South Carolina
College- and Career-Ready Assessments
2022 Technical Report

Produced for the
South Carolina Department of Education

By

Data Recognition Corporation

Table of Contents

List of Tables & Figures

Executive Summary

This report is a technical summary of the 2022 administration of the South Carolina College- and Career-Ready Assessments (SC READY) administered in grades three through eight in English language arts (ELA) and mathematics.

The SC READY assessments are grade-level tests that are designed to measure students' knowledge of ELA and mathematics and are aligned with the *South Carolina College- and Career-Ready Standards for ELA and Mathematics* (or the *SC CCR Standards for ELA and Mathematics*). There are four performance levels in ELA and mathematics: *Does Not Meet Expectations*, *Approaches Expectations*, *Meets Expectations*, and *Exceeds Expectations*.

The spring 2022 test forms were developed by Data Recognition Corporation (DRC) and directly aligned to the appropriate standards under the supervision of the South Carolina Department of Education (SCDE). All assessments were available in online and paper-and-pencil formats.

E.1 Overview of the SC READY

SC READY consists of statewide assessments in ELA and mathematics. The SC READY assessments are designed to meet all requirements of the Education Accountability Act, the Elementary and Secondary Education Act (ESEA), the Individuals with Disabilities Education Improvement Act (IDEA), and the U.S. Department of Education Peer Review of State Assessment Systems: Non-regulatory Guidance for States [Peer Review Guidance].

All students in grades three through eight are required to take the SC READY assessments except those students with significant cognitive disabilities who qualify to participate in the Alternate Assessment for students with significant cognitive disabilities.

The initial administration of the SC READY assessments was in spring 2016. Results from the test are used for federal and state accountability purposes. The SC READY assessments are available in both online and paper-and-pencil formats.

E.2 Cancellation of the Spring 2020 Assessments

South Carolina Governor Henry McMaster formally issued Executive Order 2020-09, which ordered the closing of schools in response to COVID-19. On March 17, 2020, a federal assessment waiver application to cancel testing for spring 2020 was submitted and approved.

E.3 Administration

In spring 2022, South Carolina administered summative assessments in ELA and mathematics to students in grades three through eight. The SC READY assessments were administered from April 7 to June 6, 2022. Test administration is discussed in Section 2 of this report.

The spring 2022 assessment results should be interpreted in the context of circumstances related to the COVID-19 pandemic, including school closures, nonstandard instruction delivery modes in the 2020–21 school year, potential diminished opportunity to learn for students, and other unknown effects of the pandemic on students and their families.

E.4 Student Performance

Tables E.1 and E.2 present the percentage of students classified as *Meets Expectations* or *Exceeds Expectations* in 2022 for ELA and mathematics, respectively. The percentage of students classified as *Meets Expectations* or *Exceeds Expectations* tends to be higher in the lower grade levels. More details on student performance are provided in Section 6 of this report.

Table E.1. State-Level Percentages of Students in Each Performance Level for ELA

Grade	N	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	55,905	29.0	23.0	23.6	24.4	48.0
4	56,397	28.0	21.6	19.8	30.6	50.4
5	57,065	22.1	29.7	27.9	20.4	48.3
6	57,566	23.1	31.9	25.4	19.7	45.1
7	60,109	26.9	29.8	23.1	20.1	43.2
8	60,920	27.2	26.9	29.2	16.7	45.9

Table E.2. State-Level Percentages of Students in Each Performance Level for Mathematics

Grade	N	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	55,896	26.8	22.2	26.6	24.4	51.0
4	56,404	28.7	27.9	19.6	23.7	43.4
5	57,066	29.5	27.2	23.6	19.7	43.3
6	57,662	33.8	30.5	19.3	16.5	35.7
7	60,155	35.8	33.3	15.3	15.5	30.9
8	60,987	40.8	29.0	14.1	16.0	30.2

E.5 Validity of Intended Interpretation of Test Scores

Most sections of this technical report are designed to provide validity evidence to support the use and intended interpretation of the SC READY ELA and mathematics test scores. The SC READY scores are used to identify strengths and areas for improvement in South Carolina's

student performance; to inform stakeholders (teachers, school administrators, district administrators, SCDE staff members, parents, and the public) about the status of the progress toward meeting the academic performance standards of the state; and to meet the requirements of the state's accountability program. Section 3 of this technical report provides the outline and overview of the validity framework and a summary of the validity evidence for the SC READY assessments.

Evidence of validity based on test content area was supported by the test specifications, including the test design and test blueprint. The SC READY assessments were developed in alignment with the appropriate content area standards from the *SC CCR Standards for ELA and Mathematics*. A rigorous item review and test form development process was implemented to select items from DRC's development and the prior year's field-testing. More details on test content area and test development are provided in Section 2 of this report.

The SC READY assessments were available in online and paper-and-pencil formats, as well as Braille, Large Print, and accommodated paper-based forms. The SC READY assessments were administered in a standardized manner, further supporting the validity of the intended score interpretation. Universal tools were available for all students to use, and accommodations were available to students for whom such aids were deemed appropriate and indicated in their Individualized Education Programs (IEPs). More details on test administration and the use of accommodations or universal tools are provided in Section 5 of this report.

Scoring of Technology-enhanced (TE) item types, such as drag and drop, hot spot (click to select), drop-down list, and constructed-response (type a numeric answer) items followed predefined scoring criteria. ELA assessments contained one handscored, Text-Dependent Analysis (TDA) item per grade. The multiple-choice, multi-select, and TE items were autoscored (details are included in Section 4.5 for details).

The test scaling and equating was conducted using item response theory (IRT) methodology. Students' scale scores were derived using item parameters using a pre-equated model for the spring 2022 administration. The IRT models used for SC READY test scaling were appropriate for the test data supporting the operational data analysis and ensuring that the test items, as well as the overall tests, were functioning appropriately. Details on test scaling and equating are included in Section 4.7. The cut scores used to classify students into different achievement levels and associated achievement level descriptors were established during the summer 2016 achievement level setting for ELA and mathematics. Performance level setting for both ELA and mathematics was performed in a collaborative and participatory process, further supporting the validity and interpretation of the SC READY scores (details are included in Section 6.1).

Evidence of construct-related validity—supporting the intended interpretation of test scores and their use—was provided through studies of test reliability, evaluation of internal test structure, and evaluation of the relationship of test scores with external variables. The reliability analysis results indicated that the SC READY tests produce scores that would be relatively stable if the tests were administered repeatedly under similar conditions. The assumption that the content area SC READY tests were unidimensional (that is, that each

grade-level test measured one primary dimension) was confirmed through principal component analysis. The evidence of the validity of the intended interpretation of the SC READY test scores based on the relationships with other variables was evaluated through the correlations computed between SC READY and other South Carolina state assessments. The student scores were found to be highly, but not perfectly, related to each other, suggesting that while different constructs are being measured, the two assessments may also be tapping into a similar knowledge base or general underlying ability. In addition, test fairness was evaluated through differential item functioning (DIF) analysis and analysis of differences in test performance among subgroups (details are included in Section 4).

Section 1—History of Statewide System of Standards & Assessments

1.1 The South Carolina Education Accountability Act

The Education Accountability Act of 1998, Chapter 18 of Title 59 of the 1976 South Carolina Code of Laws, provided for the establishment of a performance-based accountability system. The State Board of Education was required to develop a statewide assessment program to measure student performance on state standards. The Palmetto Achievement Challenge Tests (PACT) assessment program was developed in accordance with this legislation. English language arts (ELA) and mathematics tests were administered the first time in April 1999 to all students in grades three through eight. In 2001 and 2002, ELA and mathematics field tests, respectively, were added to the statewide program. In 2003, PACT included operational tests for all four subject areas—ELA, mathematics, science, and social studies—for all students in grades three through eight. Effective with the 2007 administration, only students in grades four and seven were administered both the science and the social studies tests. Students in grades three, five, six, and eight were randomly assigned to take either the science or the social studies test. All students in grades three through eight participated in the ELA and mathematics tests. The last administration of the PACT was in spring 2008.

1.2 The Education Accountability Act of 2008

On May 29, 2008, the South Carolina General Assembly ratified a bill to amend the Education Accountability Act. This bill revised the manner in which students, schools, and districts are assessed and how school academic performance is designated. The General Assembly noted that the PACT "no longer meets the requirements" of the amended legislation.

As stated in Section 59-18-100, the purpose of the revised Education Accountability Act legislation is "to establish a performance-based accountability system for public education which focuses on improving teaching and learning so that students are equipped with a strong academic foundation." The legislation in Section 59-18-100 states in part that the accountability system must do the following:

(1) use academic achievement standards to push schools and students toward higher performance by aligning the state assessment to those standards and linking policies and criteria for performance standards, accreditation, reporting, school rewards, and targeted assistance;
(2) provide an annual report card with a performance indicator system that is logical, reasonable, fair, challenging, and technically defensible, which furnishes clear and specific information about school and district academic performance and other performance to parents and the public;
(3) require all districts to establish local accountability systems to stimulate quality teaching and learning practices and target assistance to low performing schools;

(4) provide resources to strengthen the process of teaching and learning in the classroom to improve student performance and reduce gaps in performance;
(5) support professional development as integral to improvement and to the actual work of teachers and school staff; and
(6) expand the ability to evaluate the system to conduct in-depth studies on implementation, efficiency, and the effectiveness of academic improvement efforts.

Section 59-18-310 of the Education Accountability Act requires the Department of Education to develop or adopt a statewide assessment program to promote student learning and to measure student performance on state standards and do the following:

(1) identify areas in which students, schools, or school districts need additional support;
(2) indicate the academic achievement for schools, districts, and the State;
(3) satisfy federal reporting requirements; and
(4) provide professional development to educators.

The Education Accountability Act also mandated a standards-based assessment in which "an individual's performance is compared to specific performance standards and not to the performance of other students." The new assessment must be an "objective and reliable statewide assessment" meaning that the assessment yields "consistent results" and measures "the cognitive knowledge and skills specified in the state-approved academic standards." According to this legislation, a student's score on this assessment may not be the sole criterion for placing the student on academic probation, retaining the student in his current grade, or requiring the student to attend summer school.

In accordance with this legislation, the Palmetto Assessment of State Standards (SCPASS) was developed to achieve these goals. Beginning with the 2008–09 school year, SCPASS test results were used for school, district, and state accountability purposes.

1.3 Act 200

In 2014, Act No. 200 amended Section 59-18-325 of the 1976 Code by adding a "C" subsection to the law. The Act directs the procurement of grades three through eight ELA and mathematics assessments for the 2014–15 school year. The parameters for the statewide assessment system are outlined in this section; excerpts are provided below.

> (C) To maintain a comprehensive and cohesive assessment system that signals a student's preparedness for the next educational level and ultimately culminates in a clear indication of a student's preparedness for postsecondary success in a college or career and to satisfy federal and state accountability purposes, the State Department of Education shall procure and maintain a summative assessment system.

(1) 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 assessment or 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) be a vertically scaled, benchmarked, standards-based system of summative assessments;
(c) measures a student's preparedness for the next level of their educational matriculation and individual student performance against the state standards in English/language arts, reading, writing, 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) be administered to all students in a computer-based format except for students with disabilities as specified in the student's IEP or 504 Accommodation Plan, and unless the use of a computer by these students is prohibited due to the vendor's restrictions on computer-based test security, in which case the paper version must be made available; and
(h) assists school districts and schools in aligning assessment, curriculum, and instruction.

The assessments must meet the definition of a high-quality assessment provided in the 2012 United States Department of Education ESEA Flexibility document (which can be found at http://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc). The extended definition of a "high-quality assessment" is given in the excerpt below.

High-Quality Assessment: A 'high-quality assessment' is an assessment or a system of assessments that is valid, reliable, and fair for its intended

purposes; and measures student knowledge and skills against college- and career-ready standards in a way that—

- covers the full range of those standards, including standards against which student achievement has traditionally been difficult to measure;
- as appropriate, elicits complex student demonstrations or applications of knowledge and skills;
- provides an accurate measure of student achievement across the full performance continuum, including for high- and low-achieving students;
- provides an accurate measure of student growth over a full academic year or course;
- produces student achievement data and student growth data that can be used to determine whether individual students are college and career ready or on track to being college and career ready;
- assesses all students, including English Learners and students with disabilities;
- provides for alternate assessments based on grade-level academic achievement standards or alternate assessments based on alternate academic achievement standards for students with the most significant cognitive disabilities, consistent with 34 C.F.R. § 200.6(a)(2); and
- produces data, including student achievement data and student growth data, that can be used to inform: determinations of school effectiveness for purposes of accountability under Title I; determinations of individual principal and teacher effectiveness for purposes of evaluation; determinations of principal and teacher professional development and support needs; and teaching, learning, and program improvement.

1.4 Groups Involved with the SC READY

The SCDE developed the SC READY assessments both directly and through private contractors. In addition, the SCDE has managed the yearly administration of the SC READY and disseminated the results to the schools and to the public.

1.4.1 Education Oversight Committee

The Education Oversight Committee (EOC) was established through Section 56-6-10 of the South Carolina Code of Laws. According to the mandate of the Education Accountability Act of 1998, "the Education Oversight Committee… will review the state assessment program and the course assessments for alignment with the state standards, level of difficulty and validity, and for the ability to differentiate levels of achievement, and will make recommendations for needed changes, if any" (S.C. Code Ann. § 59-18-320(A)). The EOC is composed of eighteen

members from state government, business, and education. The EOC was charged to set achievement standards for the SC READY assessments.

1.4.2 Technical Advisory Committee

The Technical Advisory Committee (TAC) makes recommendations to the SCDE on issues regarding field test design, item analysis, linking issues, the item response theory (IRT) model for data analysis, procedures for standard setting and data reporting, and other relevant psychometric issues. Experts from national, state, and local organizations are included in the membership of the TAC.

1.4.3 Contractors and Other Groups

In addition to SCDE staff members, contractors and SC educators were involved in SC READY development and administration. DRC was contracted to provide test administration, scoring, and reporting services. Under the current contract, DRC also develops SC READY items and test forms.

1.5 South Carolina Academic Standards and Indicators

South Carolina academic standards consist of statements indicating the most important and consensually determined expectations for student learning in a particular discipline. They indicate what schools are expected to teach and what students are expected to learn. In accordance with the Education Accountability Act, the purpose of academic standards is to provide the basis for the development of local curricula and statewide assessments. Further, the standards are to promote the goals of providing every student with the competencies to do the following:

(1) read, view, and listen to complex information in the English language;
(2) write and speak effectively in the English language;
(3) solve problems by applying mathematics;
(4) conduct research and communicate findings;
(5) understand and apply scientific concepts;
(6) obtain a working knowledge of world, United States, and South Carolina history, government, economics, and geography; and
(7) use information to make decisions.

As emphasized by the Education Accountability Act (S.C. Code Ann. § 59-18), the standards

> must be reflective of the highest level of academic skills with the rigor necessary to improve the curriculum and instruction in South Carolina's schools so that students are encouraged to learn at unprecedented levels and must be reflective of the highest level of academic skills at each grade-level.

The South Carolina standards also include multiple indicators for each standard. Indicators are specific statements of the cognitive processes and the content area knowledge and skills that

students must demonstrate in order to meet the standard. The main verb in each indicator specifies the particular aspect of the particular cognitive processes that are described in the revised Bloom's taxonomy (Anderson, Lorin, & Krathwohl, 2001). Use of the taxonomic verbs will allow teachers to identify the kind of knowledge addressed by an indicator and therefore enable them to teach the content area in an effective manner.

