psychometric standards and supports the legal defensibility of the assessment program. The ratings for each of the eight legislative criteria with key comments are presented in Table 35.

Summary: ***Overall, the SC READY ELA and Mathematics assessment system meets all of the eight minimum legislative criteria prescribed in Section 59-18-325.*** Policymakers, educators and the public can have confidence that the scores South Carolina students obtain on the SC READY assessments accurately reflect their current achievement of state standards and provide meaningful guidance about their readiness for the academic content of the next grade level. The assessment system effectively utilizes a variety of item types and a comprehensive development and review process to screen, assemble and analyze items aligned to the state content standards. Psychometrically appropriate standard setting procedures were used to establish four student achievement levels labeled *does not meet expectations*, *approaches expectations*, *meets expectations,* and *exceeds expectations*. Online and paper/pencil Test Administration, Testing Accommodations and Test Security policies are detailed, clear and designed to produce psychometrically valid and reliable student scores. Individual student reports present test information clearly and concisely and contain appropriate caveats for interpreting test scores. The best available evidence links the test performance of South Carolina students to the performance of students in other states and to college- and career-readiness. Useful information is provided for aligning curricula/instruction with the assessments.

Table 35. Ratings and Comments for the Eight SC READY Legislative Criteria

RATING	LEGISLATIVE CRITERIA Comments
Meets	**1. LINKED SCALES FOR COMPARISON TO OTHER STATES WITH COMPARABLE STANDARDS** comparison groups are best available but may be nationally unrepresentative, of inadequate size, or have insufficiently aligned content standards
Meets	**2. VERTICALLY-SCALED, BENCHMARKED, STANDARDS-BASED, SUMMATIVE ASSESSMENT SYSTEM** system of grade level, standards-aligned, end-of-year tests with potentially confusing vertical scale scores and *on track for CCR* benchmarks
Meets –	**3. PERFORMANCE AGAINST STATE STANDARDS IN ELA, READING, WRITING AND MATHEMATICS; PREPAREDNESS FOR THE NEXT GRADE; GROWTH** validity studies linking test scores to performance at the next grade level not yet done; vertical scale scores may show negative growth and other growth evidence is indirect; writing is part of ELA but no subscores with achievement levels are reported
Meets –	**4. PROGRESS TOWARD NATIONAL CCR BENCHMARKS FROM EMPIRICAL RESEARCH AND STATE STANDARDS** available CCR evidence is indirect but persuasive; direct CCR predictions for elementary students are ill-advised due to imprecision and unproven validity; inchoate validity studies linking Grade 8 test scores to admissions test CCR benchmarks
Meets +	**5. ESTABLISHMENT OF AT LEAST FOUR STUDENT ACHIEVEMENT LEVELS** appropriate and well-documented standard setting procedures and performance level descriptors for 4 levels (*does not meet, approaches, meets*, & *exceeds expectations*)
Meets +	**6. USE OF A VARIETY OF ITEM TYPES REQUIRING DEMONSTRATION OF CONTENT UNDERSTANDING** mixture of item types; multiple-select, evidence-based & text-dependent analysis essay items simulate the type of thinking and analysis typically associated with CCR
Meets	**7. AVAILABILITY OF ONLINE AND PAPER/PENCIL ADMINISTRATIONS** paper form and easy-to-use online testing platform with appropriate accommodations; online testing goals and capabilities (e.g., TE items; adaptive testing) not yet fully attained
Meets	**8. REPORTS INFORMATION TO ASSIST EDUCATORS IN ALIGNING CURRICULA WITH ASSESSMENTS** summative assessments useful for global curricular alignment; reporting categories guide educators to areas for more in-depth evaluation

As with any new testing program, there are many supporting research studies and procedural decisions yet to be finalized for future test administrations to maintain the quality, equivalence, alignment and usefulness of the test forms. The SCDE has a knowledgeable Assessment TAC and experienced contractor staff to aid them in appropriately constructing and analyzing future test forms and in designing and conducting useful research studies. In the spirit of improving and strengthening the assessment program as these future actions are deliberated, the next section provides specific recommendations related to each legislative criterion. Addressing these recommendations and the suggestions provided in prior sections of this report will further support the psychometric and legal defensibility of the SC READY assessment system.

Task 7: Recommendations

This section of the chapter provides recommendations for improvement. Each recommendation is associated with one of the eight legislative criteria and has been assigned a priority rating of *urgent*, *high*, *medium* or *low* as described in Table 36. The recommendations presented below the table are grouped by priority rating and are identified with the applicable legislative criteria. As indicated by their inclusion in earlier chapters, in addition to improving legal defensibility, many of these recommendations also support improved psychometric defensibility.

Table 36. Priority Ratings for Recommendations

PRIORITY	DESCRIPTION
Urgent	Definitely needs to be considered and addressed now
High	Needs to be considered and addressed as soon as possible
Medium	Should be considered and addressed as time and circumstances permit
Low	Might be considered and addressed as part of long term planning

Urgent Priority

Legislative Criteria 1 & 2: Request that the contractor provide South Carolina with additional validity information about the participating states and the methods used to derive the reported *other states with comparable standards* percentile rank norms. Consider requesting that the contractor organize alignment information similar to a textbook crosswalk (e.g., from the Achieve Report or published state content standards) to confirm the comparability of the other states' standards to those of South Carolina. Also consider exploring the option of reporting percentile ranks for *other states* independent of South Carolina data.

Legislative Criteria 2 & 3: Weigh the advantages against the potential misinterpretations of using the current, vertical scale, and consider adopting a more traditional vertical scale before reporting 2018 SC READY scores to provide reasonable growth score interpretations and avoid the appearance of negative growth. Now is an ideal time to make this change before a second year of comparative data is reported. Score reports for 2018 could report revised 2017 scale scores on the new vertical scale for comparison.

Legislative Criterion 5: Urge the State Board of Education (SBE), with the advice and consent of the Education Oversight Committee (EOC) per Section 59-18-320(D), to officially adopt the SC READY cut scores.

Legislative Criterion 7: Create a backup test form for each grade/subject to be held in reserve in case the operational test form is compromised before all schools have finished testing.

Legislative Criterion 8: Provide additional explanatory text in the Score Report User's Guide identifying the standard error of measurement (SEM) type and size actually used to calculate the scale score ranges reported on the individual student reports, and if necessary, revise the sample reports to be consistent with the actual data.

High Priority

Legislative Criteria 1-8: Consolidate scattered program documents and information into a single, expanded Technical Manual with summarized material and data, relevant appendices, and references to supporting documents.

Legislative Criterion 2: For the Grades 3-8 ELA Reading subscores, report decision consistency estimates and reliabilities obtained using the same methodology as for the total ELA scores. Revise, if necessary, when scores become more stable.

Legislative Criterion 2: To be consistent with the 2014 *Test Standards,* provide estimated reliability data for the reporting category scores now and reconfirm and revise them later, if necessary, when scores are more stable.

Legislative Criterion 4: Consider creating an ELA Writing subscore and reporting performance levels similar to what is currently being done for ELA Reading.

Legislative Criterion 6: Document the frequency of item usage across years and use this information to target items for replacement based on prior exposure.

Legislative Criterion 6: Calculate ethnic differential item functioning (DIF) for Hispanics which represent about 9% of the South Carolina Grades 3-8 student population. Special rules/procedures for small samples may be appropriate for some grade/subject combinations.

Legislative Criterion 6: Consider routine replication of psychometric processing by an independent third party as an additional quality check. This will require more detailed documentation of procedures.

Legislative Criteria 6 & 7: As long as significant numbers of schools continue to census test with paper/pencil, conduct annual mode equating studies for ELA to ensure comparable scores and deter incentives for avoiding online testing. Also do so at least once for Mathematics to confirm that the differences are too small to warrant adjustment.

Legislative Criterion 7: Reconsider whether oral test administrations of the ELA Reading subtest should continue to be classified as a standard accommodation in Grades 4-8 given the reading skills specified by the state content standards.

Medium Priority

Legislative Criterion 2: Design and conduct empirical research studies to validate CCR benchmarks using South Carolina data.

Legislative Criterion 3: Print numerical values next to point estimates on the lexile® and quantile® score report graphs to make year-to-year growth comparisons easier.

Legislative Criterion 3: Conduct research studies to empirically confirm that SC READY proficiency scores indicate adequate preparation for the next grade level for South Carolina students.

Legislative Criteria 3 & 4: Consider placing error bands around the reported lexile® and quantile® growth trajectories using ± 1 SEM estimated from the longitudinal sample. Also consider strengthening the cautionary statements at the bottom of the score reports. Develop a research plan to collect validity evidence to support CCR claims for South Carolina students.

Legislative Criterion 5: For future standard settings, select a wider representation of stakeholders to serve on the vertical moderation panels.

Legislative Criterion 6: Use an index of readability or total word counts to track the reading load for ELA passages and ELA and Mathematics test forms within and across grade levels.

Legislative Criterion 6: Ask the fairness/sensitivity educator committee to re-examine items with gender or ethnic DIF when deciding whether to retain or revise them.

Legislative Criterion 6: Report demographic information for fairness/sensitivity and content review committees similar to that reported for standard setting committees.

Legislative Criterion 7: Expand the number of annual site visits to increase coverage and deterrence. Develop a site visit plan and seek Assessment TAC advice. Select schools where violations are suspected and randomly select others so each District receives at least one unannounced visit over a several year period.

Legislative Criterion 8: Resolve the conflict between the sample score report confidence intervals and standard errors reported in the Technical Manual by expanding the description in the Score Report User's Guide and revising the sample report if appropriate.

Low Priority

Legislative Criterion 2: To be consistent with the 2014 *Test Standards*, report preliminary reliability estimates for the reporting category indicator scores (low, middle, high) now and then revisit and revise them later, as appropriate, when scores are more stable.

Legislative Criteria 2 & 6: Consider convening an experienced educator panel to reconsider the assessment of inquiry skills for ELA and blueprint weights for Mathematics.

Legislative Criterion 6: Consider specifying target depth of knowledge (DOK) levels in the test blueprints to support greater consistency with the content standards, especially for ELA which exhibited the greatest variability.

Legislative Criterion 6: Superimpose cut scores on the item maps and identify the content of the items within each performance level to refine the PLDs and further-strengthen the standards-based validity evidence for the SC READY assessment system.