The following is an example of a standard and an indicator for ELA.

Grade 4—Standard 1: The student will read and comprehend a variety of literary texts in print and non-print formats.
Indicator 1.2: Analyze literary texts to draw conclusions and make inferences.

The academic standards for each subject are not presented in an instructional sequence. All the standards and their indicators carry equal weight and should be taught in an integrated manner. Links to the academic standards and supporting documents can be found on the "Standards" page of the SCDE website (http://ed.sc.gov/instruction/standards-learning/).

The South Carolina academic standards are reviewed on a cyclical basis using procedures agreed upon by the SCDE and the Education Oversight Committee (EOC). Procedures for the review of all newly revised South Carolina academic standards are published in the document *Procedures for the Cyclical Review of Current South Carolina K–12 Academic Standards and for the Development of New Academic Standards*.

1.6 Alignment of South Carolina Standards with other Standards

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, and the Third International Mathematics and Science Standards. More specifically, resources used by each subject are provided in the following paragraphs.

1.6.1 ELA

In the fall of 2014, the SCDE developed new ELA standards. This process began with a review of a number of resources including the following:

- the 2014 ACT College and Career Readiness Standards;
- the Common Core State Standards for English Language Arts (CCSS);
- college- and career-ready standards from other states including Indiana, Nebraska, and Texas;
- the *South Carolina Academic Standards for English Language Arts 2008*, which reference the 2001 Massachusetts standards;
- test specifications for the SAT; and

- the National Council of Teachers of English/International Reading Association (IRA) Standards.

The standards-writing team produced a document, *South Carolina Portrait of a College- and Career-Ready English Language Arts Student,* that served as the basis for the development of the new standards.

1.6.2 Mathematics

In the summer of 2016, the SCDE developed new mathematics standards. This process began with a review of a number of resources including the following:

- the South Carolina Department of Education (2007) *South Carolina Mathematics Academic Standards*;
- the ACT, Incorporated (2014) *ACT College and Career Readiness Standards: Mathematics*;
- the *SAT: Math Concepts*;
- college- and career-ready standards from other states including Indiana (2014), Nebraska (2009) *Nebraska Mathematics Standards*, and Texas (2009, 2014);
- the Minnesota Department of Education (2007) *Minnesota K–12 Academic Standards in Mathematics*;
- the National Governors Association Center for Best Practices & Council of Chief State School Officers (2010) *Common Core State Standards for Mathematics*; and
- National High School Center (2012) *College and Career Development Organizer*.

The standards-writing team produced the document *South Carolina College- and Career-Ready Standards for Mathematics.*

1.7 Data Reporting

The SC READY student data are reported using a vertical (across-grade) score system. Students were placed into one of four performance level categories:

(1) *Does Not Meet Expectations*;
(2) *Approaches Expectations*;
(3) *Meets Expectations*; or
(4) *Exceeds Expectations*.

Possible scale scores range from 100–950 across the entire grade three through eight scale for both ELA and mathematics. In addition to the total scale scores, students' performance on every reporting category for ELA and standard for mathematics is classified in one of three ordinal categories: *Low*, *Middle*, and *High*. Section 6.3 includes a full explanation.

1.8 Policies for Including All Students in Assessments

It is the state's policy to include all students in state assessments. The participation of local school districts in the statewide testing program is required under Section 59-20-60(7)(c) of the South Carolina Education Finance Act and the South Carolina Education Accountability Act of 1998.

For students with documented disabilities, the decision about a student's participation in the SC READY or the South Carolina Alternate Assessment (SC-Alt) is made by the student's IEP team and documented in the IEP for the student. All students with documented disabilities with IEPs or 504 Accommodation Plans must have necessary accommodations documented. The IEP or 504 Accommodation Plan must state any accommodations to be used.

With few exceptions, all students in grades three through eight attending South Carolina public schools are required to participate in either the SC READY assessments or the SC-Alt assessments to fulfill the mandates of federal and state law. This testing requirement includes all students with IEPs or 504 Accommodation Plans, suspended students, homeschool students who are registered through the district or local school board, homebound students, and homebased students. Also included are Multilingual learners (ML) students, charter school students, and students who are incarcerated.

Students who are not tested include the following:

1. students who are expelled (unless the student has an IEP);
2. homebound students for whom the district has documentation indicating that the student is not physically and/or mentally able to take the tests;
3. homeschool students who are registered through one of the professional homeschool organizations; and
4. students who attend a private school.

1.9 Participation Data

Demographic data were collected for each student. These data included information about gender, test mode, grade, race/ethnicity, English language proficiency (LEP, limited English proficiency), IEP, migrant status, gifted/talented status, 504 Accommodation Plan status, and accommodations.

Beginning in the 2010–11 school year, the SCDE adopted new federally mandated guidelines for collecting and reporting racial and ethnic data. The student database collects racial and ethnic data using a two-part question. The first part of the question asks whether or not the student is Hispanic/Latino. The second part asks whether the student is from one or more races using one of the following five racial groups: "American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," and "White." Students are not offered a choice of selecting "two or more races" or "other." A student who is Hispanic/Latino is classified as Hispanic/Latino regardless of how many races were selected. A student who is not Hispanic/Latino and indicates only one race is classified as that race. If the student is not

classified as Hispanic/Latino and indicates two or more races, the student is classified as "two or more races." State demographic reports display race/ethnicity using the eight categories that appear in the tables below: "Hispanic or Latino," "American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," "White," "Two or More Races," and "Unknown." Tables 1.1 through 1.6 present the combined student participation in the spring 2022 administration by the demographic variables.

Table 1.1. Summary of Student Demographics, Grade 3

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	55,905	100.00	55,896	100.00
Gender	Male	28,337	50.69	28,324	50.67
	Female	27,310	48.85	27,308	48.86
	Unknown	258	0.46	264	0.47
Ethnicity	Hispanic or Latino	6,632	11.86	6,633	11.87
	American Indian or Alaska Native	166	0.30	165	0.30
	Asian	1,019	1.82	1,019	1.82
	Black or African American	17,226	30.81	17,211	30.79
	Native Hawaiian or Other Pacific Islander	67	0.12	66	0.12
	White	26,782	47.91	26,732	47.82
	Two or More Races	3,317	5.93	3,311	5.92
	Unknown	696	1.24	759	1.36
IEP Status	Yes	8,036	14.37	8,023	14.35
	No or unknown	47,869	85.63	47,873	85.65
Gifted Status	Academic only	4,602	8.23	4,605	8.24
	Artistic only	234	0.42	234	0.42
	Both	73	0.13	73	0.13
	No or unknown	50,996	91.22	50,984	91.21
504 Plan Status	Yes	1,131	2.02	1,123	2.01
	No or unknown	54,774	97.98	54,773	97.99
English Proficiency Status	Active Multilingual Learners	2,367	4.23	2,356	4.21
	Met ELP Exit Criteria in Monitoring Period	1,821	3.26	1,820	3.26
	Title III Exited	311	0.56	308	0.55
	English Speaker II	51,388	91.92	51,395	91.95
	All others	18	0.03	17	0.03
Migrant Status	Yes	42	0.08	43	0.08
	No or unknown	55,863	99.92	55,853	99.92
Customized Material	Braille	6	0.01	1	0.00
	Sign Language	0	0.00	10	0.02
	Sign Language signed administration	107	0.19	72	0.13
	Large print	42	0.08	24	0.04
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	965	1.73	6,346	11.35
	Total	1,096	1.96	6,386	11.42

Note: *N* = All students who attempted the test except home school students.

Table 1.2. Summary of Student Demographics, Grade 4

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	56,397	100.00	56,404	100.00
Gender	Male	28,486	50.51	28,495	50.52
	Female	27,663	49.05	27,663	49.04
	Unknown	248	0.44	246	0.44
Ethnicity	Hispanic or Latino	6,798	12.05	6,799	12.05
	American Indian or Alaska Native	153	0.27	153	0.27
	Asian	1,013	1.80	1,013	1.80
	Black or African American	17,434	30.91	17,432	30.91
	Native Hawaiian or Other Pacific Islander	77	0.14	77	0.14
	White	26,891	47.68	26,879	47.65
	Two or More Races	3,273	5.80	3,274	5.80
	Unknown	758	1.34	777	1.38
IEP Status	Yes	8,047	14.27	8,049	14.27
	No or unknown	48,350	85.73	48,355	85.73
Gifted Status	Academic only	7,654	13.57	7,654	13.57
	Artistic only	430	0.76	430	0.76
	Both	219	0.39	219	0.39
	No or unknown	48,094	85.28	48,101	85.28
504 Plan Status	Yes	1,487	2.64	1,485	2.63
	No or unknown	54,910	97.36	54,919	97.37
English Proficiency Status	Active Multilingual Learners	2,339	4.15	2,338	4.15
	Met ELP Exit Criteria in Monitoring Period	590	1.05	590	1.05
	Title III Exited	1,683	2.98	1,684	2.99
	English Speaker II	51,762	91.78	51,770	91.78
	All others	23	0.04	22	0.04
Migrant Status	Yes	38	0.07	38	0.07
	No or unknown	56,359	99.93	56,366	99.93
Customized Material	Braille	5	0.01	2	0.00
	Sign Language	5	0.01	6	0.01
	Sign Language signed administration	133	0.24	92	0.16
	Large print	34	0.06	27	0.05
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	4,956	8.79	6,397	11.34
	Total	5,047	8.95	6,465	11.46

Note: *N* = All students who attempted the test except home school students.

Table 1.3. Summary of Student Demographics, Grade 5

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	57,065	100.00	57,066	100.00
Gender	Male	29,077	50.95	29,071	50.94
	Female	27,753	48.63	27,760	48.65
	Unknown	235	0.41	235	0.41
Ethnicity	Hispanic or Latino	6,932	12.15	6,930	12.14
	American Indian or Alaska Native	137	0.24	137	0.24
	Asian	956	1.68	955	1.67
	Black or African American	17,774	31.15	17,773	31.14
	Native Hawaiian or Other Pacific Islander	72	0.13	72	0.13
	White	27,280	47.81	27,268	47.78
	Two or More Races	3,159	5.54	3,163	5.54
	Unknown	755	1.32	768	1.35
IEP Status	Yes	8,179	14.33	8,182	14.34
	No or unknown	48,886	85.67	48,884	85.66
Gifted Status	Academic only	8,562	15.00	8,562	15.00
	Artistic only	620	1.09	620	1.09
	Both	341	0.60	341	0.60
	No or unknown	47,542	83.31	47,543	83.31
504 Plan Status	Yes	1,801	3.16	1,800	3.15
	No or unknown	55,264	96.84	55,266	96.85
English Proficiency Status	Active Multilingual Learners	2,308	4.04	2,306	4.04
	Met ELP Exit Criteria in Monitoring Period	958	1.68	958	1.68
	Title III Exited	1,675	2.94	1,675	2.94
	English Speaker II	52,106	91.31	52,109	91.31
	All others	18	0.03	18	0.03
Migrant Status	Yes	51	0.09	51	0.09
	No or unknown	57,014	99.91	57,015	99.91
Customized Material	Braille	3	0.01	2	0.00
	Sign Language	7	0.01	11	0.02
	Sign Language signed administration	117	0.21	100	0.18
	Large print	38	0.07	36	0.06
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	4,702	8.24	6,043	10.59
	Total	4,781	8.38	6,113	10.71

Note: *N* = All students who attempted the test except home school students.

Table 1.4. Summary of Student Demographics, Grade 6

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	57,566	100.00	57,662	100.00
Gender	Male	29,106	50.56	29,142	50.54
	Female	28,133	48.87	28,188	48.88
	Unknown	327	0.57	332	0.58
Ethnicity	Hispanic or Latino	7,108	12.35	7,117	12.34
	American Indian or Alaska Native	165	0.29	165	0.29
	Asian	974	1.69	974	1.69
	Black or African American	18,291	31.77	18,350	31.82
	Native Hawaiian or Other Pacific Islander	58	0.10	58	0.10
	White	27,101	47.08	27,116	47.03
	Two or More Races	3,049	5.30	3,052	5.29
	Unknown	820	1.42	830	1.44
IEP Status	Yes	7,879	13.69	7,911	13.72
	No or unknown	49,687	86.31	49,751	86.28
Gifted Status	Academic only	10,453	18.16	10,457	18.13
	Artistic only	767	1.33	767	1.33
	Both	559	0.97	558	0.97
	No or unknown	45,787	79.54	45,880	79.57
504 Plan Status	Yes	2,090	3.63	2,097	3.64
	No or unknown	55,476	96.37	55,565	96.36
English Proficiency Status	Active Multilingual Learners	2,352	4.09	2,356	4.09
	Met ELP Exit Criteria in Monitoring Period	1,727	3.00	1,724	2.99
	Title III Exited	1,283	2.23	1,281	2.22
	English Speaker II	52,171	90.63	52,268	90.65
	All others	33	0.06	33	0.06
Migrant Status	Yes	49	0.09	48	0.08
	No or unknown	57,517	99.91	57,614	99.92
Customized Material	Braille	3	0.01	12	0.02
	Sign Language	12	0.02	15	0.03
	Sign Language signed administration	71	0.12	71	0.12
	Large print	38	0.07	24	0.04
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	3,963	6.88	4,859	8.43
	Total	4,030	7.00	4,908	8.51

Note: *N* = All students who attempted the test except home school students.

Table 1.5. Summary of Student Demographics, Grade 7

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	60,109	100.00	60,155	100.00
Gender	Male	30,266	50.35	30,281	50.34
	Female	29,528	49.12	29,555	49.13
	Unknown	315	0.52	319	0.53
Ethnicity	Hispanic or Latino	7,054	11.74	7,056	11.73
	American Indian or Alaska Native	177	0.29	179	0.30
	Asian	985	1.64	984	1.64
	Black or African American	19,118	31.81	19,152	31.84
	Native Hawaiian or Other Pacific Islander	71	0.12	71	0.12
	White	28,740	47.81	28,743	47.78
	Two or More Races	3,100	5.16	3,098	5.15
	Unknown	864	1.44	872	1.45
IEP Status	Yes	7,625	12.69	7,642	12.70
	No or unknown	52,484	87.31	52,513	87.30
Gifted Status	Academic only	11,322	18.84	11,325	18.83
	Artistic only	1,094	1.82	1,094	1.82
	Both	758	1.26	759	1.26
	No or unknown	46,935	78.08	46,977	78.09
504 Plan Status	Yes	2,440	4.06	2,439	4.05
	No or unknown	57,669	95.94	57,716	95.95
English Proficiency Status	Active Multilingual Learners	2,006	3.34	2,007	3.34
	Met ELP Exit Criteria in Monitoring Period	1,886	3.14	1,891	3.14
	Title III Exited	1,596	2.66	1,595	2.65
	English Speaker II	54,605	90.84	54,645	90.84
	All others	16	0.03	17	0.03
Migrant Status	Yes	49	0.08	48	0.08
	No or unknown	60,060	99.92	60,107	99.92
Customized Material	Braille	3	0.00	3	0.00
	Sign Language	6	0.01	8	0.01
	Sign Language signed administration	53	0.09	60	0.10
	Large print	25	0.04	25	0.04
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	3,586	5.97	4,217	7.01
	Total	3,631	6.04	4,264	7.09

Note: *N* = All students who attempted the test except home school students.

Table 1.6. Summary of Student Demographics, Grade 8

Demographic	Group	ELA		Mathematics	
		N	%	*N*	%
All Students	**All**	60,920	100.00	60,987	100.00
Gender	Male	30,734	50.45	30,762	50.44
	Female	29,874	49.04	29,909	49.04
	Unknown	312	0.51	316	0.52
Ethnicity	Hispanic or Latino	7,332	12.04	7,341	12.04
	American Indian or Alaska Native	185	0.30	186	0.30
	Asian	987	1.62	988	1.62
	Black or African American	19,842	32.57	19,895	32.62
	Native Hawaiian or Other Pacific Islander	82	0.13	82	0.13
	White	28,762	47.21	28,771	47.18
	Two or More Races	2,903	4.77	2,896	4.75
	Unknown	827	1.36	828	1.36
IEP Status	Yes	7,551	12.39	7,565	12.40
	No or unknown	53,369	87.61	53,422	87.60
Gifted Status	Academic only	11,495	18.87	11,499	18.85
	Artistic only	1,251	2.05	1,252	2.05
	Both	880	1.44	880	1.44
	No or unknown	47,294	77.63	47,356	77.65
504 Plan Status	Yes	2,631	4.32	2,637	4.32
	No or unknown	58,289	95.68	58,350	95.68
English Proficiency Status	Active Multilingual Learners	2,366	3.88	2,368	3.88
	Met ELP Exit Criteria in Monitoring Period	2,319	3.81	2,322	3.81
	Title III Exited	891	1.46	894	1.47
	English Speaker II	55,317	90.80	55,377	90.80
	All others	27	0.04	26	0.04
Migrant Status	Yes	32	0.05	32	0.05
	No or unknown	60,888	99.95	60,955	99.95
Customized Material	Braille	4	0.01	3	0.00
	Sign Language	6	0.01	9	0.01
	Sign Language signed administration	36	0.06	43	0.07
	Large print	23	0.04	23	0.04
	Loose leaf	0	0.00	0	0.00
	Form A oral administration	3,380	5.55	3,913	6.42
	Total	3,415	5.61	3,955	6.48

Note: *N* = All students who attempted the test except home school students.

Section 2—Assessment System Operations

2.1 Test Design and Development

This section of the report provides a high-level description of the South Carolina content area standards and a description of how those content area standards are being measured on the SC READY tests. Content-related evidence of the validity of the intended score interpretations in SC READY testing is supported by the degree of correspondence or alignment between the assessments and the specifications of the standards that are assessed (i.e., what students should know and be able to do at a given grade and content area). In this section, evidence of content-related validity is demonstrated through each SC READY assessment's consistent adherence to the assessment blueprints, which were constructed by South Carolina educators based on the South Carolina Academic Standards.

According to the most recent edition of the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). As stated above, essential validity evidence supporting the development of the SC READY assessments is well documented through the item and test development process, including the review of the assessment items for alignment to the *SC CCR Standards for ELA and Mathematics* that the SC READY measures.

The information found in this section provides an overview of the *SC CCR Standards for ELA and Mathematics* and the process used for the development of the blueprints for ELA and mathematics for grades three through eight. This section also includes a description of the involvement of educators, which serves to demonstrate adherence to AERA, APA, & NCME (2014) Standards 3.1, 3.2, 4.0, 4.1, 4.7, and 4.12.

2.1.1 Development of Test Blueprint and Specifications

The purpose of this section is to document the test development process used for the SC READY tests. Test development and psychometric staff worked extensively with the SCDE in the construction of all test forms to support post-administration equating. All form construction activities focused on: (1) ensuring that test blueprint requirements were met and (2) concurrently matching test characteristic curves and test information functions to previous forms. After items had been selected and reviewed by the test development and psychometric specialists for content area excellence and technical quality, test maps for each subject were created. Each anchor set was selected to match the test blueprint and the test difficulty of previous operational administrations. Once the anchor set had been selected, the remainder of the assessment was selected according to the same rigorous content area and psychometric specifications.

AERA, APA, & NCME (2014) Standard 4.0 states the following:

> 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. (p. 84)

Operational tests were designed based on the test specifications by combining expert review with intensive test construction processes. Once test selections had been made, content area experts reviewed the selections to confirm appropriate alignment with the test specifications, and psychometric experts reviewed the statistical summary information.

AERA, APA, & NCME (2014) Standard 4.1 states the following:

> 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). (p. 85)

The key structural aspect of the SC READY ELA and mathematics assessments is the assessment blueprint, which specifies the target score points for each grade and content area standard. The overall structure of the SC READY design is found in Table 2.1.

The 2022 SC READY operational forms matched the test blueprints found in Tables 2.2 and 2.3, including the actual point distributions. Actual point distributions on the 2022 SC READY operational forms matched blueprint targets. Item types found in the SC READY assessments can be found in Table 2.4.

Table 2.1. SC READY Test Design

Subject	Grade	No. of OP Items/Points	No. of FT Items/~~Points~~	TOTAL No. of items/Points
ELA	3	58	7	66
	4	59	7	66
	5	59	7	66
	6	60	8	68
	7	60	8	68
	8	60	8	68
Mathematics	3	50	6	56
	4	56	6	62
	5	56	6	62
	6	60	6	66
	7	60	6	66
	8	62	6	68

Table 2.2. ELA Blueprint

Grades	Reporting Strand	Key Idea	Number of Points	% of Test
3-5	Reading Literary Text	Meaning and Context	9–11	14–17
		Language, Craft, and Structure	8–10	12–15
	Reading Informational Text	Meaning and Context	9–11	14–17
		Language, Craft, and Structure	8–10	12–15
	Writing	Meaning Context, and Craft	6–8	9–12
		Language	6–8	9–12
		Text-Dependent Analysis	8	12
	Inquiry		6–8	9–12
		Total	**66**	**100**
6-8	Reading Literary Text	Meaning and Context	9–11	13–16
		Language, Craft, and Structure	8–10	12–15
	Reading Informational Text	Meaning and Context	9–11	13–16
		Language, Craft, and Structure	8–10	12–15
	Writing	Meaning Context, and Craft	6–8	9–12
		Language	6–8	9–12
		Text-Dependent Analysis	8	12
	Inquiry		6–8	9–12
	Total	**Total**	**68**	**100**

Table 2.3. Mathematics Blueprint

Grade	Reporting Category	Number of Items	% of Test
3	1. Number Sense and Base Ten	8–9	16–18
	2. Number Sense – Fractions	8–9	16–18
	3. Algebraic Thinking and Operations	12–16	24–32
	4. Geometry	8–9	16–18
	5. Measurement and Data Analysis	12–16	24–32
	Total	**50**	**100**
4	1. Number Sense and Base Ten	10–12	18–21
	2. Number Sense – Fractions	11–14	20–25
	3. Algebraic Thinking and Operations	11–14	20–25
	4. Geometry	8–10	14–18
	5. Measurement and Data Analysis	11–14	20–25
	Total	**56**	**100**
5	1. Number Sense and Base Ten	10–13	18–23
	2. Number Sense – Fractions	10–12	18–21
	3. Algebraic Thinking and Operations	10–13	18–23
	4. Geometry	10–12	18–21
	5. Measurement and Data Analysis	11–14	20–25
	Total	**56**	**100**
6	1. The Number System	12–16	20–27
	2. Ratios and Proportional Relationships	8–11	13–18
	3. Expressions, Equations, and Inequalities	12–16	20–27
	4. Geometry and Measurement	8–11	13–18
	5. Data Analysis and Statistics	11–14	18–23
	Total	**60**	**100**
7	1. The Number System	12–15	20–25
	2. Ratios and Proportional Relationships	8–10	13–17
	3. Expressions, Equations, and Inequalities	11–15	18–25
	4. Geometry and Measurement	10–13	17–22
	5. Data Analysis, Statistics, and Probability	12–15	20–25
	Total	**60**	**100**
8	1. The Number System	8–12	13–19
	2. Functions	11–14	18–23
	3. Expressions, Equations, and Inequalities	12–16	19–26
	4. Geometry and Measurement	12–16	19–26
	5. Data Analysis, Statistics, and Probability	8–12	13–19
	Total	**62**	**100**

Table 2.4. SC READY Item Types

Content Area Information	Item Type	Details
ELA (3–8) Mathematics (3–8)	Selected Response (SR)	Students select one response from four possible answer options.
ELA (3–8)	Multi-Select (MS)	Students will be prompted to select a number of correct answers (e.g., "Choose **two** answers . . ."). The multi-select items may have 5 or 6 answer choices. The number of correct answers will be more than one choice, but fewer than all choices. In order to receive credit for a correct response, students must select **all** of the correct answer choices and **only** the correct answer choices.
ELA (6–8) Mathematics (6–8)	Technology Enhanced (TE)	For online testers only, students interact with the item (in DRC INSIGHT) to provide their response (e.g., drag and drop, drop-down menu, hot spot). Comparable SR items will replace TE items on the paper tests.
ELA (3–8) Session 1 Only	Text-Dependent Analysis	Students read a piece of text or passage and draw upon that text for their extended written responses—e.g., support their responses with evidence from the text.
ELA (3–8) Session 2 Only	Evidence-Based Selected Response (EBSR)	These are two-part items. Students read a piece of text or passage and choose the best answer from the answer choices. Students will then be asked to support their response with evidence from the text.
Mathematics (5–8)	Multi-Select (MS)	Students will be prompted to select **all** of the correct answer choices. The multi-select items may have 5 or 6 answer choices. The number of correct answers will be more than one choice, but fewer than all choices. In order to receive credit for a correct response, students must select **all** of the correct answer choices and **only** the correct answer choices.
Mathematics (6–8)	Short Answer (SA) or Gridded Response (GR)	For online testers, students will key a numeric response (short answer) in DRC INSIGHT. For paper testers, students will grid a numeric response (gridded response) on their answer document.
Mathematics (6–8)	Keypad Input	For online testers, some items will have a keypad input. Students may use either the online keypad or the physical keyboard to enter their response. The keyboard will be limited by the buttons available in the online keypad. Depending on the grade-level of the item, the online keypad will contain buttons for the digits 0–9, a fraction button, a negative button, and/or a decimal point button. For paper testers, students will grid some of their response to this item type in the answer document.

2.1.2 Universal Design

SC READY assessments are universally designed to allow for the participation of the widest possible range of students, resulting in more valid inferences about student performance. Universally designed grade-level assessments may reduce the need for accommodations by reducing or eliminating access barriers associated with the tests themselves. Table 2.5 presents the elements of universal design that were implemented on the SC READY assessments (Thompson & Thurlow, 2002; Center for Universal Design, 1997).

These elements of universal design are relevant to both item development and form construction. This section addresses how the elements of universal design were incorporated in the construction of the spring ~~2022~~ test forms in compliance with AERA, APA, & NCME (2014) Standard 3.1, which states the following:

> Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)

A goal of universal design is to measure the performance of students with a wide range of abilities and skills, ensuring that students with diverse learning needs receive opportunities to demonstrate competence in the same content area. To accommodate the greatest number of students for the SC READY tests, the assessments include simple, clear, and intuitive instructions and procedures; maximum readability and comprehensibility; and maximum legibility. These design components are addressed primarily through the physical layout and formatting of the online test forms and the paper-based test forms used for accommodations. The page specifications define how directions and test items are placed on the pages, the location and appearance of headers and footers, the spacing between an item stem and the answer choices, and other page elements to ensure a consistent, legible appearance of online forms and paper-based test forms. Written instructions at the beginning of each test session are clearly and simply stated, and the wording of such instructions is standardized as much as possible across content areas and grade levels to ensure clarity and consistency.

Table 2.5. Elements of Universal Design

Element	Explanation
Inclusive Assessment Population	Tests designed for state, district, or school accountability must include every student except those in the alternate assessment, and this is reflected in assessment design and field-testing procedures.
Precisely Defined Constructs	The specific constructs tested must be clearly defined so that all construct-irrelevant cognitive, sensory, emotional, and physical barriers can be removed.
Accessible, Unbiased Items	Accessibility is built into items from the beginning, and bias review procedures ensure that quality is retained in all items.

Element	Explanation
Amenable to Accommodations	The test design facilitates the use of needed accommodations.
Simple, Clear, and Intuitive Instructions and Procedures	All instructions and procedures are simple, clear, and presented in understandable language.
Maximum Readability and Comprehensibility	Readability and plain language guidelines are followed (e.g., sentence length and number of difficult words are kept to a minimum) to produce readable and comprehensible text.
Maximum Legibility	Characteristics that ensure easy decipherability are applied to text, tables, figures, illustrations, and response formats.

2.1.3 Item Development

AERA, APA, & NCME (2014) Standard 4.12 states the following:

> Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications. (p. 89)

The SC READY item specifications are designed to ensure that the assessment items measure the assessment's domains. The purpose of the item specifications is to define the characteristics of the items that will provide the evidence to support one or more domains. To do this, the item specifications delineate the types of evidence that should be elicited for each strand within a grade level. Then, they provide explicit guidance on how to write items in order to elicit the desired evidence.

In doing this, the item specifications provide guidance on how to measure the standards. The item specifications also provide guidance on how to create items that are specific to each assessment domain or strand. In ELA and mathematics, item specifications describe the knowledge, skills, and processes being measured by each of the item types aligned to *SC CCR Standards for ELA and Mathematics*. These item specifications were developed for each grade level and standard in order to delineate the expectations of the knowledge and skills measured by the items on the SC READY tests at each grade level.

All SC READY items were developed with reference to the *SC CCR Standards for ELA and Mathematics* and measurement guidelines. All newly developed items were reviewed by committees of South Carolina educators for content area and bias and sensitivity issues; items approved by these committees and the SCDE were field-tested among South Carolina students. Items demonstrating satisfactory performance on field tests became eligible for inclusion in operational forms during the subsequent administration.

New item development evaluates items using the following criteria:

- **Content alignment**—determines whether an item measures what it is intended to measure by matching items to a standard and indicator
- **Rigor-level alignment**—determines cognitive complexity and Depth of Knowledge and examines for appropriateness to the rigor required
- **Technical design**—determines whether an item is current and accurate and whether its stem, stimuli, distractors, and answer options are clear and concise, grade-level appropriate, and considerate of students with special needs
- **Universal design**—determines whether an item provides for an accessible assessment of all students, focusing on language demand, format/complexity, and graphics/visuals
- **Fairness in testing**—determines whether an item generates valid test scores for all groups of test takers by avoiding bias in test items and/or content area and avoiding language that unduly distracts students or disrupts their performance

Content specialists also check to see that the items comply with the guidelines provided by the SCDE, including matching the items to an appropriate standard. DRC's item development work was and continues to be designed to produce reliable and instructionally valid tests that adhere to the guidelines articulated in the AERA, APA, & NCME (2014) *Standards*. In particular, the item development process discussed in this section is in compliance with AERA, APA, & NCME (2014) Standard 4.7, which states the following:

> The procedures used to develop, review, and try out items and to select items from the item pool should be documented. (p. 87)

As noted in the item specifications, the SC READY ELA and mathematics assessments include several types of items including selected-response, technology-enhanced, EBSR, and TDA. Table 2.6 summarizes the steps DRC content specialists used in the preparation of items for the SC READY program.

Table 2.6. SC READY Item Bank Development Activities

SC READY Item Bank Development Activities
Establish item/scenario development specifications and style guides and prepare item writing training manuals.
Determine item development plans.
Train item writers and/or scenario developers in the project requirements and specifications.
Develop passages and write items.
Review, edit, code, and track items and produce graphics.
Produce review forms for content and bias/fairness/sensitivity reviews by external reviewers.
Modify items based on external reviewers' recommendations.