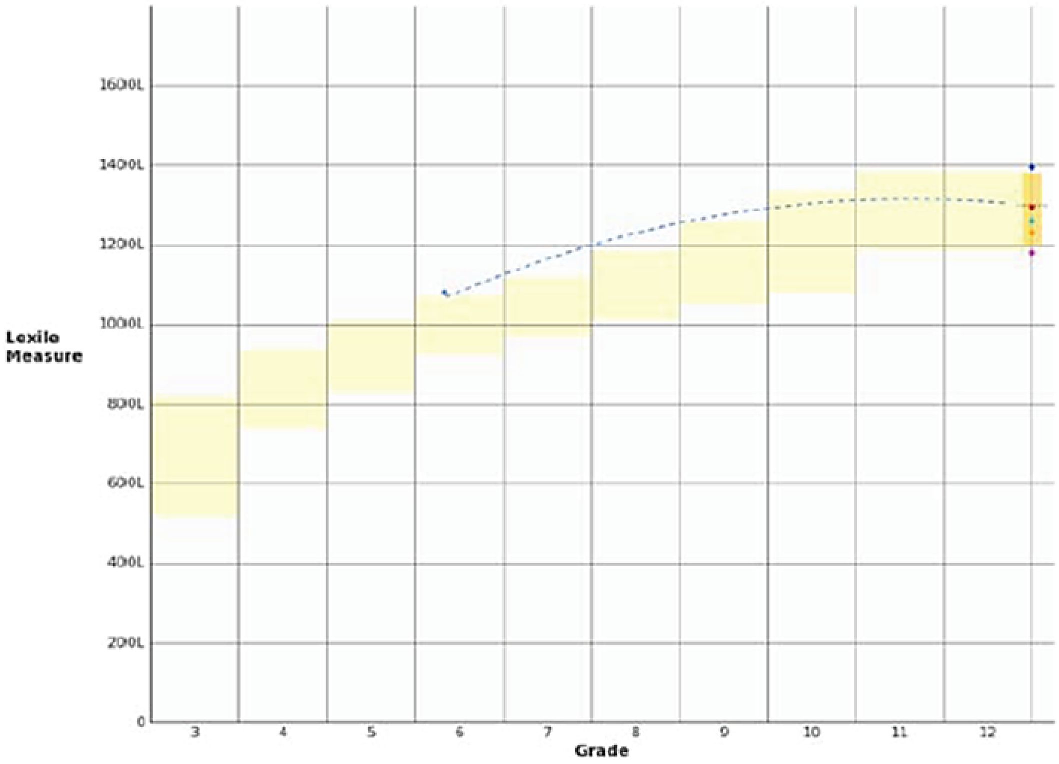
Legislative Criterion 7: Continue to expand the availability of accommodated practice materials. Develop a plan for monitoring the provision of accommodations using school/district testing coordinators and/or site visits.

Legislative Criterion 7: Continue to explore item formats that take full advantage of the technological capabilities of online testing. Consider computer adaptive testing to shorten test lengths and administration times, and speed score reporting while maintaining score accuracy.

EXHIBIT A
SC READY Sample Lexile® and Quantile® Reports

Lexile Range: 1115L-1265L
Lexile Norm Percentile: 79%

LEXILE®
The Lexile® Framework for Reading
Matching readers with texts



• Student Lexile Measures
- - - Estimated Growth Path
▬ College- and Career-Readiness Range (1200L-1380L)
▬ Grade Level Ranges

Domain CCR Estimates

• University (1395L) • Community College (1295L) • Workplace (1260L)
• Citizenship (1230L) • Military (1180L)

EXHIBIT A Cont'd

Quantile Range: 815Q-915Q
Quantile Norm Percentile: 60%

The Quantile® Framework for Mathematics
Linking assessment with mathematics instruction

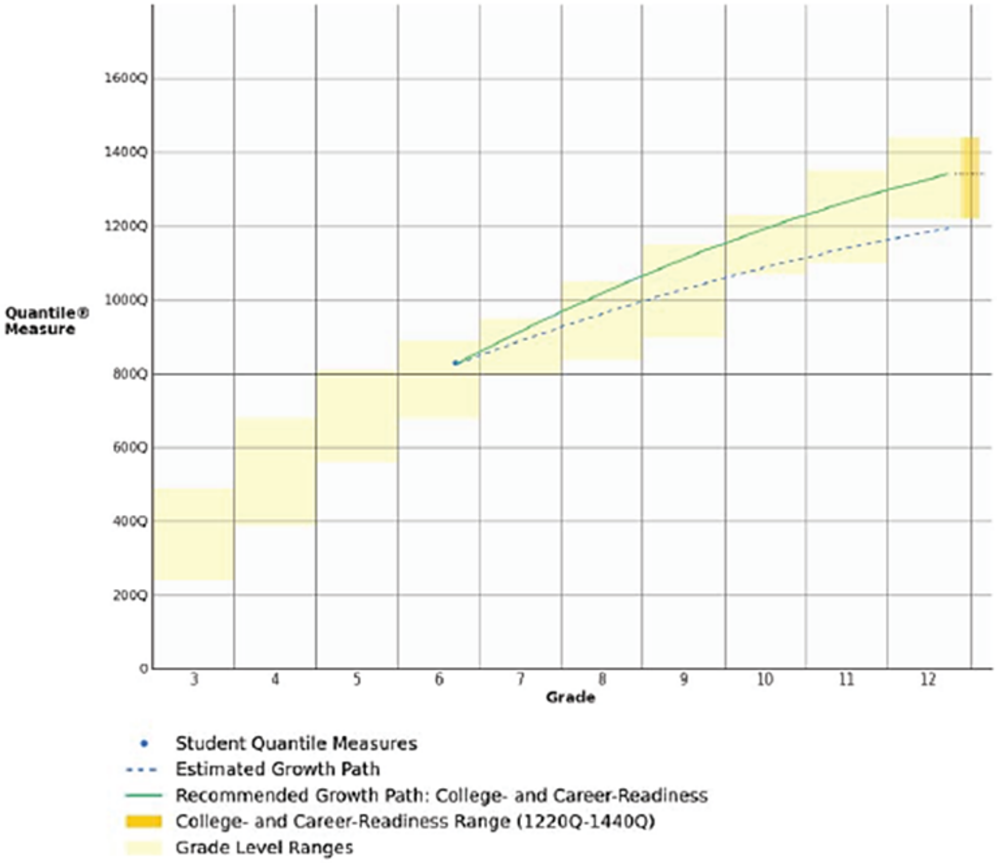
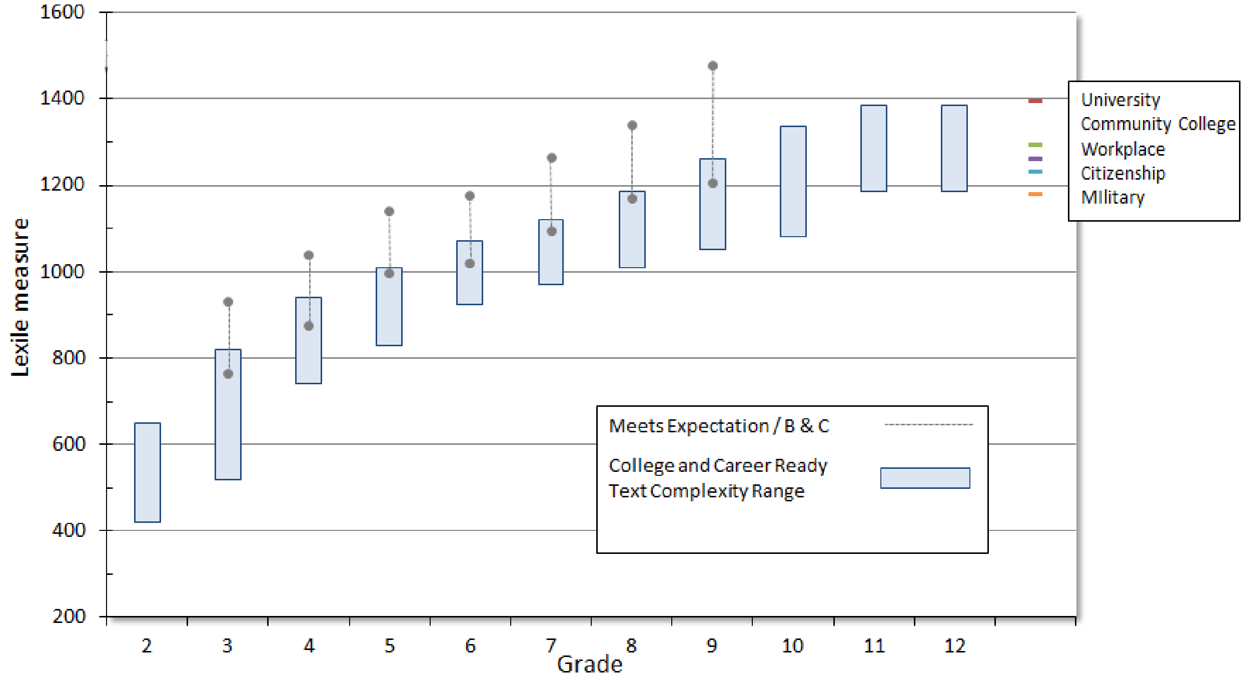


EXHIBIT A Cont'd
SC READY *Meets Expectations* and Stretch CCR (Reading) or Next Grade (Mathematics) Ranges