SC READY Item Bank Development Activities
Review and approve field-test–ready items and scenarios.
Develop field-test forms and administer field-test.
Internally review field-test item data.
Approve items to be included in the item bank.

2.1.4 Content and Bias Reviews

All newly developed items are put through a rigorous review process within DRC, and Bias and Sensitivity issues are addressed by DRC content area specialists and a DRC bias and sensitivity specialist. South Carolina educators with a background in the content area reviewed all items as a separate committee simultaneous with item review. Items put forward following approval from DRC and the SCDE were field-tested among South Carolina students. Items demonstrating satisfactory performance on field tests became eligible for inclusion in operational forms during the following administration.

New item development focuses on items using the following criteria:

- Content alignment
- Rigor-level alignment
- Technical design
- Universal design
- Fairness in testing

During the Bias and Sensitivity meeting with South Carolina educators, DRC also provided training on the procedures and forms used for item content review. The content and bias training addressed AERA, APA, & NCME (2014) Standard 3.2, which is relevant to fairness in item development:

> 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. (p. 64)

Additionally, participants were provided training on how to apply the Principles of Universal Design and the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) to ensure that each item developed was fair, reliable, and educationally sound. Committee members were grouped by grade level and content area.

The members of the review committees provided feedback for each item, and committee facilitators recorded the committee decisions on the item review rating forms provided by DRC.

Items accepted for use on the SC READY ELA and mathematics assessments constituted the pools of items from which subsequent test forms can be constructed.

Table 2.7. SC READY ELA Item Development

Item Type	Number of Items Developed at Each Grade					
	3	4	5	6	7	8
SR	140	140	132	131	120	131
EBSR	0	0	8	10	9	10
MS	0	0	0	0	1	0
TE	NA	NA	NA	0	1	0

Table 2.8. SC READY Mathematics Item Development

Item Type	Number of Items Developed at Each Grade					
	3	4	5	6	7	8
SR	96	96	82	82	82	82
EBSR	NA	NA	NA	NA	NA	NA
MS	NA	NA	NA	14	14	14
TE	NA	NA	NA	0	0	0

2.1.5 Data Review

The review of the spring 2022 field test item data was conducted internally by SCDE content area test development specialists and psychometricians. The review process involved a brief exploration of possible reasons for the statistical profile of an item (e.g., possible bias, grade inappropriateness, and instructional issues) and a decision regarding the acceptance or rejection of that item. SCDE content area experts reviewed the pool of field-tested items and made recommendations on each item, scenario, or passage. Items accepted for subsequent use in the SC READY assessments were included in the pool of items for spring 2022 operational test form selection.

2.1.6 Field-Testing

This section of the report describes the timeline and process of field-testing SC READY ELA and mathematics items for future use on SC READY. The SC READY ELA and mathematics items accepted during the content and bias reviews were field-tested prior to the spring 2022 assessment. The online test forms were spiraled at the student level within a grade and content area.

Following each field test data acquisition, the field test data analyses were conducted, with the most recent field test analysis conducted after the spring ~~2022~~ test administration. The analyses included classical item analysis, differential item functioning (DIF) analysis, and Rasch analysis. The classical item analysis included the computation and evaluation of the following statistics: item *p*-values (difficulty), item-total test correlation, percentage of students selecting incorrect responses, point-biserial correlation for incorrect responses for the selected-response

(SR) items, score point distribution for items worth more than one point, and omit rates for all items. More details on classical item analysis methodology are provided in Section 4.6.1 of this report.

DIF analysis was conducted for all field test items to examine potential item bias and to determine whether item performance differences between identifiable subgroups were due to factors other than student ability, making the items unfairly difficult for a subgroup in the student population. DIF analyses were conducted based on gender, race/ethnicity, and accommodation use. More details on the DIF methodology are provided in Section 4.3.2 of this report.

Item statistics are used as a means of detecting items that deserve closer scrutiny, rather than as mechanisms for automatic retention or rejection. To this end, a set of criteria was used as a screening tool to identify items that needed a closer review. The criteria for an item to be flagged for an additional review included the following:

- *p*-value <0.20 or >0.90
- item-total test correlation (point biserial for SR items) <0.20
- positive point biserial on a distractor for an SR item
- omit rate >5%
- item flagged for DIF

Items flagged for any of the above reasons were reviewed by the content area specialists and SCDE prior to their selection as operational items on any future SC READY test. The intent was to capture all items requiring additional review based on their statistical properties; thus, the criteria employed for item flagging tended to over-identify rather than under-identify potential item issues.

2.1.7 Form Construction Process

This section provides an overview of a very specific set of guidelines relative to the selection of items in the form construction process. DRC believes that a key factor in form construction is a solid understanding of the content area curriculum standards and the test specifications. Items selected to appear on forms must not only meet psychometric qualities for excellence, they must also meet technical quality in terms of content area and conventions of good item writing and construction. DRC uses a series of steps to determine the technical quality of each item, including reviewing each item matches the given standard. The entire pool of items was provided to the DRC content area assessment specialists. The following bulleted list serves to summarize the steps DRC used when selecting items.

The total number of operational items on each form and the number of operational items within each reporting category must follow the test blueprints approved by the SCDE. This is explained more specifically below.

- Forms should be built according to the reporting category level.

- Within each reporting category, the items should provide full coverage of the indicators that define the reporting category, according to the test blueprint.
  Note that there are indicators with '0' as the minimum number of items. Therefore, some indicators may occasionally (and appropriately) have zero representation on a test.
- Using the eligible pool of items and the most recent item performance data (operational or field test) for items of each subject and grade, DRC content area specialists first select items to match the blueprint, standards, and indicators.
- DRC content area specialists ensure that each item measures the content area standards/indicators specified in the applicable standards documents.
- DRC content area specialists check to see whether each item meets psychometric guidelines for excellence.
- DRC content area specialists will check to see that each item meets technical quality for well-crafted items, including having only one clearly correct answer for multiple-choice items; having wording that is clear and concise; being grammatically correct; being appropriate for the range of difficulty; and being free of any content that might be offensive, inappropriate, or biased.

The construction of the test forms themselves was a collaborative effort between SCDE content area staff and DRC's integrated development team of assessment specialists, a psychometrician, and scoring specialists. The content and psychometric criteria used for item selection included the following:

- Test length and item types adhered to the SCDE-approved test design.
- Content coverage adhered to the SCDE-approved test blueprint.
- Items were evaluated for technical quality, including that each item
  - had one clearly correct answer (or answers if multi-select or technology-enhanced),
  - used clear and concise wording,
  - was grammatically correct,
  - had an appropriate range of difficulty,
  - was free of any offensive, inappropriate, or biased content, and
  - met the Principles of Universal Design and maximum accessibility.
- Recommended psychometric properties of the items included
  - a *p*-value between 0.25 and 0.88,
  - an item-total test correlation >0.20,
  - omit rates $\leq$ 5%,
  - an acceptable item fit (no misfit flag), and
  - no DIF flag. If an item with DIF had to be included in the test to maintain blueprint coverage, the item was examined to determine whether any content reason exists for the DIF flag—sometimes items demonstrate statistical bias but no content reason can be determined for the bias.

The anchor (linking) items were selected from the spring 2021 operational test items for ELA and mathematics. These anchor items were common for the two operational core forms in each grade level and content area and were selected to provide the link between the two forms.

The anchor set was selected as a "mini" version of the full operational test for each grade level and content area in regard to its length, content coverage, and psychometric properties. The items included in the anchor sets met the same blueprint specifications as the full test in regard to the percentage of score points.

The non-anchor operational items had been field-tested on an SC READY form prior to spring 2022. Full form selections met the criteria for statistical equivalence within each grade level and content area and for statistical alignment with the spring 2018 operational test forms for the SC READY ELA and mathematics assessments.

In addition to the core online operational test forms, breach forms and paper-and-pencil accommodated forms, including Braille and Large Print forms, were available for administration in spring 2022. SCDE reviewed the items placed on the operational test forms during the form construction meeting.

New field test items were appended to the test forms for ELA and embedded for mathematics in the test forms. ELA test forms had 7 or 8 field test item positions per grade, and mathematics test forms had 6 field test item positions per grade (Table 2.1 contains detailed information about test design). The field test items embedded in the field test positions were accepted for field-testing during the content and bias review.

All forms were reviewed and approved by DRC psychometric staff and content area staff as well as content area staff from SCDE. At each step of the process, SCDE staff was involved in the review and approval of the forms. SCDE sign-off on each grade and content area form was required prior to proceeding to the next step of the forms process. SCDE staff reviewed both paper-and-pencil forms and online forms. The online forms were made accessible to SCDE for review in DRC's secure INSIGHT testing engine in two different stages. Upon receipt of SCDE feedback, DRC test development specialists adjusted the ELA and mathematics forms as needed. A subsequent review in the INSIGHT engine was also provided to ensure that all changes were made and to complete a final rendering check in the final production environment.

2.1.8 Multiple Assessment Forms

The majority of SC READY tests are computer-based assessments, but paper-based tests are available. All students testing online, are provided a passage booklet that contains operational as well as field test passages. Students may reference this paper-based version of the passages in place of reading the passages in the online format. Accommodated forms are available for students who require them due to disability or for students in a few uncommon situations, such as homebound students or students in group homes without adequate internet access. All SC READY tests have a version translated into American Sign Language (ASL). The computer-based

ASL version has an embedded video of a professional signer. For students who require a paper-and-pencil test, an ASL script is available. Computer-based SC READY tests can be delivered by different types of devices (e.g., desktop computers, laptops, or tablets). Information on the type of device is maintained in the online testing engine, but analysis is currently not performed by device type.

The comparability of the different versions for each subject combination (online and paper-and-pencil) is addressed by using Mode DIF when sufficient numbers of student records are available to perform the analyses. The results of the analyses would highlight the comparability of the two modes of administration. Section 4.3 has more details on the Mode DIF results. Section 2.1 has information on test design and development.

2.2 Test Administration

The next section examines how test administration procedures implemented for the SC READY assessments strengthen and support the intended score interpretations and reduce construct-irrelevant variance, which could threaten the validity of score interpretations.

This section describes how the SC READY assessments demonstrate adherence to AERA, APA, & NCME (2014) Standards 3.9, 4.15, 4.16, 6.1, 6.4, 6.6, and 7.2. Each standard will be explicated within the relevant sections.

2.2.1 Description of Target Student Population

Standard 7.2 provides general guidance that is relevant to this section:

> The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported. (p. 126)

For the purpose of this report, the South Carolina student population is defined as all students in public and charter schools who were eligible to take the SC READY assessments. Homeschool students can participate in SC READY testing, but their records are not included in the summary. The characteristics of the students who were eligible to participate in the SC READY spring 2022 assessments are presented in Tables 1.1 through 1.6. The number of students ranged from approximately 50,300 to approximately 51,500 per grade level for all content areas. For each grade level, approximately 49% of South Carolina students who participated in the SC READY assessments in spring 2022 were female and about 51% of students were male. In terms of student ethnicity, approximately 49–51% of students were White, about 29–30% of students were Black (not Hispanic), approximately 12% of students were Hispanic, 2% of students were Asian, less than 1% were Native Hawaiian or Other Pacific Islander, and less than 1% of students were American Indian or Alaska Native. In addition, approximately 4–6% of students identified themselves as Multi-racial (Two or More Races). Approximately 12–15% of students had an IEP, 12–20% of students were identified as having a gifted status, and 2–4% had a 504

Accommodation Plan. Approximately 3–5% of students were not identified as English speakers. Less than 1% of students were identified as migrants. Lastly, approximately 1–11% of students used customized materials.

2.2.2 SC READY Testing Window

The 2022 SC READY tests were administered to students between April 7 to June 6, 2022. The assessments must be administered during the last twenty school days as determined by each district's instructional calendar. Additionally, if a test session is started, it must be completed within the same day (except for students with IEPs or 504 Accommodation Plans or Multilingual learners (ML) students with scheduling accommodations).

2.2.3 Test Time

The SC READY tests were not timed; however, each session had to be administered during a single day (unless a student's IEP or 504 Accommodation Plan specifically stated that the student needed to have the test administered over several days). To ensure an accurate assessment, districts and schools were instructed that students should be given as much time as they needed to complete the test.

For online testing, start and stop times were recorded automatically. For paper-and-pencil testing, students were asked to record on their answer documents the exact times that they started and finished the test. The total elapsed time was calculated for each student. (It was not possible to calculate a total testing time for students with incomplete or invalid data.) Total elapsed time may be influenced by the presence of field test items on some forms. Table 2.9 shows that most students finished each section of the ELA test within 95 minutes. Table 2.10 shows that most students finished both sessions (note that grades three through five have one session) of the mathematics test within 125 minutes.

Table 2.9. SC READY Test Time Distribution (In Minutes), ELA

Grade	Session 1				Session 2			
	N Items	25th Percentile	Median	75th Percentile	*N* Items	25th Percentile	Median	75th Percentile
3	23	51	75	108	43	54	75	105
4	23	55	78	112	43	60	84	115
5	24	60	88	120	42	63	87	116
6	26	52	72	98	43	55	74	96
7	26	51	70	94	43	54	72	91
8	26	50	68	91	43	51	66	85

Note: The number of items includes both operational and field-test items. ELA Session 1 tests Writing and Session 2 tests Reading.

Table 2.10. SC READY Test Time Distribution (In Minutes), Mathematics

Grade	Session 1				Session 2			
	N Items	25th Percentile	Median	75th Percentile	*N* Items	25th Percentile	Median	75th Percentile
3	56	52	71	98				
4	62	61	84	114				
5	62	66	89	120				
6	46	52	69	90	20	16	22	30
7	46	50	65	83	20	11	16	22
8	47	45	59	76	21	11	15	20

Note: The number of items includes both operational and field-test items. Mathematics Session 1 allows a calculator and Session 2 does not allow a calculator.

2.2.4 Test Administration Manuals

Test administration procedures and guidelines for the SC READY assessments are included in ancillary materials and contribute to the body of evidence of the validity of score interpretation. This section examines how the test materials address the specific AERA, APA, & NCME (2014) Standard*s* related to test administration procedures.

Working with the SCDE, DRC staff drafted the administration manuals for the tests. SCDE staff reviewed and revised the manuals, and DRC finalized and printed them. Test Administration Manuals (TAM, found in Appendix A) were provided each spring administration. The TAMs are for online and paper-and-pencil testing and were available for download from both the SCDE and DRC websites. The TAMs contained the information that school test coordinators (STCs), test administrators (TAs), and monitors needed to administer the tests to students in their schools.

The TAMs included logistical and administrative procedures as well as the directions (scripts) for administering the tests that are both general and grade specific. Instructions for completing student demographic information and returning scorable and nonscorable test materials were also included. The district test coordinators (DTCs), STCs, and TAs were encouraged to offer comments and suggestions on the procedures therein.

Standard 6.1 of AERA, APA, & NCME (2014) states the following:

> Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user. (p. 114)

The TA section of the TAM outlines steps that should be followed when administering the SC READY tests. This section presents the AERA, APA, & NCME (2014) standards that are relevant to test administration and how the information in the TA section addresses these standards.

Standard 4.15 of AERA, APA, & NCME (2014) states the following:

> 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. (p. 90)

The TA section provides instructions for activities before, during, and after testing with sufficient detail and clarity to support reliable test administrations by qualified TAs. To ensure uniform administration conditions throughout the state, instructions in the TA section describe the following: general rules of online testing; pause rules; scheduling requirements for the tests; recommended order of test administration; classroom activity information; assessment duration, timing, and sequencing information; and materials that the examiner and students need for testing.

Standard 4.16 of AERA, APA, & NCME (2014) states the following:

> 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 specification or domain should be provided to the test takers prior to the administration of the test or should be included in the testing material as part of the standard administration instructions. (p. 90)

To ensure clarity of instructions to students, the TAMs include scripts that the TAs are instructed to read verbatim to students. TAs are instructed to follow the scripts and to repeat any part of the directions as many times as needed but to not modify the words used. TAs may use professional judgment to respond to student questions, but they may not reword test items, suggest answers, or evaluate student work during the testing session. A sample of a script is presented in "Administration Directions Manual" (for more information see the TAM).

Online Tools Training (OTT) tutorials and practice tests are provided, in advance of the SC READY assessment window, in both content areas to familiarize students/users with the navigation of the online systems, functionality of the testing environment, and different item types. Districts have the following options for training students on interacting with the INSIGHT testing platform and using the tools contained within INSIGHT:

- OTT gives students/users the ability to use the tools available in the INSIGHT testing platform on a variety of item types that will be used in the operational assessments. Using the OTT allows students/users to become comfortable with using the built-in system tools prior to the summative assessment. There is no limit to the amount of times a student/user can access the OTT.

- Online Tutorials give students/users the ability to watch recorded videos that demonstrate the features of INSIGHT and the tools that will be used for the operational assessments.

To ensure the usefulness and interpretability of test scores and to minimize sources of construct-irrelevant variance, it is essential that the SC READY tests are administered according to the prescribed test schedule. The STC section of the TAM includes instructions for scheduling the test within the state testing window. The TA section contains the schedule for timing each test session and indicates whether timing should be enforced.

Standard 6.4 of AERA, APA, & NCME (2014) states the following:

> The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance. (p. 116)

The STC and monitor sections of the TAM overview the conditions that TAs should meet to prepare for administering the SC READY tests. These include the following:

- Administer tests in a familiar classroom or computer lab setting to reduce student test anxiety and simplify test security.
- Ensure that each TA has created a seating chart for each testing session and that measures have been taken to provide maximum privacy for each student in the testing room.
- Use a Do Not Disturb sign on the door of the testing room.
- Ensure that subject-related materials displayed on walls, halls, desks, or windows are covered or removed prior to testing.

Standard 6.6 of AERA, APA, & NCME (2014) states the following:

> 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. (p. 116)

The TA and STC sections of the TAM present instructions for post-test activities to ensure that online tests are submitted properly and printed test materials are handled properly, ensuring the integrity of student information and test scores. Detailed instructions guide test examiners in submitting all online test records. For students who are administered a Large Print or Braille version of the SC READY tests, examiners are instructed to transcribe students' responses from the Large Print test or Braille test book into the answer document or online testing system (INSIGHT) exactly as the students responded in the Large Print or Braille test book.

TAs are given guidelines on how to handle a wide range of testing disturbances or improper activities that may occur. Testing concerns related to improper activity should be reported to the DTC. Specific cases will be handled at the school or district level, depending on district procedures.

Throughout the TAM, STCs, TAs, and monitors are reminded of test security requirements and procedures to maintain test security. Specific actions that are direct violations of test security are so noted.

2.2.5 Accommodations and Universal Tools

Universal Supports are not intended to eliminate individualization but rather to reduce the need for certain accommodations and various alternative assessments by eliminating access barriers associated with the tests themselves. Universal Supports are available to all students taking state assessments in order to address their individual accessibility needs. The available universal supports can be found in Appendix C of the TAM. Universal supports are available to all students with or without a documented disability. However, educators may determine that one or more might be distracting for a particular student and, thus, might indicate that the support should not be used for the administration of the assessment to the student.

This complies with AERA, APA, & NCME (2014) Standard 3.9, which states the following:

> Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs. (p. 67)

A student's IEP or 504 Accommodation Plan team determines how, not if, a student with disabilities participates in the SC READY assessments. Decisions about accommodations and alternate assessments must be made on an individual student basis, not on the basis of the category of disability or instructional placement.

"*Accommodations* are adaptations to test format or administration (such as changes in the way the test is presented, the setting for the test, or the way in which the student responds) that maintain the same construct and produce results that are comparable to those obtained by students who do not use accommodations." (AERA, APA, & NCME, 2014). More information related to accommodations and universal tools can be found in Section 5.3.

2.2.6 Return Material Forms and Guidelines

Due to the preponderance of online testing, the need for shipping and return of physical materials had been greatly reduced. For districts testing online, the test tickets were available for download and printing approximately two weeks prior to testing. Test ticket rosters must be used to track and monitor the distribution and receipt of student test tickets. For each day of testing, STCs collected all online test materials from TAs, including testing rosters, student test tickets, and seating charts.

Materials for paper-and-pencil tests were shipped to schools approximately two weeks before testing—in time for the DTCs to distribute school materials at least one week before the schools' test dates. Each school's shipment was boxed individually and labeled with the total number of boxes shipped to that school. For the eleven largest districts and the SC public charter school district, materials were shipped directly to schools.

The district offices were also sent a shipment of non-customized overage materials, which were to be used by the DTCs to complete any additional materials requests from the STCs. Materials in customized formats were sent only to the schools and only in the quantities ordered.

TAs were instructed to return their test materials to the STCs immediately after the test administration. The STCs then redistributed test materials to the TAs who needed them in order to administer makeup tests. Those TAs were instructed to return the makeup test materials to their STCs immediately after the makeup session. DTCs were to arrange for the pickup of all scorable materials for return to DRC within three days of testing.

Each school district could return the scorable materials to DRC in as many shipments as needed. Nonscorable materials were to be returned in one shipment within three days of the completion of makeup tests. Step-by-step instructions for returning scorable and nonscorable materials were included in the TAM and Materials Receipt and Return Supplement. These instructions listed the toll-free phone numbers of the shipping companies that the DTCs were instructed to call to schedule pickups of return materials.

2.2.7 Administrative Support and Training

To ensure that the SC READY tests are administered and scored in accordance with the AERA, APA, & NCME *Standards*, SCDE takes the primary role in communicating with and training district personnel. SCDE conveys to districts the purpose of the assessments and the importance of test administration being consistent with test industry standards. The tests and the consistent standards of administration must also meet the State Board of Education policies and the mandates of both state and federal legislation.

To accomplish these goals, the SCDE contracted with DRC to provide training for DTCs. DRC provided Technology Coordinator training on September 21, 2021. The DTC training was held via a webinar on March 3, 2022. An STC/TA training PowerPoint was posted for districts and schools on March 11, 2022.

The DTCs are responsible for training the schools within their districts. They disseminate information to each school, offer assistance with test administration, and serve as the liaisons between the SCDE and their districts.

2.3 Test Security

Test security is an important issue before, during, and after test administrations. The specific procedures to be followed during the SC READY test administration are outlined in the TAM (2021a and 2021b). The manual includes an excerpt from Section 59-1-445 of the South Carolina Code of Laws, a summary of Section 59-1-447 of the Code of Laws, and the entirety of State Board of Education Regulation 43-100.

Section 59-1-445 states the following in part:

> It is unlawful for anyone knowingly and willfully [*sic*] to violate security procedures regulations promulgated by the State Board of Education for

mandatory tests administered by or through the State Board of Education to students or educators, or knowingly and willfully to:

(a) Give examinees access to test questions prior to testing;
(b) Copy, reproduce, or use in any manner inconsistent with test security regulations all or any portion of any secure test booklet;
(c) Coach examinees during testing or alter or interfere with examinees' responses in any way;
(d) Make answer keys available to examinees;
(e) Fail to follow security regulations for distribution and return of secure test [materials] as directed, or fail to account for all secure test materials before, during, and after testing;
(f) Participate in, direct, aid, counsel, assist in, encourage, or fail to report any of the acts prohibited in this section.

Regulation 43-100 mandates that "each local school board must develop and adopt a district test security policy" with procedures for the storage and handling of all test materials and that each district superintendent must annually designate a DTC. The regulation and the TAM provide specific security guidelines regarding various aspects of the test administration process (e.g., the storage and handling of test materials, the responsibility of administrators to monitor students during testing and to remove supplemental materials from the testing room, and the requirement that administrators refrain from interference with student responses).

Following the test administration and the return of materials, DRC generated a missing- document report, listing the identification numbers of unreturned secure materials. The report was used to notify districts of missing materials. A toll-free telephone line was provided to answer questions regarding missing documents, and follow-up procedures were employed until all materials were accounted for. Subsequently, the districts located and returned the materials or sent signed statements indicating that all secure materials had been returned.

2.3.1 Secure Materials

Secure materials—each assigned a human- and machine-readable security identification number—are test booklets, answer documents, customized test materials, and administration scripts. For online testing, secure materials consist of passage booklets, student test tickets, student rosters, and any materials containing student writing. Secure materials are locked in storage until the day of the test administration and are signed out when they are to be used and signed in when they are returned. These materials are not to be left unattended at any time.

2.3.2 Monitoring Test Administration

The Office of Assessment staff conducted on-site monitoring of test administrations to verify district and school compliance with policies and procedures outlined in the TAMs. Monitoring is defined as an unannounced visit to a selected school. Monitoring includes documenting the

school's adherence to test security guidelines and the appropriate use of accommodations as specified in a student's IEP or 504 Accommodation Plan.

This section describes how the SC READY assessments demonstrate adherence to AERA, APA, & NCME (2014) Standard 6.7, which states the following:

> Test users have the responsibility of protecting the security of test materials at all times. (p. 117)

Before each testing window, the Security Committee developed a list of school sites to be monitored. The list of prioritized sites was constructed based on data forensics with additional consideration of the other named sources of information. Qualitative and quantitative data were examined and compared. Data was used to determine improbable gains in test scores as well as an unusual number of erasures resulting in correct answers. Surveys, test security violations, training attendance records, calls, and emails were also considered in site selection. Individual members of the Security Committee were called upon to share their knowledge of the events that suggested a site's addition to the monitoring list.

Selected SCDE monitors were encouraged to identify additional schools near the selected sites for routine visits if it was deemed feasible. Visits to those additional schools were approached as an opportunity for school staff to clarify test practices, reinforce the culture of good administration procedures, and provide feedback to the SCDE Office of Assessment.

As monitors observed, they were guided by the Observation Checklist that is completed online in a Google Form. If a security violation was observed by the SCDE monitor, in some cases, a Test Security Violation Action Form was completed. In other cases, the SCDE monitor requested that the STC submit documentation to the chair of the Security Committee. At the end of the visit, the monitors may have discussed their findings with the STC and the school principal if time and circumstances allowed.

Monitors or monitoring teams made visits to school systems as assigned. The teams not only made visits to individual schools but also sometimes visited the district office to review school system testing plans and local procedures. Visits may have occurred at any time during the testing window, and more than one visit may have been made to a specific school if needed. Additional test observation or additional documentation review and interviews with personnel may have been needed. Each individual monitor or monitoring team completed a checklist to document observations and make general notes related to test administration policy and procedures that were observed during the monitored test administration. In addition to the checklist, a short summary of the visit was required.

Following the on-site monitoring visit, monitoring teams reviewed checklists and notes to discuss an appropriate follow-up. If, during a monitoring visit, a potential test security violation was observed, a test security violation report was completed. Next, notes were reviewed for schools that did not exhibit potential security violations; strengths and weaknesses of the processes and procedures observed were noted and suggestions for further action were

included on the checklist. Next, the pertinent facts of the monitoring visit were entered in the Office of Assessment Monitoring Database. Finally, when the test window ended, Test Security personnel held a monitoring and initial security debriefing with the testing program manager to discuss any information that may have impacted changes to the TAM or protocol or procedures for that particular test.

2.3.3 Systems for Protecting Data Integrity & Privacy

Data security policies and procedures are based on state laws, regulations, and federal policies, such as CEPA, COPPA, FERPA, and others. The state law for data use and governance policy, SC Code of Laws 1976 as amended, is provided in Section 59-1-490 under "Data Use and Governance Policy." Data security and policies are located on the SCDE Web site in a section entitled "Data Security and Privacy." The Web address is https://ed.sc.gov/data/data-security-privacy/.

The Web page contains 13 documents covering security policies and standards for different areas of operation, including *Data Protection and Privacy Policy; HR and Security Awareness Policy; Information Security, IT Compliance,* and *Threat Vulnerability Management Policy.*

All personally identifiable information (PII) is stored on secure servers at SCDE under a PII policy (CE2.6B.RES – CE2.6G.RES). When reporting data, the website notes that all cells where the *N* count is less than 10 students are suppressed.

2.4 Summary

In summary, the overall purpose of this section is to explicate the procedures used in the development and administration of the SC READY assessments. The efforts by SCDE and DRC in developing the SC READY assessments are in alignment with multiple best practices of the assessment industry but, in particular, support the following AERA, APA, & NCME (2014) standards:

- Standard 3.1—Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
- 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.
- Standard 3.9—Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.
- 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.

- 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).
- Standard 4.7—The procedures used to develop, review, and try out items and to select items from the item pool should be documented.
- Standard 4.12—Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.
- 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.
- 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 specification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.
- 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.
- Standard 6.4—The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.
- 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.
- Standard 6.7—Test users have the responsibility of protecting the security of test materials at all times.
- Standard 7.2—The population for whom a test is intended and specifications for the test should be documented. If normative data are provided, the procedures used to gather the data should be explained; the norming population should be described in terms of relevant demographic variables; and the year(s) in which the data were collected should be reported.

Section 3 – Technical Quality (Validity)

3.1 Validity Evidence

The *Standards for Educational and Psychological Testing* defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA, APA, & NCME, 2014, p. 11). The purpose of test score validation is not to validate the test itself but to validate interpretations of the test scores for particular purposes or uses. Test score validation is not a quantifiable property but an ongoing process, beginning at initial conceptualization and continuing throughout the entire assessment process. Every aspect of an assessment provides evidence that either supports or challenges its validity, including design, content area specifications, item development, psychometric quality, and inferences made from the results.

Validity is the overarching component of the SC READY assessments. The following excerpt is from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014):

> Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence . . . include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question. (p. 22)

The validity of score interpretations for the SC READY assessments is supported by multiple sources of evidence. Section 1 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) specifies the following sources of validity evidence that are important to gather and document in order to support validity claims for an assessment:

- Test content
- Response processes
- Internal test structure
- Relation to other variables
- Consequences of test use

It is important to note that these categories are not mutually exclusive. One source of validity evidence often falls into more than one category, as discussed in more detail in this section. The process of gathering evidence of the validity of score interpretations is best characterized as ongoing throughout test development, administration, scoring, reporting, and beyond. As the technical report has progressed, it has covered the different phases of the testing cycle. Each part of the technical report has detailed the procedures and processes applied in South Carolina and the corresponding results. Each part has also highlighted the meaning and significance of the procedures, processes, and results in terms of validity and their relationship to specific

sections of the *Standards*. The current section now addresses these issues in validity: test content, response processes, internal test structure, relation to other variables, and consequences of test use.

3.2 Minimization of Construct-Irrelevant Variance and Construct Underrepresentation

Minimization of construct-irrelevant variance and construct underrepresentation is addressed in the following steps of the test development process: 1) specification, 2) item writing, 3) review, 4) field-testing, 5) test construction, and 6) item calibration. Section 2 contains more information on steps 1 through 5 and Section 4 contains more information on calibration.