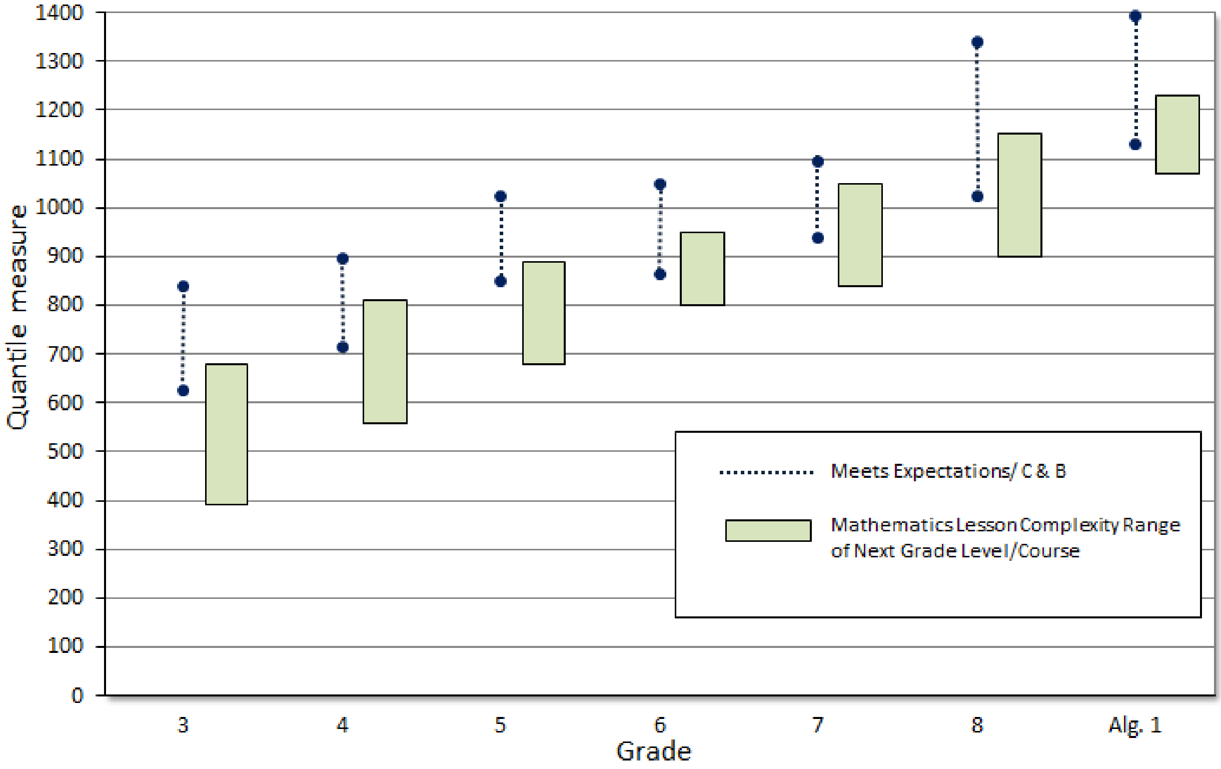
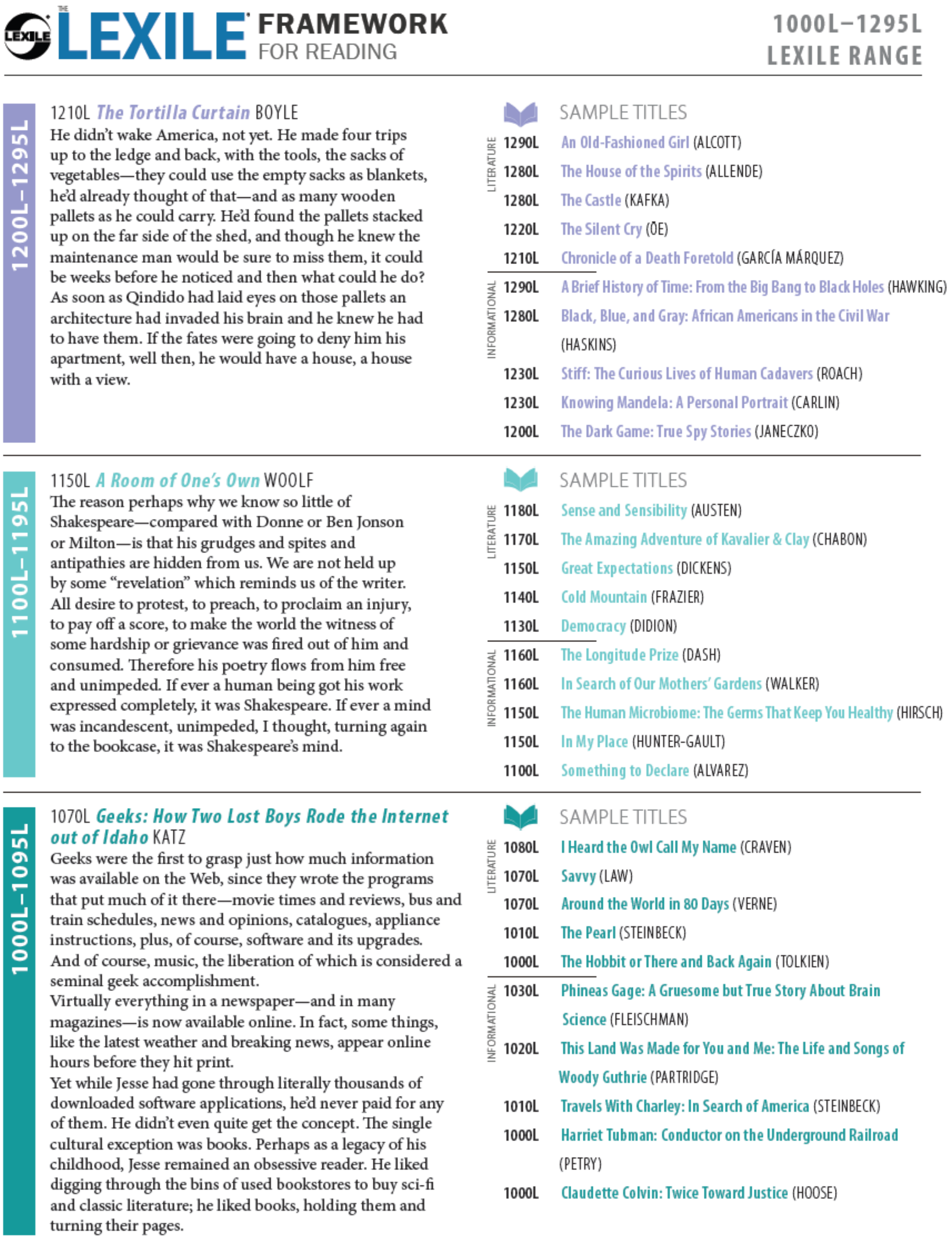


EXHIBIT A Cont'd
Lexile® & Quantile® Framework Maps



THE LEXILE FRAMEWORK FOR READING

1000L–1295L LEXILE RANGE

1200L–1295L

1210L *The Tortilla Curtain* BOYLE

He didn't wake America, not yet. He made four trips up to the ledge and back, with the tools, the sacks of vegetables—they could use the empty sacks as blankets, he'd already thought of that—and as many wooden pallets as he could carry. He'd found the pallets stacked up on the far side of the shed, and though he knew the maintenance man would be sure to miss them, it could be weeks before he noticed and then what could he do? As soon as Qindido had laid eyes on those pallets an architecture had invaded his brain and he knew he had to have them. If the fates were going to deny him his apartment, well then, he would have a house, a house with a view.

SAMPLE TITLES

LITERATURE	1290L	An Old-Fashioned Girl (ALCOTT)
	1280L	The House of the Spirits (ALLENDE)
	1280L	The Castle (KAFKA)
	1220L	The Silent Cry (ŌE)
	1210L	Chronicle of a Death Foretold (GARCÍA MÁRQUEZ)
INFORMATIONAL	1290L	A Brief History of Time: From the Big Bang to Black Holes (HAWKING)
	1280L	Black, Blue, and Gray: African Americans in the Civil War (HASKINS)
	1230L	Stiff: The Curious Lives of Human Cadavers (ROACH)
	1230L	Knowing Mandela: A Personal Portrait (CARLIN)
	1200L	The Dark Game: True Spy Stories (JANECZKO)

1100L–1195L

1150L *A Room of One's Own* WOOLF

The reason perhaps why we know so little of Shakespeare—compared with Donne or Ben Jonson or Milton—is that his grudges and spites and antipathies are hidden from us. We are not held up by some "revelation" which reminds us of the writer. All desire to protest, to preach, to proclaim an injury, to pay off a score, to make the world the witness of some hardship or grievance was fired out of him and consumed. Therefore his poetry flows from him free and unimpeded. If ever a human being got his work expressed completely, it was Shakespeare. If ever a mind was incandescent, unimpeded, I thought, turning again to the bookcase, it was Shakespeare's mind.

SAMPLE TITLES

LITERATURE	1180L	Sense and Sensibility (AUSTEN)
	1170L	The Amazing Adventure of Kavalier & Clay (CHABON)
	1150L	Great Expectations (DICKENS)
	1140L	Cold Mountain (FRAZIER)
	1130L	Democracy (DIDION)
INFORMATIONAL	1160L	The Longitude Prize (DASH)
	1160L	In Search of Our Mothers' Gardens (WALKER)
	1150L	The Human Microbiome: The Germs That Keep You Healthy (HIRSCH)
	1150L	In My Place (HUNTER-GAULT)
	1100L	Something to Declare (ALVAREZ)

1000L–1095L

1070L *Geeks: How Two Lost Boys Rode the Internet out of Idaho* KATZ

Geeks were the first to grasp just how much information was available on the Web, since they wrote the programs that put much of it there—movie times and reviews, bus and train schedules, news and opinions, catalogues, appliance instructions, plus, of course, software and its upgrades. And of course, music, the liberation of which is considered a seminal geek accomplishment.

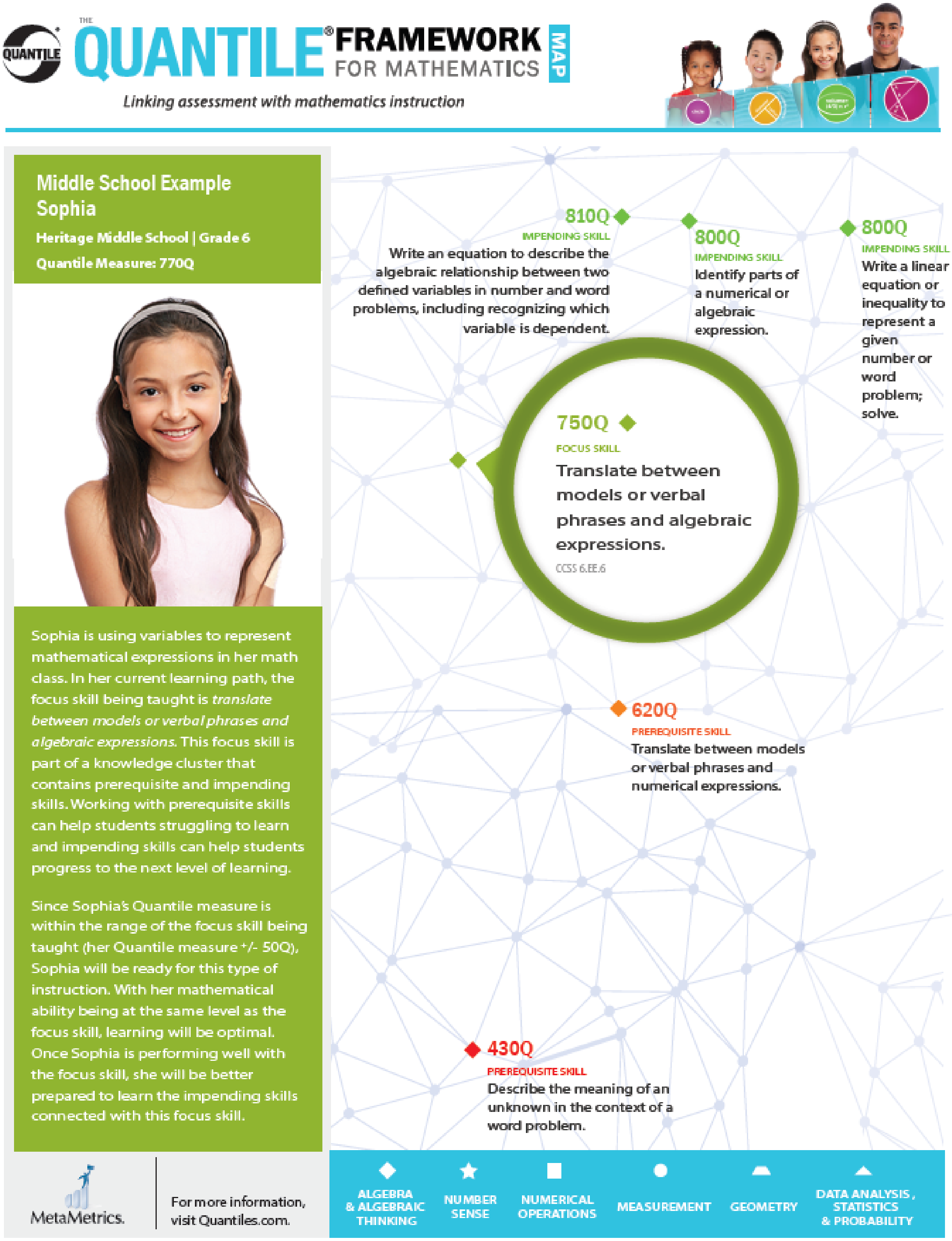
Virtually everything in a newspaper—and in many magazines—is now available online. In fact, some things, like the latest weather and breaking news, appear online hours before they hit print.

Yet while Jesse had gone through literally thousands of downloaded software applications, he'd never paid for any of them. He didn't even quite get the concept. The single cultural exception was books. Perhaps as a legacy of his childhood, Jesse remained an obsessive reader. He liked digging through the bins of used bookstores to buy sci-fi and classic literature; he liked books, holding them and turning their pages.

SAMPLE TITLES

LITERATURE	1080L	I Heard the Owl Call My Name (CRAVEN)
	1070L	Savvy (LAW)
	1070L	Around the World in 80 Days (VERNE)
	1010L	The Pearl (STEINBECK)
	1000L	The Hobbit or There and Back Again (TOLKIEN)
INFORMATIONAL	1030L	Phineas Gage: A Gruesome but True Story About Brain Science (FLEISCHMAN)
	1020L	This Land Was Made for You and Me: The Life and Songs of Woody Guthrie (PARTRIDGE)
	1010L	Travels With Charley: In Search of America (STEINBECK)
	1000L	Harriet Tubman: Conductor on the Underground Railroad (PETRY)
	1000L	Claudette Colvin: Twice Toward Justice (HOOSE)

EXHIBIT A Cont'd



THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

Middle School Example
Sophia

Heritage Middle School | Grade 6

Quantile Measure: 770Q

Sophia is using variables to represent mathematical expressions in her math class. In her current learning path, the focus skill being taught is *translate between models or verbal phrases and algebraic expressions.* This focus skill is part of a knowledge cluster that contains prerequisite and impending skills. Working with prerequisite skills can help students struggling to learn and impending skills can help students progress to the next level of learning.

Since Sophia's Quantile measure is within the range of the focus skill being taught (her Quantile measure +/- 50Q), Sophia will be ready for this type of instruction. With her mathematical ability being at the same level as the focus skill, learning will be optimal. Once Sophia is performing well with the focus skill, she will be better prepared to learn the impending skills connected with this focus skill.

810Q ◆
IMPENDING SKILL
Write an equation to describe the algebraic relationship between two defined variables in number and word problems, including recognizing which variable is dependent.

◆ 800Q
IMPENDING SKILL
Identify parts of a numerical or algebraic expression.

◆ 800Q
IMPENDING SKILL
Write a linear equation or inequality to represent a given number or word problem; solve.

750Q ◆
FOCUS SKILL
Translate between models or verbal phrases and algebraic expressions.
CCSS 6.EE.6

◆ 620Q
PREREQUISITE SKILL
Translate between models or verbal phrases and numerical expressions.

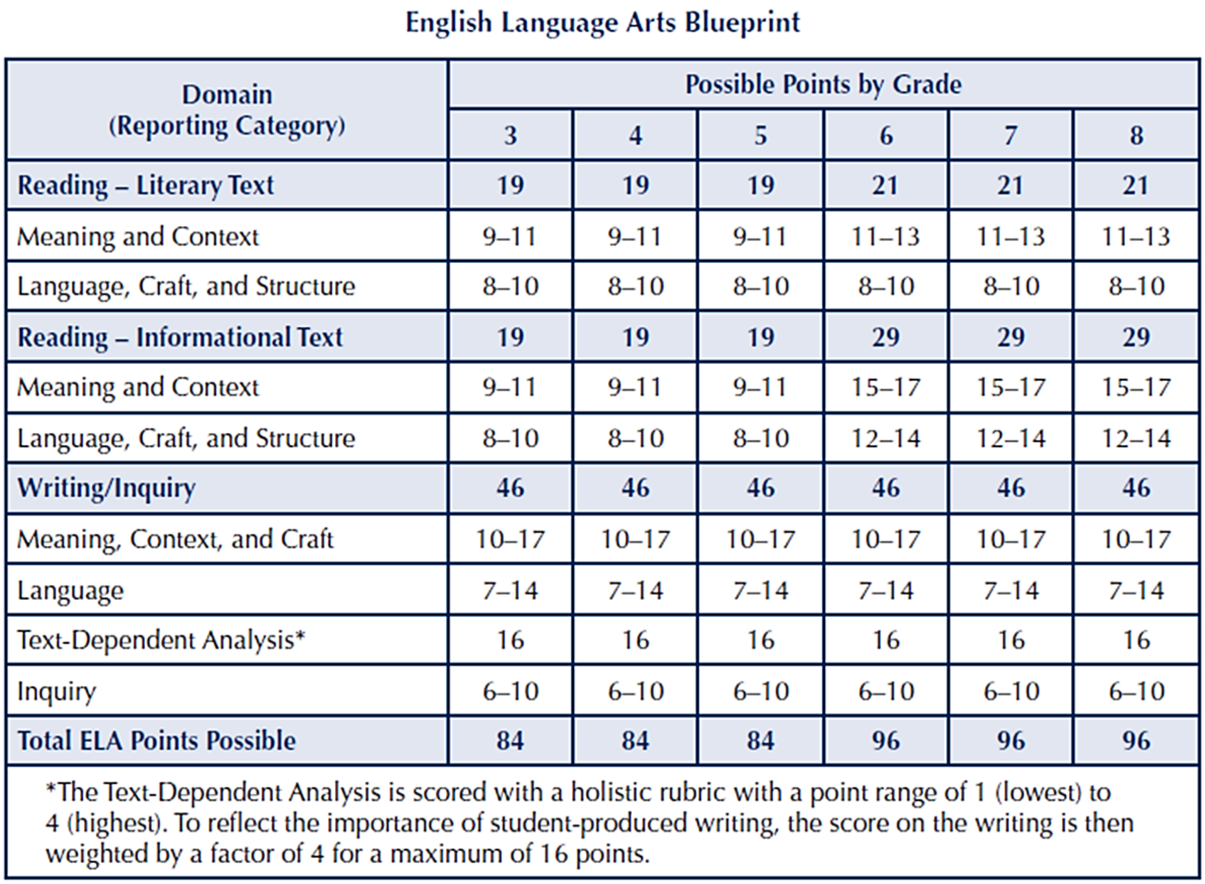
◆ 430Q
PREREQUISITE SKILL
Describe the meaning of an unknown in the context of a word problem.

MetaMetrics.

For more information, visit Quantiles.com.

◆ ALGEBRA & ALGEBRAIC THINKING
★ NUMBER SENSE
■ NUMERICAL OPERATIONS
● MEASUREMENT
◓ GEOMETRY
▲ DATA ANALYSIS, STATISTICS & PROBABILITY

EXHIBIT B
SC READY Test Blueprints

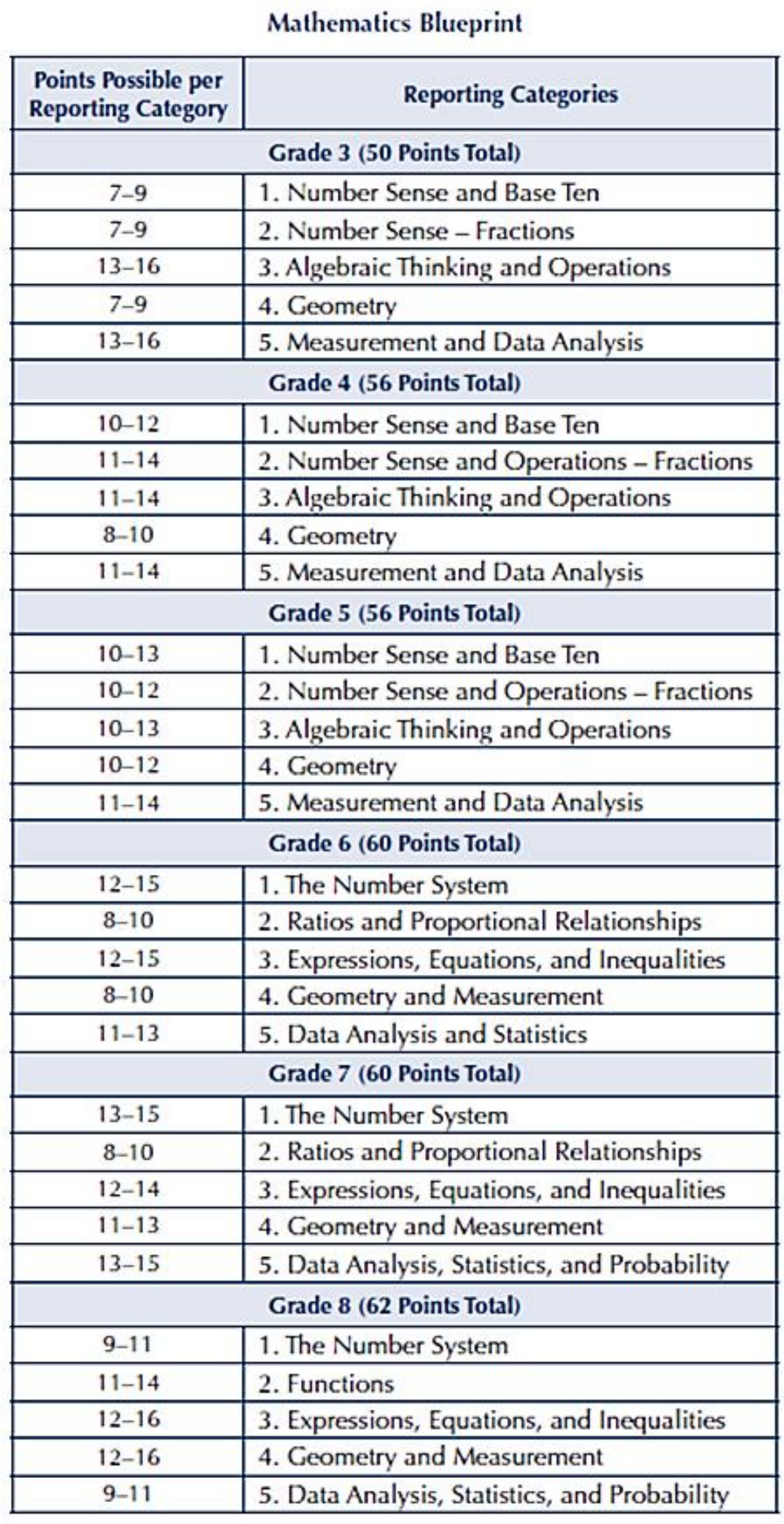


English Language Arts Blueprint

Domain (Reporting Category)	Possible Points by Grade					
	3	4	5	6	7	8
Reading – Literary Text	**19**	**19**	**19**	**21**	**21**	**21**
Meaning and Context	9–11	9–11	9–11	11–13	11–13	11–13
Language, Craft, and Structure	8–10	8–10	8–10	8–10	8–10	8–10
Reading – Informational Text	**19**	**19**	**19**	**29**	**29**	**29**
Meaning and Context	9–11	9–11	9–11	15–17	15–17	15–17
Language, Craft, and Structure	8–10	8–10	8–10	12–14	12–14	12–14
Writing/Inquiry	**46**	**46**	**46**	**46**	**46**	**46**
Meaning, Context, and Craft	10–17	10–17	10–17	10–17	10–17	10–17
Language	7–14	7–14	7–14	7–14	7–14	7–14
Text-Dependent Analysis*	16	16	16	16	16	16
Inquiry	6–10	6–10	6–10	6–10	6–10	6–10
Total ELA Points Possible	**84**	**84**	**84**	**96**	**96**	**96**

*The Text-Dependent Analysis is scored with a holistic rubric with a point range of 1 (lowest) to 4 (highest). To reflect the importance of student-produced writing, the score on the writing is then weighted by a factor of 4 for a maximum of 16 points.