Construct-irrelevant variance refers to error variance that is caused by factors unrelated to the constructs measured by the test. For example, when tests are not administered under standardized conditions (e.g., one administration may be timed, but another administration may not be timed), differences in student performance may be partially associated with the different administration conditions. Careful specification of content and review of the items representing that content are the first steps in minimizing construct-irrelevant variance. Then, empirical evidence, especially item-level data, is used to infer construct irrelevance.

Construct underrepresentation occurs when the content of the assessment does not reflect the full range of content that the assessment is expected to cover. Specification and review, in which test blueprints are developed and reviewed, are primary steps in the development process and are designed to ensure that content is appropriately represented.

3.3 Overall Validity, Including Validity Based on Content

According to the *Standards*, evidence based on test content "can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores" (AERA, APA, & NCME, 2014, p. 14). Documentation of the content domains, how the content is sampled and represented, and alignment of items to the content were discussed in Section 2. The documentation showed how test specification documents, which were derived from earlier developmental activities, guided the final phases of test development and ultimately yielded the test forms that were administered to students.

Section 2 also showed that the participation of South Carolina educators in that process provided a solid rationale for having confidence in the content and design of the SC READY assessments as a tool from which to derive valid inferences about South Carolina student performance. The test development process and the involvement of South Carolina educators in that process formed an important part of the validity of the SC READY ELA and mathematics assessments.

3.4 Validity Based on Cognitive Processes

Validity evidence based on response process relies to large degree on the evaluation of the cognitive processes of examinees responding to various types of items and the relationship between these processes and the construct being measured. Direct evidence based on response processes typically comes from analyses of individual responses or from test takers from various groups making up the intended test-taking population about their performance strategies or responses to specific items (AERA, APA, & NCME, 2014, p. 15). Such evidence can be gathered through cognitive labs conducted as part of the field test data analysis. Validity evidence based on response process is also supported by a relationship between the item type, format, and content and the construct being measured. For example, if a test is intended to measure a certain set of skills, it is important to determine whether the items included in the test are, in fact, designed to measure these skills or knowledge. As discussed in Section 2.1, SC READY items go through internal review processes within DRC followed by review by the SCDE.

3.5 Validity Based on Internal Structure—Construct Validity

The term "construct validity" refers to the degree to which the test score is a measure of the educational domain (i.e., construct) of interest. A construct is an individual characteristic that is assumed to exist to explain some aspect of behavior (Linn & Gronlund, 1995). When an individual characteristic from the assessment results is inferred, a generalization or interpretation of some construct is made. For example, problem-solving is a construct. An inference that students who master the mathematical reasoning portion of an assessment are "good problem-solvers" implies an interpretation of the results of the assessment in terms of a construct. It is important to demonstrate that it is a reasonable and valid use of the results to make such an inference.

Validity evidence based on internal test structure refers to the fact that "analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, & NCME, 2014, p. 16). Such analyses may include statistical analyses of items and subscores conducted to investigate the dimensionality of an assessment. Procedures for gathering such evidence may include factor analysis for single assessments. Internal test structure can also be evaluated using indices of measurement precision such as test reliability, decision accuracy and consistency, generalizability coefficients, and standard errors of measurement. Evaluation of the correlation coefficients that measure the relationship between the content area strand (domain) scores and studies of whether test items may function differently for different subgroups of students are additional sources of validity evidence based on internal test structure.

The collection of construct-related evidence is a continuous and ongoing process, and construct-related validity evidence can come from many sources. The *Standards* (AERA, APA, & NCME, 2014) provides the following list of possible sources:

- High intercorrelations among assessment items or tasks attest that the items are measuring the same trait, such as a content area objective, sub-domain, or construct.
- Substantial relationships between the assessment results and other measures of the same defined construct.
- Little or no relationship between the assessment results and other measures that are clearly not of the defined construct.
- Substantial relationships between different methods of measurement regarding the same defined construct.
- Relationships to non-assessment measures of the same defined construct.

Four indicators of construct validity for the spring 2022 SC READY assessment are item-total correlations, Rasch item fit statistics, reporting category intercorrelations, measurement invariance, and test dimensionality. The psychometric details of each are discussed in Section 4. The following summarizes the findings in terms of the value each has in supporting the validity of the SC READY assessments.

3.5.1 Item-Total Correlations

An item-total correlation is the correlation between an item and the total test score, excluding that item score. Conceptually, if an item has a high item-total correlation (i.e., 0.40 or above), it indicates that students who performed well on the test overall usually answered the item correctly and students who performed poorly on the test overall usually answered the item incorrectly. That is, the item did a good job discriminating between high performing and low performing students. Assuming the total test score represents the extent to which a student possesses the construct being measured by the test, high item-total correlations indicate that the items on the test require knowledge of this construct to be answered correctly. Item-total correlations for items across the SC READY spring 2022 administration can be found in Section 4.6.1.3. Most items have item-total correlations over 0.30 as shown in Table 3.1. These high item-total correlations provide evidence for construct validity.

Table 3.1. Item-Total Correlation Summary for SC READY ELA and Mathematics

Content Area	Grade	N Items	N Items Rit >0.30	% Items Rit>0.30
ELA	3	59	57	96.61
	4	59	57	96.61
	5	59	56	94.92
	6	61	59	96.72
	7	61	59	96.72
	8	61	57	93.44
Mathematics	3	50	48	96.00
	4	56	54	96.43
	5	56	53	94.64
	6	64	57	89.06
	7	64	55	85.94

Content Area	Grade	N Items	N Items Rit >0.30	% Items Rit>0.30
	8	67	61	91.04

3.5.2 Fit Statistics and Model Fit

In addition to item-total correlations, Rasch fit statistics also provide good evidence of construct validity. The Rasch model requires unidimensional data. Therefore, statistics showing that the items fit the measurement model also provide evidence of construct validity. Fit statistics for the spring 2022 SC READY assessments can be found in Section 4.6.2.2. The spring 2022 SC READY item fit information is provided to aid in interpretation. The vast majority of items on both the SC READY ELA and mathematics forms had infit and outfit mean square statistics within the acceptable range of 0.7 to 1.3. Items that fell outside of that range were further reviewed by DRC psychometric staff. The Rasch model-data fit indicates further evidence of good construct validity.

In addition to fit statistics, Section 4.6.2.3 also examines residual item correlations to assess the local dependence among SC READY ELA and mathematics items. Most of the correlations are very small, suggesting local item independence generally holds for the SC READY ELA and mathematics tests. The assumption of local independence being met is additional evidence of the construct validity of the SC READY ELA and mathematics assessments.

3.5.3 Reporting Category Intercorrelations

A third indicator of construct validity is the intercorrelations between the content area total scale scores and the subscale reporting category scale scores. This information is contained in Section 4.6.3 and is reported by subject. Moderate correlations were observed for all pairs of reporting categories across all grades. However, the correlation between two reporting category subscores may be artificially low because of measurement error. The intercorrelation corrected for attenuation was also examined, and the domain scores were found to be highly related, which also supports the validity of the SC READY ELA and mathematics assessments.

3.5.4 Item Distribution Across Content Domains

The SC READY operational and implementation test forms were constructed according to the test specifications and the test blueprints. These items measured the specific assessment standards that were approved by the SCDE. All items in the test forms were reviewed by the content review committee and the sensitivity review committee and were approved by the SCDE. Additional details of the item distributions across content area domains are provided in Section 2.1.1 and Section 4.6.3.1.

3.5.5 Validity Evidence for Measurement Invariance Different Student Populations

The primary evidence for the validity of the SC READY assessments lies in the content and constructs being measured. Because the test assesses the statewide content area standards required to be taught to all students, the test should not be more or less valid for use with one subpopulation of students over another subpopulation. In other words, because the SC READY

assessments are measuring what is required to be taught to all students and are given under the same standardized conditions to all students, the validity of score interpretations should apply to all students. A summary of student demographic information for the SC READY spring 2022 administration is presented in Tables 1.1 through 1.6.

A summary of student accommodation information is presented in Section 5.3. Great care has been taken to ensure that the items included on the SC READY assessments are fair and representative of the content area domain expressed in the content area standards. Much scrutiny is applied to the items and their possible impact on minority or other subpopulations making up the population of South Carolina. Every effort is made to eliminate items that may have gender, ethnic, or cultural biases. Section 2.1.4 contains discussion of how potential item bias is identified.

3.5.6 Dimensionality Assessment

Evidence presented in Section 4.4 assesses the dimensionality of the SC READY assessments using Principal Components Analysis (PCA). The findings support the claim that there is a dominant dimension underlying the items/tasks in each test and that scores from each test represent performance that is primarily determined by that ability. Construct-irrelevant variance, such as factual knowledge that is irrelevant to doing well in a subject, does not appear to create significant nuisance factors.

3.6 Validity Based on Relations to Other Variables

The SC READY test score relationship with other variables was examined to further support the validity of the intended score interpretation. This was done using two measures: evaluation of correlations between the SC READY content area scores and other South Carolina assessments and comparisons of the percentages of students classified in different proficiency levels (impact data) on the state assessment and on the NAEP assessment.

3.6.1 Correlations between Content Area Test Scores

Measures of different constructs should not be highly correlated with each other. The relationship between the scores from tests measuring different constructs can be assessed by the extent to which measures of constructs that theoretically should not be related to each other are, in fact, observed as not related to each other. Typically, correlation coefficients among measures of unrelated or distantly related constructs are examined in support of divergent evidence.

To assess the relationship between the SC READY content area scores and the scores on other South Carolina assessments, the correlations between SC READY ELA and mathematics scale scores and SC READY ELA and mathematics and SCPASS science scores were calculated for students who took two or three subject area tests in 2022. These correlations were based on the reportable census data, and the results are shown in Table 3.2. For the total population of students, the correlation coefficients ranged from 0.77 to 0.81 between SC READY ELA and SC READY mathematics scale scores. Comparing SC READY ELA to SCPASS, the correlation

coefficients was 0.84 for science. When comparing SC READY mathematics to SCPASS, the correlation coefficients ranged from 0.82 to 0.83 for science.

Despite moderate to high correlations, the tests are not perfectly related to each other, suggesting that different constructs are being tapped; however, the test scores do appear to be highly related to one another, suggesting they may be tapping into a similar knowledge base or general underlying ability.

Table 3.2: Inter-Correlations of SC Assessment Content Areas

Content Area	Grade	SC READY Mathematics	SC PASS Science
SC READY ELA	3	0.79	–
	4	0.78	0.84
	5	0.79	–
	6	0.80	0.84
	7	0.81	–
	8	0.77	–
SC READY Mathematics	3	–	–
	4	–	0.82
	5	–	–
	6	–	0.83
	7	–	–
	8	–	–

3.6.2 Comparison of the SC READY and South Carolina NAEP Impact Data

The NAEP is the largest nationally representative and continuing assessment of what America's students know and can do in various subject areas. Assessments in several content areas, including reading, mathematics, and science, are periodically administered to students in grades four, eight, and twelve. Representative samples of students from different states, including South Carolina, participated in the latest NAEP assessments, which occurred in spring 2019 for reading and mathematics and in spring 2015 for science.

The main NAEP assessments are constructed using detailed frameworks that result from a comprehensive national process in which teachers, curriculum experts, policymakers, and members of the public work to create a unified vision of how each subject should be assessed. This vision is based on current educational research on achievement and its measurement; it is also based on good educational practices. These frameworks are updated about every decade to keep them current (for details, refer to https://nces.ed.gov).

The NAEP results are reported for all assessed content areas and for all participating grades at the national level. At the state level, the results for reading, mathematics, science, and writing are reported for grades four and eight. The results may also be reported at the district level (within a state) for these four content areas. No results are reported at the student level.

South Carolina students participated in the latest reading and mathematics assessment in spring 2022. The SC READY ELA and mathematics assessment results are compared to the NAEP results in grades four and eight. The percentages of South Carolina students classified in different proficiency levels on the SC READY ELA and mathematics assessments and the corresponding NAEP assessments are presented in Table 3.3.

As presented in Table 3.3, for both ELA and mathematics, lower percentages of students were classified in the *Proficient or above* level in NAEP grades four and eight compared to the *Meets Expectations or above* level in SC READY grades four and eight. For ELA, the difference was approximately 10% for grade four and 17% for grade eight. For mathematics, the difference was approximately 9% for grade four and 8% for grade eight.

Similarly, as shown in Table 3.3, lower percentages of students in grades four and eight were classified in the *Basic or above* level in NAEP Reading compared to the *Approaches Expectations or above* level in SC READY ELA. For ELA, the difference was approximately 11% for grade four and 10% for grade eight. For mathematics, lower percentages of students in grades four and eight were classified in the *Basic or above* level in NAEP compared to the *Approaches Expectations or above* level in SC READY. The difference was approximately 3% for grade four and 3% for grade eight.

Table 3.3. Comparison of 2022 South Carolina NAEP and Spring 2022 SC READY Impact Data

Content Area	Grade	NAEP 2022			SC READY 2022		
		At or Above Basic	**At or Above *Proficient***	**At Advanced**	**At or Above *Approaches***	**At or Above *Meets***	**At *Exceeds***
Reading/ELA	**4**	60.9	32.5	8.7	72.0	50.4	30.6
Reading/ELA	**8**	62.7	26.6	2.4	72.8	45.9	16.7
Mathematics	**4**	74.4	34.0	6.3	71.3	43.4	23.7
Mathematics	**8**	56.4	22.0	5.2	59.2	30.2	16.0

3.7 Evidence Based on the Consequences of Test Use

The *Standards* incorporates the intended and unintended consequences of test use into the concept of validity. It indicates that information about the consequences of testing does not in and of itself detract from the validity of intended test interpretations (AERA, APA, & NCME, 2014, p. 19). Rather, according to the *Standards*, a more searching inquiry into the sources of those consequences given the intended purposes of an assessment is a basis for evaluating the quality of the validity evidence. The test data alone do not provide sufficient verification of this type of evidence. For this reason, it is not straightforward to measure and collect evidence on the consequential aspects of validity.

To address the intended consequences of the SC READY assessments, the purposes of the assessments must be specified. SCDE has carefully articulated the intended purposes of SC READY as driving features of the development of the ELA and mathematics tests and the implementation of the testing program. The specific purposes associated with SC READY include the following:

- SC READY accurately describes both student achievement (i.e., how much students know at the end of the year) to inform program evaluation and school, district, and state accountability systems and to provide valid, reliable, and fair measures of students' progress toward, and attainment of, the knowledge and skills required to be "Meet" the South Carolina content area standards.
- SC READY informs state and federal accountability.
- SC READY assessments are fair for all students, including those with disabilities or limited English proficiency, at all levels of achievement.

3.8 Summary

In summary, most sections of this technical report are designed to provide validity evidence to support the use and intended interpretation of the SC READY ELA and mathematics test scores. SC READY scores are used to identify strengths and areas for improvement in South Carolina's student performance; to inform stakeholders (teachers, school administrators, district administrators, SCDE staff members, parents, and the public) about the status of the progress toward meeting the academic performance standards of the state; and to meet the requirements of the state's accountability program.

Section 4 – Technical Quality (Other)

Evidence for construct-related validity—the meaning of test scores and the inferences they support—is the central concept underlying the SC READY ELA and mathematics assessment validation process. In this section, DRC presents additional evidence of construct-related validity, which includes the minimization of construct-irrelevant variance and construct-underrepresentation in the test development process, as well as through studies of internal consistency, psychometric analyses of fairness, model fit, dimensionality analyses, analyses by reporting category, and scale evaluation and model fit. All analyses in this section are based on final data used after standard setting.