EXHIBIT B Cont'd

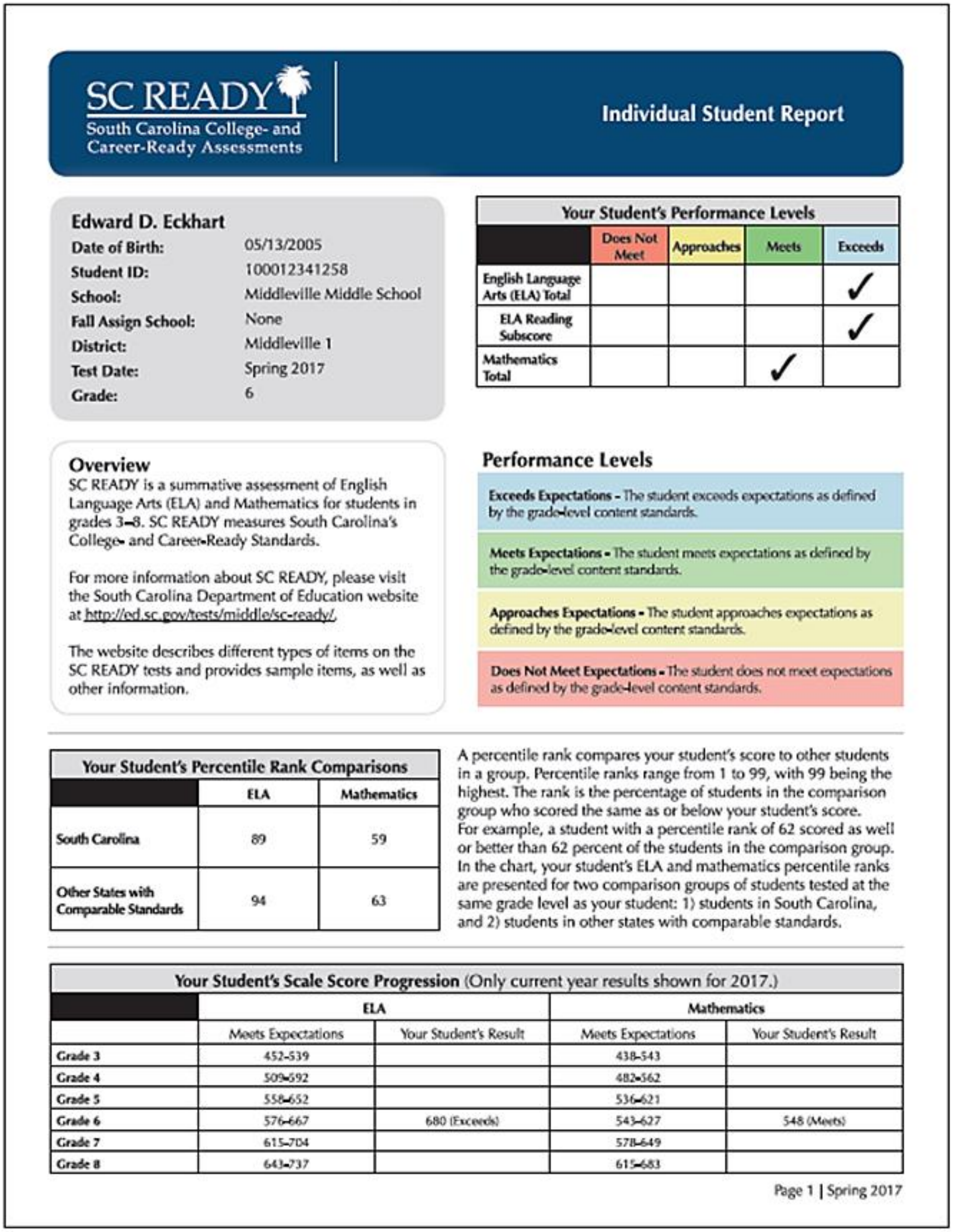


Mathematics Blueprint

Points Possible per Reporting Category	Reporting Categories
Grade 3 (50 Points Total)	
7–9	1. Number Sense and Base Ten
7–9	2. Number Sense – Fractions
13–16	3. Algebraic Thinking and Operations
7–9	4. Geometry
13–16	5. Measurement and Data Analysis
Grade 4 (56 Points Total)	
10–12	1. Number Sense and Base Ten
11–14	2. Number Sense and Operations – Fractions
11–14	3. Algebraic Thinking and Operations
8–10	4. Geometry
11–14	5. Measurement and Data Analysis
Grade 5 (56 Points Total)	
10–13	1. Number Sense and Base Ten
10–12	2. Number Sense and Operations – Fractions
10–13	3. Algebraic Thinking and Operations
10–12	4. Geometry
11–14	5. Measurement and Data Analysis
Grade 6 (60 Points Total)	
12–15	1. The Number System
8–10	2. Ratios and Proportional Relationships
12–15	3. Expressions, Equations, and Inequalities
8–10	4. Geometry and Measurement
11–13	5. Data Analysis and Statistics
Grade 7 (60 Points Total)	
13–15	1. The Number System
8–10	2. Ratios and Proportional Relationships
12–14	3. Expressions, Equations, and Inequalities
11–13	4. Geometry and Measurement
13–15	5. Data Analysis, Statistics, and Probability
Grade 8 (62 Points Total)	
9–11	1. The Number System
11–14	2. Functions
12–16	3. Expressions, Equations, and Inequalities
12–16	4. Geometry and Measurement
9–11	5. Data Analysis, Statistics, and Probability

EXHIBIT C
SC READY Sample Score Report

SAMPLE INDIVIDUAL STUDENT REPORT (GRADE 6 EXAMPLE—PAGE 1)



SC READY
South Carolina College- and Career-Ready Assessments

Individual Student Report

Edward D. Eckhart

Date of Birth:	05/13/2005
Student ID:	100012341258
School:	Middleville Middle School
Fall Assign School:	None
District:	Middleville 1
Test Date:	Spring 2017
Grade:	6

Your Student's Performance Levels

	Does Not Meet	Approaches	Meets	Exceeds
English Language Arts (ELA) Total				✓
ELA Reading Subscore				✓
Mathematics Total			✓	

Overview

SC READY is a summative assessment of English Language Arts (ELA) and Mathematics for students in grades 3–8. SC READY measures South Carolina's College- and Career-Ready Standards.

For more information about SC READY, please visit the South Carolina Department of Education website at http://ed.sc.gov/tests/middle/sc-ready/.

The website describes different types of items on the SC READY tests and provides sample items, as well as other information.

Performance Levels

Exceeds Expectations - The student exceeds expectations as defined by the grade-level content standards.

Meets Expectations - The student meets expectations as defined by the grade-level content standards.

Approaches Expectations - The student approaches expectations as defined by the grade-level content standards.

Does Not Meet Expectations - The student does not meet expectations as defined by the grade-level content standards.

Your Student's Percentile Rank Comparisons

	ELA	Mathematics
South Carolina	89	59
Other States with Comparable Standards	94	63

A percentile rank compares your student's score to other students in a group. Percentile ranks range from 1 to 99, with 99 being the highest. The rank is the percentage of students in the comparison group who scored the same as or below your student's score. For example, a student with a percentile rank of 62 scored as well or better than 62 percent of the students in the comparison group. In the chart, your student's ELA and mathematics percentile ranks are presented for two comparison groups of students tested at the same grade level as your student: 1) students in South Carolina, and 2) students in other states with comparable standards.

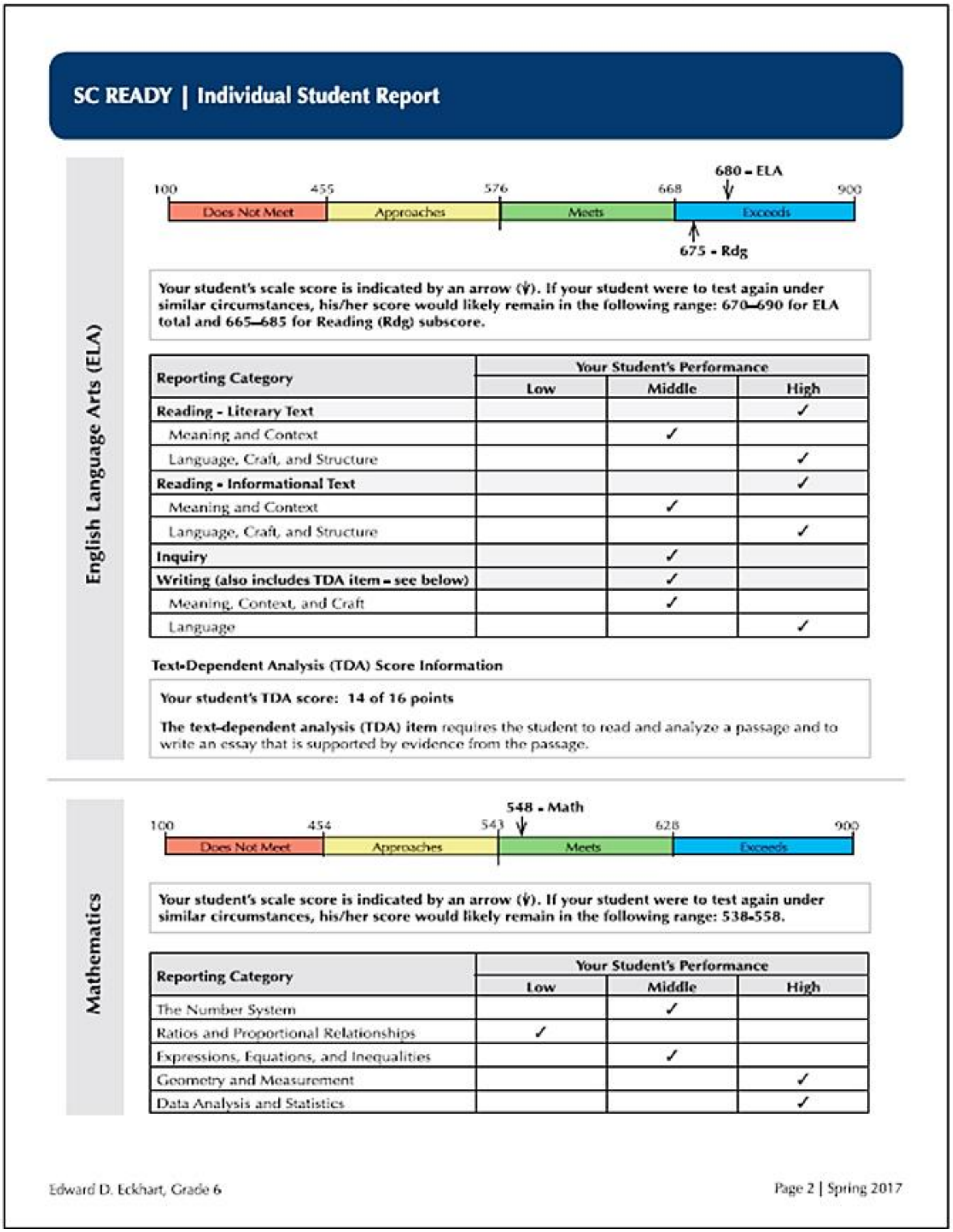
Your Student's Scale Score Progression (Only current year results shown for 2017.)

	ELA		Mathematics	
	Meets Expectations	Your Student's Result	Meets Expectations	Your Student's Result
Grade 3	452-539		438-543	
Grade 4	509-592		482-562	
Grade 5	558-652		536-621	
Grade 6	576-667	680 (Exceeds)	543-627	548 (Meets)
Grade 7	615-704		578-649	
Grade 8	643-737		615-683	

Page 1 | Spring 2017

EXHIBIT C Cont'd

SAMPLE INDIVIDUAL STUDENT REPORT (GRADE 6 EXAMPLE—PAGE 2)



SC READY | Individual Student Report

English Language Arts (ELA)

680 – ELA

100 | 455 | 576 | 668 | 900

Does Not Meet | Approaches | Meets | Exceeds

675 – Rdg

Your student's scale score is indicated by an arrow (↓). If your student were to test again under similar circumstances, his/her score would likely remain in the following range: 670—690 for ELA total and 665—685 for Reading (Rdg) subscore.

Reporting Category	Your Student's Performance		
	Low	Middle	High
Reading – Literary Text			✓
Meaning and Context		✓	
Language, Craft, and Structure			✓
Reading – Informational Text			✓
Meaning and Context		✓	
Language, Craft, and Structure			✓
Inquiry		✓	
Writing (also includes TDA item – see below)		✓	
Meaning, Context, and Craft		✓	
Language			✓

Text-Dependent Analysis (TDA) Score Information

Your student's TDA score: 14 of 16 points

The text-dependent analysis (TDA) item requires the student to read and analyze a passage and to write an essay that is supported by evidence from the passage.

Mathematics

548 – Math

100 | 454 | 543 | 628 | 900

Does Not Meet | Approaches | Meets | Exceeds

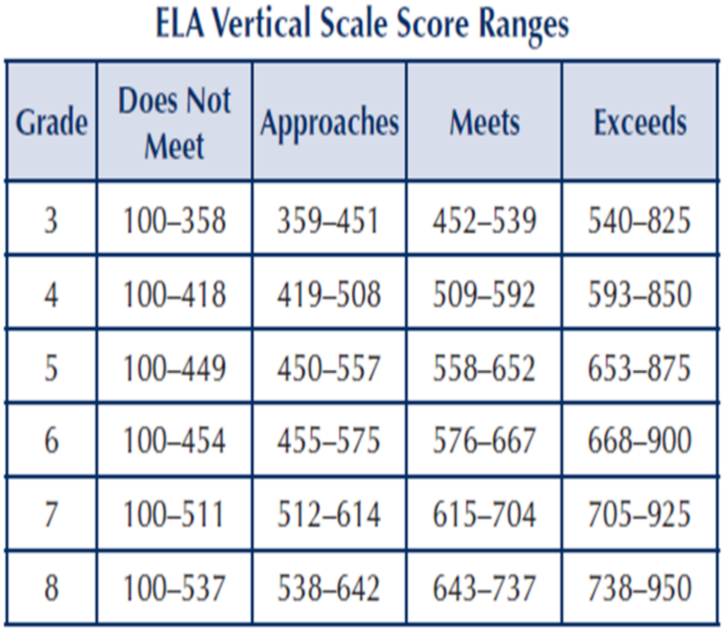
Your student's scale score is indicated by an arrow (↓). If your student were to test again under similar circumstances, his/her score would likely remain in the following range: 538–558.

Reporting Category	Your Student's Performance		
	Low	Middle	High
The Number System		✓	
Ratios and Proportional Relationships	✓		
Expressions, Equations, and Inequalities		✓	
Geometry and Measurement			✓
Data Analysis and Statistics			✓

Edward D. Eckhart, Grade 6

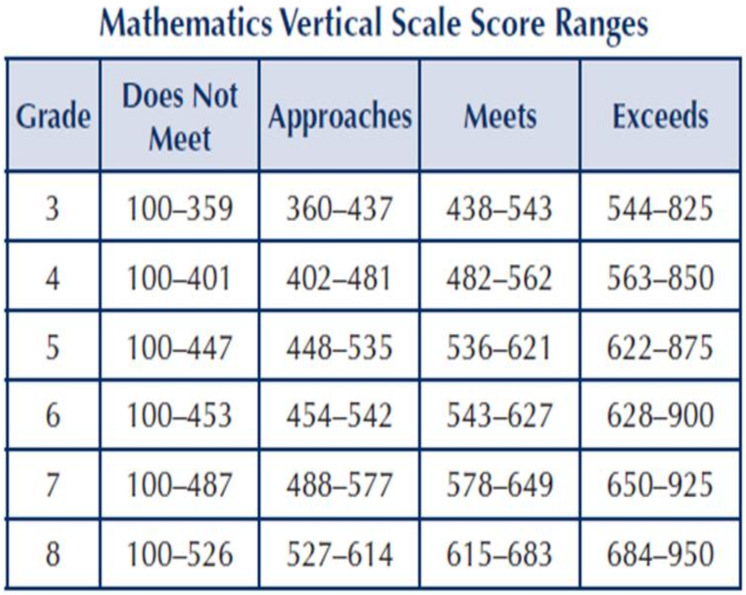
Page 2 | Spring 2017

EXHIBIT D
SC READY ELA and Mathematics Vertical Scale Score Ranges



ELA Vertical Scale Score Ranges

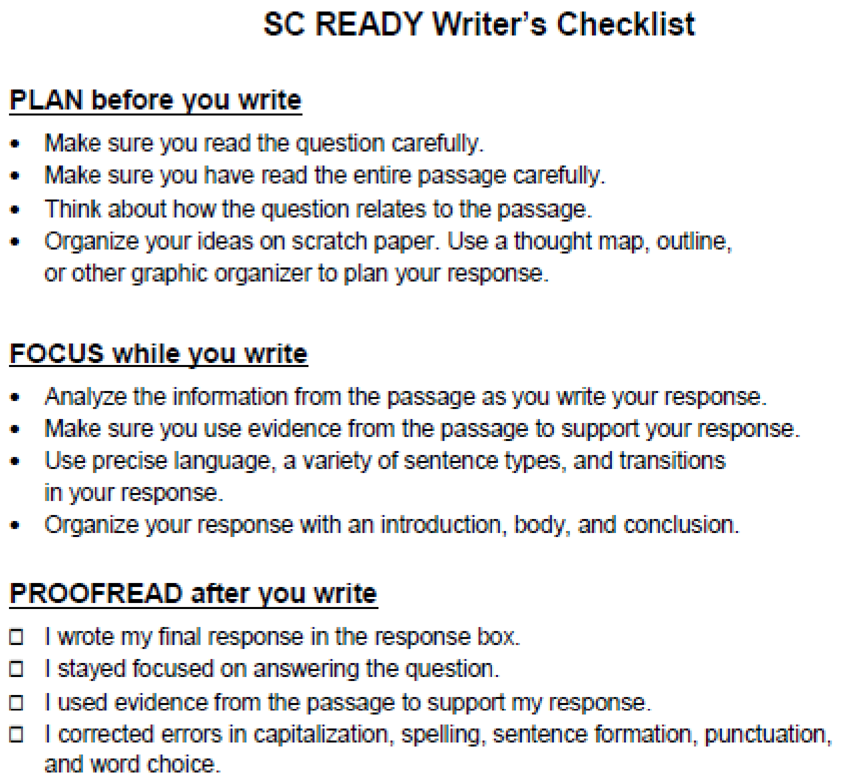
Grade	Does Not Meet	Approaches	Meets	Exceeds
3	100–358	359–451	452–539	540–825
4	100–418	419–508	509–592	593–850
5	100–449	450–557	558–652	653–875
6	100–454	455–575	576–667	668–900
7	100–511	512–614	615–704	705–925
8	100–537	538–642	643–737	738–950



Mathematics Vertical Scale Score Ranges

Grade	Does Not Meet	Approaches	Meets	Exceeds
3	100–359	360–437	438–543	544–825
4	100–401	402–481	482–562	563–850
5	100–447	448–535	536–621	622–875
6	100–453	454–542	543–627	628–900
7	100–487	488–577	578–649	650–925
8	100–526	527–614	615–683	684–950

EXHIBIT E
Writer's Checklist for TDA Essay Item



SC READY Writer's Checklist

PLAN before you write

- Make sure you read the question carefully.
- Make sure you have read the entire passage carefully.
- Think about how the question relates to the passage.
- Organize your ideas on scratch paper. Use a thought map, outline, or other graphic organizer to plan your response.