Section 4 of this report demonstrates the SC READY ELA and mathematics assessments' adherence to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). Section 4 is related to Standards 1.8, 1.13, 1.21, 2.0, 2.3, 2.11, 2.13, 2.14, 2.16, 2.19, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 4.14, 4.18, 4.20, 5.2, 5.13, 5.15, 6.8, 6.9, and 7.2. Each standard will be discussed in the pertinent section.

4.1 Reliability

Reliability refers to the consistency of students' test scores on parallel forms of a test. A reliable test is one that produces scores that are expected to be relatively stable if the test is administered repeatedly under similar conditions. Often, however, it is impractical to administer multiple forms of the test, and reliability is estimated on a single administration of the test instead. This type of reliability, known as internal consistency, provides an estimate of how consistently examinees perform across items within a test during a single test administration (Crocker & Algina, 1986). Reliability is a necessary but not sufficient condition of validity.

The AERA, APA, & NCME (2014) *Standards* states the following:

> The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency). (p. 33)

In accordance with the AERA, APA, & NCME (2014) *Standards* and in order to develop and maintain tests of the highest quality, DRC has calculated the reliability of each SC READY ELA and mathematics test in a variety of ways: reliability of raw scores, overall standard error of measurement (SEM), IRT-based conditional standard error of measurement (CSEM), and decision consistency of performance level classifications.

There are several specific AERA, APA, & NCME (2014) Standards that this section addresses. These include Standards 2.0, 2.3, 2.13, and 2.19, which are included below.

Standard 2.0 states the following:

> Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use. (p. 42)

Standard 2.3 states the following:

> For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported. (p. 43)

The total score reliabilities are discussed in Section 4.1.1 of this report. The SEM of the total score is discussed in Section 4.1.3.

AERA, APA, & NCME (2014) Standard 2.13 states the following:

> The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (p. 45)

The raw score-based SEM is discussed in Section 4.1.3 and is reported in raw score units. The CSEM is discussed in Section 4.1.4 and is presented in scale score units.

Standard 2.19 states the following:

> Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported. (p. 47)

Section 4.1.4 discusses different ways of measuring test reliability, including reliability of raw scores and test form SEM, IRT-based CSEM, and decision consistency of performance level classifications. These statistics were computed based on the data used for operational analyses.

4.1.1 Test Reliability

True-score theory considers all measures as having a "true" component and an error component. Errors occur as a natural part of the measurement process and can never be eliminated entirely. For example, uncontrollable factors such as differences in the physical world and changes in examinee disposition may work to increase error and decrease reliability. This is the fundamental premise of true-score reliability analysis and measurement theory. Stated explicitly, this relationship can be shown as follows:

$$X = T + E, \tag{4.1}$$

where *X* represents the observed test score, *T* represents the student's true score, and *E* represents random error.

If the variance of the observed measures is denoted by σ_X^2 and the variance of error is denoted by σ_E^2, then the reliability (ρ_{xx}) is given by

$$\rho_{XX} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}. \quad (4.2)$$

The variance of the observed measures can be estimated from the variance of the raw scores using the usual variance formula, and the error variance can be estimated by

$$\Sigma\, p_i\,(1\text{-}p_i), \quad (4.3)$$

where p_i is the proportion correct for each item.

The reliability index used for the 2022 administration of the SC READY assessments was the Coefficient Alpha (Cronbach, 1951):

$$\alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right), \quad (4.4)$$

where *k* is the number of items, σ_i^2 is the variance of the set of scores associated with item *i*, and σ_X^2 is the variance of the set of observed total scores.

Acceptable α values generally range in the high 0.80s to low 0.90s. When there is no error, the reliability index reduces to the true score variance divided by the true score variance, which is one.

Tables 4.1 to 4.3 show the test form reliability coefficients and standard error of measurement (SEM) by content area and grade for student race/ethnicity, student gender, student English proficiency status, student disability status, and whether a student used any testing accommodations. Note that for spring 2022, the accommodated and non-accommodated forms for ELA consisted of the same items. Additionally, the mathematics accommodated and non-accommodated forms for grades three through five consisted of the same items.

The overall reliability coefficients for the SC READY assessments are reported in Tables 4.1 through 4.3 and ranged from 0.94 to 0.95 for ELA and was 0.94 for all grades in mathematics. These results indicate acceptable reliability coefficients for the SC READY assessments.

4.1.2 Test Reliability by Subgroup

AERA, APA, & NCME (2014) Standard 2.11 states the following:

> Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended. (p. 45)

The reliability coefficients by subgroup, reported in Tables 4.1 through 4.3, ranged from 0.92 to 0.95 for ELA. and from 0.89 to 0.94 for mathematics across all grades, forms, accommodation statuses, and subgroups. The analysis of the test reliability by subgroup shows that the test reliability is of acceptable magnitude for all subgroups.

4.1.3 Standard Error of Measurement

The standard error of measurement uses the information from the test along with an estimate of reliability to make statements about the degree to which error is impacting individual scores. The standard error of measurement is based on the premise that underlying traits, such as academic achievement, cannot be measured exactly. The standard error expresses unreliability in terms of the raw score metric. With the standard error of measurement, an error band can be placed around an individual score, indicating the degree to which error might be affecting that score. In true-score test theory, the standard error of measurement can be calculated by

$$SEM = \sigma_X \sqrt{1 - \rho_{XX}}\ , \tag{4.5}$$

where σ_X is the standard deviation of the total test (observed measure scores) and ρ_{XX} is the reliability estimate (Coefficient Alpha) for the test.

The true-score test theory approach to judging a test's consistency can be useful for making overall comparisons between alternate forms. However, it is not very useful for judging the precision with which a specific student's score is known. The Rasch measurement model provides asymptotic standard errors that pertain to each unique ability estimate (i.e., raw score).

Table 4.1. Classical Reliability Indices and SEM by Subgroup for ELA Forms

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
3	All Students	55,905	0.95	3.35
	Female	27,310	0.94	3.32
	Male	28,337	0.95	3.36
	Black or African American	17,226	0.93	3.53
	White	26,782	0.94	3.17
	Hispanic or Latino	6,632	0.94	3.47
	ML	2,367	0.93	3.53
	SWD	9,116	0.94	3.53
4	All Students	56,397	0.95	3.42
	Female	27,663	0.94	3.39

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
	Male	28,486	0.95	3.43
	Black or African American	17,434	0.93	3.56
	White	26,891	0.94	3.27
	Hispanic or Latino	6,798	0.94	3.50
	ML	2,339	0.93	3.54
	SWD	9,457	0.94	3.55
5	All Students	57,065	0.94	3.46
	Female	27,753	0.94	3.44
	Male	29,077	0.94	3.47
	Black or African American	17,774	0.92	3.57
	White	27,280	0.93	3.34
	Hispanic or Latino	6,932	0.93	3.54
	ML	2,308	0.92	3.59
	SWD	9,904	0.93	3.58
6	All Students	57,566	0.94	3.49
	Female	28,133	0.94	3.44
	Male	29,106	0.94	3.51
	Black or African American	18,291	0.93	3.63
	White	27,101	0.94	3.34
	Hispanic or Latino	7,108	0.94	3.54
	ML	2,352	0.92	3.62
	SWD	9,917	0.93	3.64
7	All Students	29,528	0.94	3.56
	Female	60,109	0.94	3.62
	Male	30,266	0.94	3.64
	Black or African American	19,118	0.93	3.73
	White	28,740	0.94	3.48
	Hispanic or Latino	7,054	0.93	3.72
	ML	2,006	0.90	3.81
	SWD	9,987	0.92	3.73
8	All Students	60,920	0.94	3.64
	Female	29,874	0.94	3.56
	Male	30,734	0.94	3.68
	Black or African American	19,842	0.93	3.73
	White	28,762	0.94	3.49
	Hispanic or Latino	7,332	0.94	3.70

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
	ML	2,366	0.92	3.77
	SWD	10,118	0.93	3.74

Note: ALL = All students who attempted the test except home school students and students who used Braille or sign language test booklets.

Table 4.2. Classical Reliability Indices and SEM by Subgroup for Mathematics Grades 3-5 And Non-Accommodated Forms Grades 6-8

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
3	All Students	55,896	0.94	2.96
	Female	27,308	0.93	2.97
	Male	28,324	0.94	2.93
	Black or African American	17,211	0.91	3.14
	White	6,633	0.93	3.05
	Hispanic or Latino	3,311	0.93	2.97
	ML	2,356	0.93	3.04
	SWD	9,095	0.93	3.09
4	All Students	56,404	0.94	3.18
	Female	27,663	0.93	3.20
	Male	28,495	0.94	3.15
	Black or African American	17,432	0.90	3.33
	White	6,799	0.93	3.25
	Hispanic or Latino	3,274	0.93	3.20
	ML	2,338	0.92	3.26
	SWD	9,457	0.92	3.29
5	All Students	57,066	0.94	3.19
	Female	27,760	0.93	3.21
	Male	29,071	0.94	3.17
	Black or African American	17,773	0.91	3.32
	White	6,930	0.93	3.25
	Hispanic or Latino	3,163	0.94	3.21
	ML	2,306	0.92	3.29
	SWD	9,906	0.92	3.28
6	All Students	48,041	0.94	3.37
	Female	23,961	0.93	3.39

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
	Male	23,819	0.94	3.34
	Black or African American	14,488	0.89	3.50
	White	5,673	0.92	3.43
	Hispanic or Latino	2,590	0.93	3.40
	ML	2,035	0.89	3.51
	SWD	5,279	0.92	3.44
7	All Students	50,998	0.93	3.38
	Female	25,480	0.92	3.40
	Male	25,261	0.93	3.35
	Black or African American	15,443	0.88	3.50
	White	5,800	0.91	3.44
	Hispanic or Latino	2,689	0.92	3.41
	ML	1,723	0.87	3.52
	SWD	5,829	0.91	3.45
8	All Students	52,370	0.94	3.42
	Female	26,055	0.93	3.42
	Male	26,049	0.94	3.40
	Black or African American	16,268	0.90	3.52
	White	6,056	0.93	3.47
	Hispanic or Latino	2,509	0.93	3.43
	ML	2,097	0.90	3.53
	SWD	6,240	0.92	3.47

Note: ALL = All students who attempted the test except home school students and students who used Braille or sign language test booklets.

Table 4.3. Classical Reliability Indices and SEM by Subgroup for Mathematics Grades 6-8 Accommodated Forms

Grade	Subgroup	*N* Count	Cronbach's Alpha	SEM
6	All Students	9,621	0.94	3.37
	Female	4,227	0.94	3.38
	Male	5,323	0.94	3.37
	Black or African American	3,862	0.91	3.44
	White	1,444	0.90	3.44
	Hispanic or Latino	462	0.94	3.38
	ML	321	0.89	3.47
	SWD	4,675	0.83	3.43
7	All Students	9,155	0.93	3.40
	Female	4,074	0.93	3.40
	Male	5,019	0.93	3.39
	Black or African American	3,707	0.89	3.47
	White	1,256	0.88	3.46
	Hispanic or Latino	409	0.93	3.40
	ML	284	0.80	3.49
	SWD	4,172	0.81	3.46
8	All Students	8,617	0.94	3.43
	Female	3,854	0.94	3.44
	Male	4,713	0.94	3.41
	Black or African American	3,627	0.90	3.50
	White	1,285	0.90	3.49
	Hispanic or Latino	387	0.94	3.45
	ML	271	0.85	3.52
	SWD	3,898	0.85	3.45

Note: ALL = All students who attempted the test except home school students and students who used Braille or sign language test booklets.

4.1.4 Conditional Standard Error of Measurement

In contrast to the SEM, the conditional standard error of measurement (CSEM) expresses the degree of measurement error in scale score units and is conditioned on the ability of the student. The CSEM is reported in support of AERA, APA, & NCME (2014) Standard 2.14, which states the following:

> When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is

constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. (p. 46)

The CSEM of each cut score is reported in Tables 4.4 and 4.5.

The CSEMs are defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985)

$$\mathrm{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}, \tag{4.6}$$

where $I(\vartheta_i)$ is the test information function, as a sum of item information function 2, obtained as

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i)\, q_{ij}(\theta_i)}, \tag{4.7}$$

where $p'_{ij}(\theta_i)$ is the derivative of $p_{ij}(\theta_i)$ and $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$. Note that the CSEMs vary in magnitude across the entire range of student ability estimates (i.e., scale scores) and are lower in the middle of the score distribution and higher at the tails. This pattern is seen for all SC READY CSEMs and is to be expected when IRT methods are used. The CSEMs at the three cut scores that define the performance levels are presented in Table 4.4 for ELA and Table 4.5 for mathematics. For ELA, the CSEM at the *Approaches* cut score ranged between 24 and 25 scale score points, the CSEM at the *Meets* cut score ranged between 26 and 28 scale score points, and the CSEM at the *Exceeds* cut score ranged between 31 and 34 scale score points across all grades. For mathematics, the CSEM at the *Approaches* cut score ranged between 23 and 25 scale score points, the CSEM at the *Meets* cut score ranged between 24 and 26 scale score points, and the CSEM at the *Exceeds* cut score ranged between 28 and 34 scale score points across all grades.

Table 4.4. CSEM at SC READY Scale Score Cuts, ELA

Grade	Form	*Approaches*	*Meets*	*Exceeds*
3	Online	25	28	34
4	Online	25	27	31
5	Online	25	26	31
6	Online	25	27	33
7	Online	24	26	32
8	Online	24	27	34
3	Paper-and-pencil	25	28	34
4	Paper-and-pencil	25	27	31

Grade	Form	*Approaches*	*Meets*	*Exceeds*
5	Paper-and-pencil	25	26	31
6	Paper-and-pencil	25	27	33
7	Paper-and-pencil	24	26	32
8	Paper-and-pencil	24	27	34

Table 4.5. CSEM at SC READY Scale Score Cuts, Mathematics

Grade	Form	*Approaches*	*Meets*	*Exceeds*
3	Online	25	26	34
4	Online	25	25	29
5	Online	24	24	30
6	Online	23	24	29
7	Online	23	24	28
8	Online	23	24	28
3	Paper-and-pencil	25	26	34
4	Paper-and-pencil	25	25	29
5	Paper-and-pencil	24	24	30
6	Paper-and-pencil	23	24	29
7	Paper-and-pencil	23	24	28
8	Paper-and-pencil	23	24	29

4.2 Indicators of Consistency

Classification Consistency: Classification consistency (also known as decision consistency) is defined as the extent to which the classifications of students agree on the basis of two independent administrations of the test or one administration of two parallel test forms. It is difficult, however, to obtain data from repeated administrations of the same form because of cost, time, and students' recall of the first administration. Also, it is difficult to construct two parallel forms. A common practice, therefore, is to estimate decision consistency from one administration of a test. These analyses directly address AERA, APA, & NCME (2014) Standard 2.16:

> When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure. (p. 46)

Classification Accuracy: Classification accuracy is defined as the extent to which the actual classifications of test takers agree with classifications that would be made based on the test takers' true scores (Livingston & Lewis, 1995). It is common to estimate classification accuracy by utilizing a psychometric model to find true scores corresponding to observed scores.