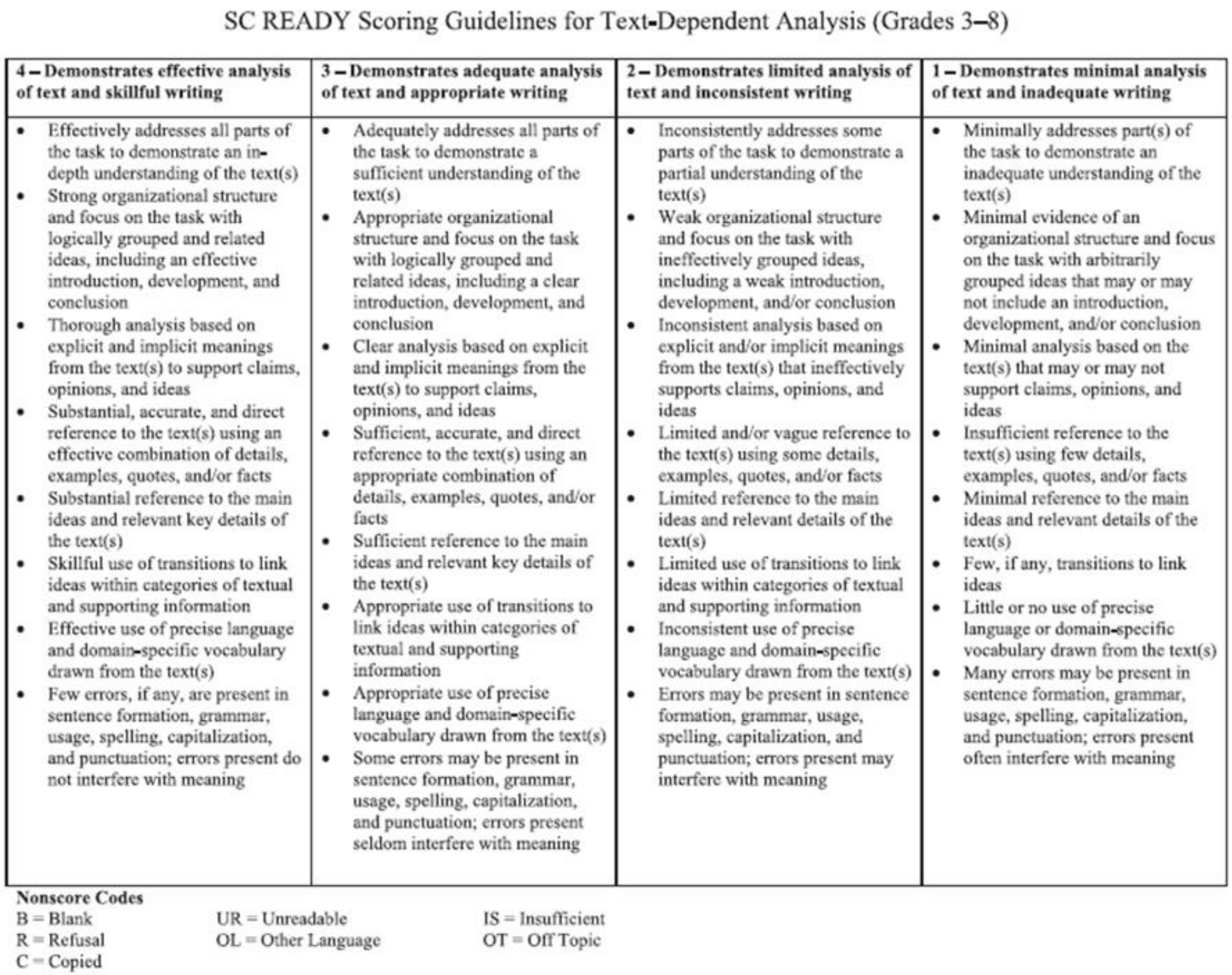
FOCUS while you write

- Analyze the information from the passage as you write your response.
- Make sure you use evidence from the passage to support your response.
- Use precise language, a variety of sentence types, and transitions in your response.
- Organize your response with an introduction, body, and conclusion.

PROOFREAD after you write

□ I wrote my final response in the response box.
□ I stayed focused on answering the question.
□ I used evidence from the passage to support my response.
□ I corrected errors in capitalization, spelling, sentence formation, punctuation, and word choice.

EXHIBIT F
SC READY TDA Item Scoring Rubric



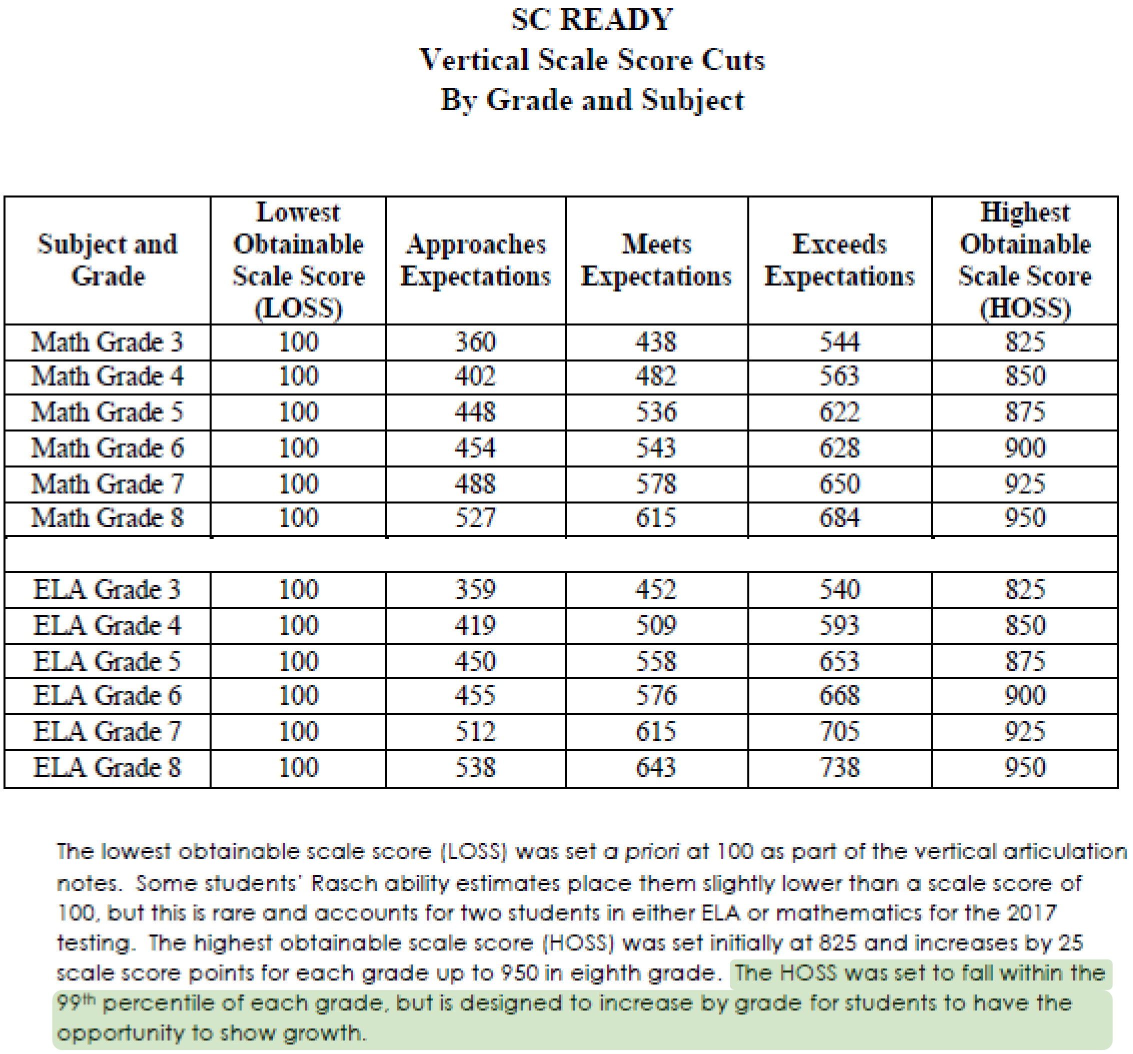
SC READY Scoring Guidelines for Text-Dependent Analysis (Grades 3–8)

4 – Demonstrates effective analysis of text and skillful writing	3 – Demonstrates adequate analysis of text and appropriate writing	2 – Demonstrates limited analysis of text and inconsistent writing	1 – Demonstrates minimal analysis of text and inadequate writing
• Effectively addresses all parts of the task to demonstrate an in-depth understanding of the text(s) • Strong organizational structure and focus on the task with logically grouped and related ideas, including an effective introduction, development, and conclusion • Thorough analysis based on explicit and implicit meanings from the text(s) to support claims, opinions, and ideas • Substantial, accurate, and direct reference to the text(s) using an effective combination of details, examples, quotes, and/or facts • Substantial reference to the main ideas and relevant key details of the text(s) • Skillful use of transitions to link ideas within categories of textual and supporting information • Effective use of precise language and domain-specific vocabulary drawn from the text(s) • Few errors, if any, are present in sentence formation, grammar, usage, spelling, capitalization, and punctuation; errors present do not interfere with meaning	• Adequately addresses all parts of the task to demonstrate a sufficient understanding of the text(s) • Appropriate organizational structure and focus on the task with logically grouped and related ideas, including a clear introduction, development, and conclusion • Clear analysis based on explicit and implicit meanings from the text(s) to support claims, opinions, and ideas • Sufficient, accurate, and direct reference to the text(s) using an appropriate combination of details, examples, quotes, and/or facts • Sufficient reference to the main ideas and relevant key details of the text(s) • Appropriate use of transitions to link ideas within categories of textual and supporting information • Appropriate use of precise language and domain-specific vocabulary drawn from the text(s) • Some errors may be present in sentence formation, grammar, usage, spelling, capitalization, and punctuation; errors present seldom interfere with meaning	• Inconsistently addresses some parts of the task to demonstrate a partial understanding of the text(s) • Weak organizational structure and focus on the task with ineffectively grouped ideas, including a weak introduction, development, and/or conclusion • Inconsistent analysis based on explicit and/or implicit meanings from the text(s) that ineffectively supports claims, opinions, and ideas • Limited and/or vague reference to the text(s) using some details, examples, quotes, and/or facts • Limited reference to the main ideas and relevant details of the text(s) • Limited use of transitions to link ideas within categories of textual and supporting information • Inconsistent use of precise language and domain-specific vocabulary drawn from the text(s) • Errors may be present in sentence formation, grammar, usage, spelling, capitalization, and punctuation; errors present may interfere with meaning	• Minimally addresses part(s) of the task to demonstrate an inadequate understanding of the text(s) • Minimal evidence of an organizational structure and focus on the task with arbitrarily grouped ideas that may or may not include an introduction, development, and/or conclusion • Minimal analysis based on the text(s) that may or may not support claims, opinions, and ideas • Insufficient reference to the text(s) using few details, examples, quotes, and/or facts • Minimal reference to the main ideas and relevant details of the text(s) • Few, if any, transitions to link ideas • Little or no use of precise language or domain-specific vocabulary drawn from the text(s) • Many errors may be present in sentence formation, grammar, usage, spelling, capitalization, and punctuation; errors present often interfere with meaning

Nonscore Codes

B = Blank	UR = Unreadable	IS = Insufficient
R = Refusal	OL = Other Language	OT = Off Topic
C = Copied		

EXHIBIT G
SC READY Vertical Scale Minimums, Maximums and Cut Scores



SC READY
Vertical Scale Score Cuts
By Grade and Subject

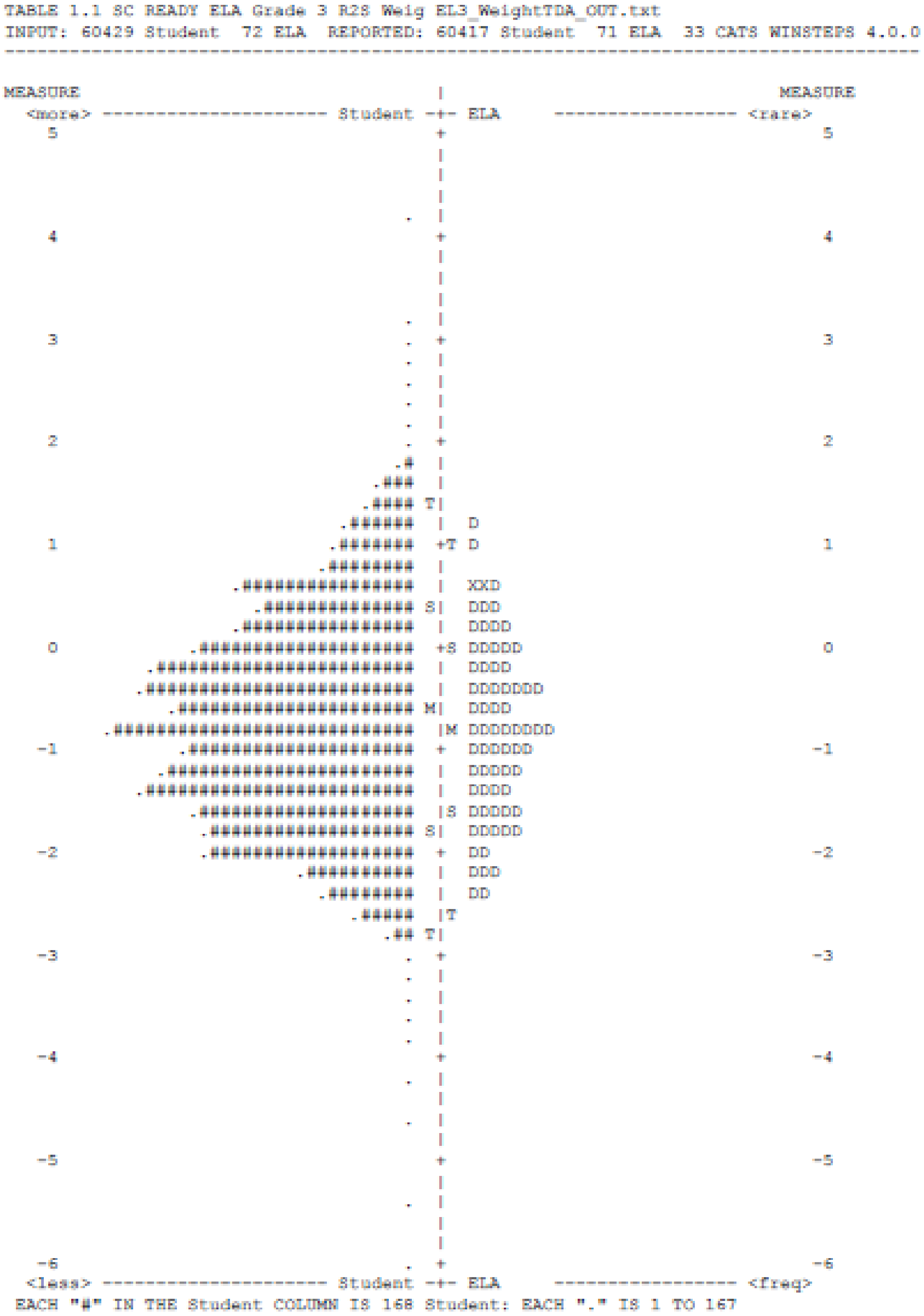
Subject and Grade	Lowest Obtainable Scale Score (LOSS)	Approaches Expectations	Meets Expectations	Exceeds Expectations	Highest Obtainable Scale Score (HOSS)
Math Grade 3	100	360	438	544	825
Math Grade 4	100	402	482	563	850
Math Grade 5	100	448	536	622	875
Math Grade 6	100	454	543	628	900
Math Grade 7	100	488	578	650	925
Math Grade 8	100	527	615	684	950
ELA Grade 3	100	359	452	540	825
ELA Grade 4	100	419	509	593	850
ELA Grade 5	100	450	558	653	875
ELA Grade 6	100	455	576	668	900
ELA Grade 7	100	512	615	705	925
ELA Grade 8	100	538	643	738	950

The lowest obtainable scale score (LOSS) was set a *priori* at 100 as part of the vertical articulation notes. Some students' Rasch ability estimates place them slightly lower than a scale score of 100, but this is rare and accounts for two students in either ELA or mathematics for the 2017 testing. The highest obtainable scale score (HOSS) was set initially at 825 and increases by 25 scale score points for each grade up to 950 in eighth grade. The HOSS was set to fall within the 99th percentile of each grade, but is designed to increase by grade for students to have the opportunity to show growth.

Source: SC READY Scale Score Cuts with LOSS and HOSS_101617.pdf

EXHIBIT H
SC READY Rasch Item Maps

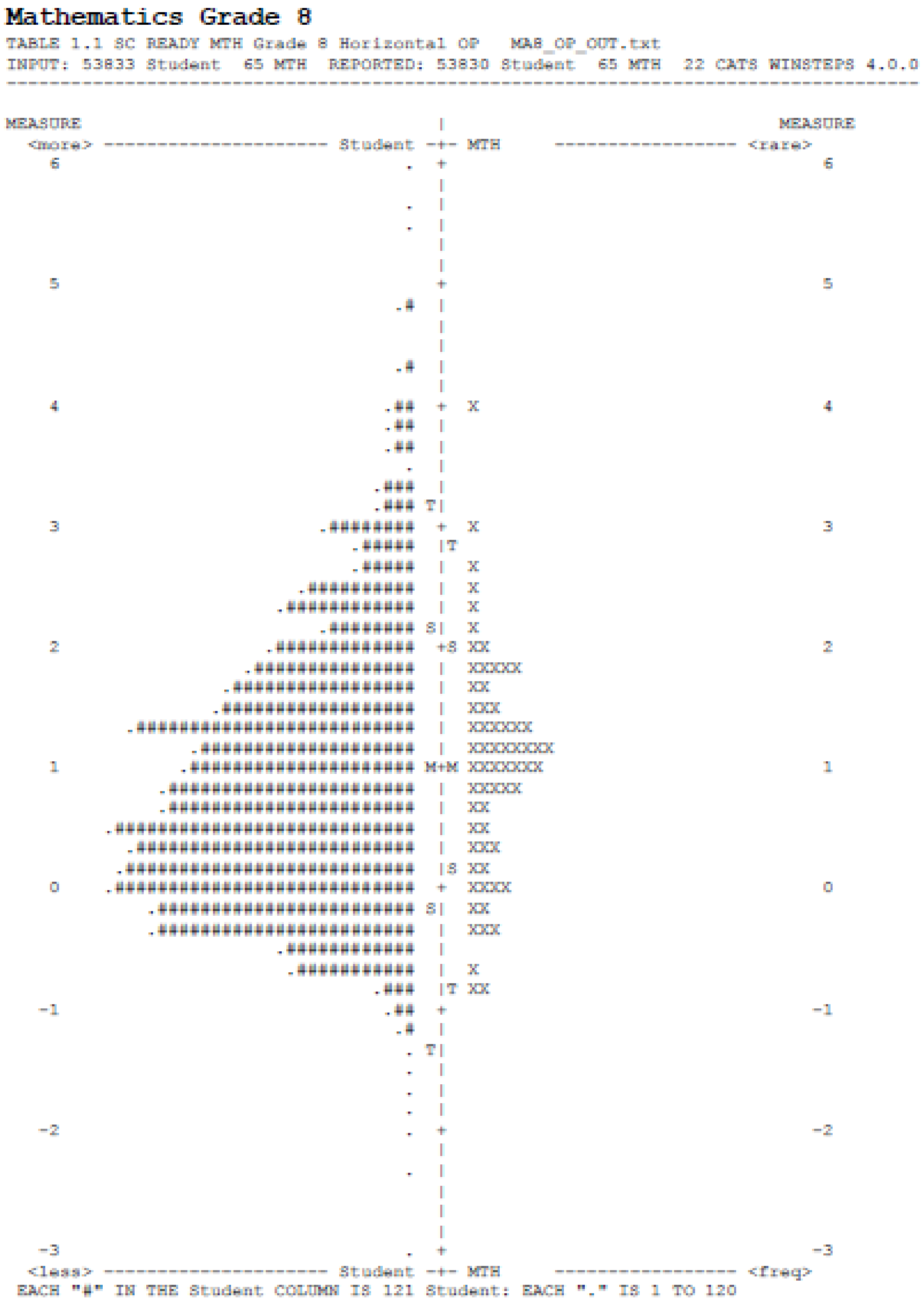
ELA Grade 3



```
TABLE 1.1 SC READY ELA Grade 3 R2S Weig EL3_WeightTDA_OUT.txt
INPUT: 60429 Student  72 ELA  REPORTED: 60417 Student  71 ELA  33 CATS WINSTEPS 4.0.0
-------------------------------------------------------------------------------------

MEASURE                                 |                               MEASURE
  <more> --------------------- Student -+- ELA     ----------------- <rare>
    5                                   +                                   5
                                        |
                                        |
                                        |
                                     .  |
    4                                   +                                   4
                                        |
                                        |
                                        |
                                     .  |
    3                                .  +                                   3
                                     .  |
                                     .  |
                                     .  |
                                     .  |
    2                                .  +                                   2
                                    .#  |
                                  .###  |
                                 .#### T|
                               .######  |  D
    1                         .#######  +T D                                1
                             .########  |
                     .################  |  XXD
                       .############## S|  DDD
                     .################  |  DDDD
    0            .####################  +S DDDDD                            0
             .########################  |  DDDD
            .#########################  |  DDDDDDD
               .###################### M|  DDDD
         .############################  |M DDDDDDDD
   -1           .#####################  +  DDDDDD                          -1
              .#######################  |  DDDDD
            .#########################  |  DDDD
                 .####################  |S DDDDD
                  .################### S|  DDDDD
   -2             .###################  +  DD                              -2
                           .##########  |  DDD
                             .########  |  DD
                                .#####  |T
                                   .## T|
   -3                                .  +                                  -3
                                     .  |
                                     .  |
                                     .  |
                                     .  |
   -4                                   +                                  -4
                                     .  |
                                        |
                                     .  |
                                        |
   -5                                   +                                  -5
                                        |
                                     .  |
                                        |
                                        |
   -6                                .  +                                  -6
  <less> --------------------- Student -+- ELA     ----------------- <freq>
 EACH "#" IN THE Student COLUMN IS 168 Student: EACH "." IS 1 TO 167
```

EXHIBIT H Cont'd



Mathematics Grade 8

```
TABLE 1.1 SC READY MTH Grade 8 Horizontal OP   MA8_OP_OUT.txt
INPUT: 53833 Student  65 MTH  REPORTED: 53830 Student  65 MTH  22 CATS WINSTEPS 4.0.0
-------------------------------------------------------------------------------------

MEASURE                                  |                               MEASURE
  <more> ---------------------- Student -+- MTH     ----------------- <rare>
    6                                 .  +                                   6
                                         |
                                      .  |
                                      .  |
                                         |
                                         |
    5                                    +                                   5
                                     .#  |
                                         |
                                         |
                                     .#  |
                                         |
    4                               .##  +  X                                4
                                    .##  |
                                    .##  |
                                      .  |
                                   .###  |
                                   .### T|
    3                         .########  +  X                                3
                                .#####  |T
                                .#####  |  X
                            .##########  |  X
                          .############  |  X
                              .######## S|  X
    2                    .#############  +S XX                               2
                       .###############  |  XXXXX
                     .#################  |  XX
                    .##################  |  XXX
            .##########################  |  XXXXXX
                  .####################  |  XXXXXXXX
    1            .##################### M+M XXXXXXX                          1
               .#######################  |  XXXXX
               .#######################  |  XX
          .############################  |  XX
            .##########################  |  XXX
           .###########################  |S XX
    0     .############################  +  XXXX                             0
              .######################## S|  XX
              .########################  |  XXX
                          .############  |
                           .###########  |  X
                                   .###  |T XX
   -1                               .##  +                                  -1
                                     .#  |
                                      . T|
                                      .  |
                                      .  |
                                      .  |
   -2                                 .  +                                  -2
                                         |
                                      .  |
                                         |
                                         |
                                         |
   -3                                 .  +                                  -3
  <less> ---------------------- Student -+- MTH     ----------------- <freq>
 EACH "#" IN THE Student COLUMN IS 121 Student: EACH "." IS 1 TO 120
```