4.2.1 Classification Consistency Index

SC READY adheres to an extension of the two-parameter beta-binomial model (Huynh, 1976) to polytomous constructed-response items. This extension was used in these computations. Table 4.6 depicts the general framework of multiple decisions.

Table 4.6. Multiple Decisions—General Framework

	Category 1	Category 2	Category 3	Category 4	Total
Category 1	p_{11}				$p_{1.}$
Category 2		p_{22}			$p_{2.}$
Category 3			p_{33}		$p_{3.}$
Category 4				p_{44}	$p_{4.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{.4}$	$p_{..}$

From this general framework, the reliability index can be computed:

$$\kappa = \frac{1-p}{p-p_c}, \tag{4.8}$$

where $p = p_{..}$,

$$p_c = \sum_i p_{i.}^2, \tag{4.9}$$

$$p_{11} = \sum_{x,y=c_1}^{n} f(x,y), \tag{4.10}$$

and

$$p_1 = \sum_{x=c_1}^{n} f(x). \tag{4.11}$$

4.2.2 Classification Accuracy Index

To solve the problem of a complex assessment, Livingston and Lewis (1995) proposed an effective test length

$$n = \frac{(\mu_x - X_{\min})(X_{\max} - \mu_x) - r\sigma_x^2}{\sigma_x^2(1-r)}, \tag{4.12}$$

which transforms the original raw score random variable from $X = 0,...,K$ into a new random variable $X' = 0,...,n$, where *n* is the number of dichotomous, locally independent, equally difficult items required to produce a raw score of the same reliability. Then, using the transformed observed distribution X', parameters are estimated for a four-parameter beta-binomial model where the conditional error distribution is assumed to be binomial. The X' distribution is then converted back onto the original *X* scale using interpolation. This method is designed only to estimate a contingency table, not a full bivariate distribution, which means the probability of a consistent decision by chance, and subsequently kappa, cannot be estimated.

The BB-Class (Brennan, 2004) program was used to calculate assessment consistency. The results of all consistency analyses are presented in Table 4.7.

Table 4.7. Decision Consistency Indices

Content Area	Grade	Huynh				Livingston Lewis	
		Two Achievement Levels		Four Achievement Levels		Two Achievement Levels	Four Achievement Levels
		Proportion of Agreement	Kappa	Proportion of Agreement	Kappa	Proportion of Agreement	Proportion of Agreement
ELA	3	0.918	0.835	0.739	0.650	0.916	0.732
	4	0.921	0.842	0.761	0.678	0.920	0.759
	5	0.909	0.818	0.738	0.648	0.908	0.735
	6	0.913	0.824	0.741	0.651	0.912	0.738
	7	0.913	0.823	0.751	0.665	0.913	0.749
	8	0.914	0.826	0.721	0.623	0.912	0.715
Mathematics All Forms	3	0.912	0.823	0.742	0.655	0.911	0.741
	4	0.912	0.821	0.743	0.655	0.912	0.743
	5	0.913	0.822	0.751	0.665	0.913	0.752
Mathematics Non-Accommodated Forms	6	0.916	0.822	0.748	0.658	0.916	0.750
	7	0.915	0.807	0.743	0.643	0.915	0.744
	8	0.921	0.817	0.763	0.668	0.921	0.763
Mathematics Accommodated Forms	6	0.952	0.863	0.822	0.714	0.952	0.824
	7	0.945	0.839	0.808	0.692	0.946	0.811
	8	0.947	0.843	0.825	0.712	0.947	0.825

4.3 Reliability of Fairness & Accessibility

As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), there are varying definitions of fairness. In this section, we examine fairness as it relates to minimizing bias on a test. We then look at test performance among varying subgroups assessed by the ELA and mathematics tests. It should be noted that differences in test

performance among subgroups do not mean that a test is unfair—they simply mean that groups perform differently on the test. Even when a test is carefully and properly constructed, differences may exist among subgroups as a result of differences in curriculum or learning by students in the subgroup.

This section is particularly relevant to AERA, APA, & NCME (2014) Standards 3.1 through 3.6. These standards are from Chapter 3 of the AERA, APA, & NCME (2014) *Standards*, "Fairness in Testing." Each of these standards will be presented, as will the way the standard is addressed, in this section. Standard 3.6 states the following:

> Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws. (p. 65)

There is no particular research on the SC READY assessments showing that the test scores of examinee subgroups differ in meaning; however, this is an ongoing concern in any large-scale testing program. To lessen the possibility of differences in test score meaning, DRC has several steps that are followed in the item development and selection processes, as are explained in Section 4.3.1. In addition, SCDE and DRC conduct content and bias reviews on items, as explained in Section 2. These practices adhere to 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. (p. 64)

DRC conducts differential item functioning (DIF) studies following the operational administration of the SC READY assessments. Typically, items are evaluated for possible DIF in the field test phase of test development, and items flagged for DIF are typically further examined for possible bias. During the ELA and mathematics test development, DRC content area experts avoided including items that may potentially favor one demographic group over another. Section 4.3.2 explains the steps taken to evaluate SC READY items through the use of DIF in order to adhere to Standard 3.3.

In addition, standardized test administration and the training of test readers for SC READY comply with Standards 3.4 and 3.5:

> **Standard 3.4** Test takers should receive comparable treatment during the test administration and scoring process. (p. 65)

> **Standard 3.5** Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population. (p. 65)

Section 4.3.1 is directly relevant to Standards 3.1 and 3.2:

> **Standard 3.1** Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p. 63)
>
> **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. (p. 64)

In this section, we describe the steps taken by DRC to minimize words, phrases, and content that may be regarded as offensive by members of demographic subgroups. Section 2 discusses content considerations during item development and item reviews for items included in the SC READY assessments. These reviews are also critical in fulfilling the guidelines established in Standards 3.1 and 3.2.

4.3.1 Minimizing Bias through Test Development

The development of a test that is fair for all examinees begins in the early stages of planning and development. The item and test development processes that were used to minimize bias are summarized below.

First, careful attention was paid to content-related validity during the item development and item selection processes. Bias can occur only if the test is measuring different things for different groups. By eliminating irrelevant skills or knowledge that may be tested in the items, the possibility of bias is reduced.

Second, DRC and SCDE item writers followed DRC's internal bias and sensitivity guidelines to help ensure that the items are fair for all groups of test takers, despite differences in characteristics, including but not limited to disability status, ethnic group, gender, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Test developers reviewed all items included in the SC READY assessments and other testing materials with these guidelines in mind.

Finally, careful attention is typically given to item statistics (if available) throughout the test development process. As part of the test assembly process, attempts are made to avoid using or reusing items with poor statistical fit or distractors with positive point biserial correlations, since poor statistics may indicate that an item is tapping an ability that is irrelevant to the construct being measured. Additional steps to reduce bias, including the use of content and bias committees comprising South Carolina participants, are described in more detail in Section 2 of this report.

4.3.2 Evaluating Bias through Differential Item Functioning (DIF) Statistics

After administering the test, an empirical approach known as DIF was used to examine the items. The DIF statistics indicate the degree to which members of a particular subgroup perform better or worse than expected on each item as compared to the members of the reference group. The DIF procedures used and the results of these analyses are detailed in this section. It should be noted, though, that all items included on the SC READY ELA and mathematics assessments have been thoroughly reviewed for content and bias issues by South Carolina educators and DRC content area experts to ensure that they do not tap knowledge or specific abilities irrelevant to the construct the test intends to measure. Therefore, DIF flags do not necessarily indicate that an item is biased; rather, DIF flags indicate that the item functions differently for equally able members of different groups (Camilli & Shepard, 1994). Items are not necessarily suppressed from operational scoring if they are flagged for DIF.

The position of DRC concerning test bias is based on two general propositions. First, students may differ in their background knowledge, cognitive and academic skills, language, attitudes, and values. To the degree that these differences are large, no one curriculum and no one set of instructional materials will be equally suitable for all. Therefore, no one test will be equally appropriate for all. Furthermore, it is difficult to specify what amount of difference can be called large and to determine how these differences will affect the outcome of a particular test. Second, schools have been assigned the tasks of developing certain basic cognitive skills and supporting development of these skills equitably among all students. Therefore, there is a need for tests that measure the common skills and bodies of knowledge that are common to all learners. The test publisher's task is to develop assessments that measure these key cognitive skills without introducing extraneous or construct-irrelevant elements into the performances on which the measurement is based. If these tests require that students have culturally specific knowledge and skills not taught in school, differences in performance among students can occur because of differences in student background and out-of-school learning. Such tests are measuring different things for different groups and can be called biased (Camilli & Shepard, 1994; Green, 1975).

In order to lessen such biases, DRC strives to minimize the role of extraneous elements, thereby increasing the number of students for whom the test is appropriate. As discussed above, careful attention is given during the test development and test construction processes to lessen the influence of these elements for large numbers of students. Content and bias review committees are used to detect these elements and lessen their influence on students. Unfortunately, in some cases, these elements may continue to play a substantial role. To assess the extent to which items may be performing differently for various subgroups of interest, DIF analyses are conducted after each operational test administration.

DIF statistics are used to quantify differences in item performance between two groups after controlling for examinees' overall achievement level. Two DIF statistics that are commonly used for this purpose are the Mantel-Haenszel (MH) statistic (1959) and the standardized mean difference (SMD) between the reference and focal groups, proposed by Dorans and Schmitt (1991).

The MH procedure, as implemented by DRC, compared the observed and expected totals of a two-by-two-by-four contingency table (Holland & Thayer, 1986). The MH statistic is computed as follows (Zwick et al., 1993):

$$\text{Mantel } \chi^2 = \frac{\left(\sum_k F_k - \sum_k E(F_k)\right)^2}{\sum_k \text{Var}(F_k)}, \qquad (4.13)$$

where F_k is the sum of scores for the focal group at the kth level of the matching variable. Note that the MH statistic is sensitive to the case count such that larger sample sizes increase the value of chi-square.

In addition to the MH chi-square statistic, the delta statistic (MH-D DIF) was computed for all items. The Educational Testing Service first developed the MH-D DIF statistic.

To compute delta, alpha (the odds ratio) is first computed using the following equation:

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} N_{r1k} N_{f0k} / N_k}{\sum_{k=1}^{K} N_{f1k} N_{r0k} / N_k}, \qquad (4.14)$$

where N_{r1k} is the number of correct responses in the reference group at ability level k, N_{f0k} is the number of incorrect responses in the focal group at ability level k, N_k is the total number of responses, N_{f1k} is the number of correct responses in the focal group at ability level k, and N_{r0k} is the number of incorrect responses in the reference group at ability level k. MH-D DIF is then computed using the following equation:

$$\text{MH-D DIF} = -2.35 \ln(\alpha_{MH}). \qquad (4.15)$$

For selected-response items, the MH (χ^2_{MH}) statistic was used to evaluate potential DIF items. In the MH procedure, subgroups are matched by their raw total test scores using a contingency table with k ability levels. When applying the MH procedure, the log-odds ratio α is assumed to be constant across the K matched levels. The χ^2_{MH} , then, estimates a pooled common-odds ratio. Taking the natural logarithm of the common-odds ratio and its confidence limits and multiplying these by the constant −2.35 allows the resulting values to then be placed on the MH delta metric (Δ_{MH}) for interpretive purposes. Items were flagged for DIF using the following criteria:

- Moderate DIF: significant MH chi-square statistic ($p < 0.05$) and $1.0 \leq |\text{MH D-DIF}| < 1.5$
- Large DIF: significant MH chi-square statistic ($p < 0.05$) and $|\text{MH D-DIF}| \geq 1.5$

A positive DIF value indicates that the item favors the focal group, while a negative value indicates that the item disadvantages the focal group. DIF statistics were computed for the following subgroups:

- ***Gender:*** The focal group is females; the reference group is males.
- ***Race/Ethnicity:*** The focal groups are students whose race/ethnicity is reported as Black or African American, Hispanic, Asian, American Indian, or Two or More Races; the reference group is students whose race/ethnicity is reported as White.
- ***English Language Learner Status.*** The focal group are students whose EL status is reported as an English Language Learner; the reference group is a students whose EL status is reported as a Native English Speaker or English Proficient.
- ***Disability Status.*** The focal group are students whose EL status is reported as a student with a disability status of "Yes"; the reference group is students whose disability status is "No" or unknown.
- ***Form Type:*** The focal group is students who had taken the paper-and-pencil test form; the reference group is all others.

A negative SMD value implies that the focal group has a lower mean item score than the reference group, whereas a positive value implies that the focal group has a higher mean item score than the reference group, conditioned on the matching test score.

The minimum case count for the focal group was set at 200, and the minimum case count for the reference group was set at 400. The DIF analyses are not performed for subgroups of fewer than 200 students. In these cases, the statistical procedures do not have sufficient power to detect differences, should they exist.

Tables 4.8 and 4.9 summarize the number of moderate and large DIF flags by grade for each focal group that included at least 200 students for ELA and mathematics respectively. For example, consider grade four ELA. In this grade, two items were flagged for DIF in favor of the female group over the male subgroup (moderate DIF).

Overall, the number of items flagged for DIF in the SC READY spring 2022 tests was very small. Again, any items included on the SC READY assessments (including those items flagged for DIF) have been thoroughly reviewed for content and bias issues by South Carolina teachers, SCDE staff, and DRC test development experts.

Table 4.8. Operational DIF Summary, ELA

Grade	Reference Group	Focal Group	Total *N* of Items	DIF Classification				
				A	B+	B-	C+	C-
3	Male	Female	59	58	0	1	0	0
	White	Black or African American	59	58	1	0	0	0
	White	Hispanic	59	58	0	0	0	1
	White	Asian	59	56	0	1	1	1
	White	Two or More Races	59	59	0	0	0	0
	Non-ML	ML	59	58	0	0	0	1

Grade	Reference Group	Focal Group	Total *N* of Items	DIF Classification				
				A	B+	B-	C+	C-
	Non-SWD	SWD	59	59	0	0	0	0
	Paper-and-pencil	Online	59	58	0	1	0	0
4	Male	Female	59	58	0	1	0	0
	White	Black or African American	59	59	0	0	0	0
	White	Hispanic	59	59	0	0	0	0
	White	Asian	59	55	0	3	1	0
	White	Two or More Races	59	59	0	0	0	0
	Non-ML	ML	59	59	0	0	0	0
	Non-SWD	SWD	59	58	0	1	0	0
	Paper-and-pencil	Online	59	56	1	2	0	0
5	Male	Female	59	58	0	1	0	0
	White	Black or African American	59	59	0	0	0	0
	White	Hispanic	59	59	0	0	0	0
	White	Asian	59	55	2	1	0	1
	White	Two or More Races	59	59	0	0	0	0
	Non-ML	ML	59	58	0	1	0	0
	Non-SWD	SWD	59	59	0	0	0	0
	Paper-and-pencil	Online	59	59	0	0	0	0
6	Male	Female	61	57	2	2	0	0
	White	Black or African American	61	61	0	0	0	0
	White	Hispanic	61	61	0	0	0	0
	White	Asian	61	57	0	3	0	1
	White	Two or More Races	61	61	0	0	0	0
	Non-ML	ML	61	61	0	0	0	0
	Non-SWD	SWD	61	61	0	0	0	0
	Paper-and-pencil	Online	61	61	0	0	0	0
7	Male	Female	61	58	1	2	0	0
	White	Black or African American	61	60	0	1	0	0
	White	Hispanic	61	61	0	0	0	0
	White	Asian	61	60	0	1	0	0
	White	Two or More Races	61	61	0	0	0	0
	Non-ML	ML	61	61	0	0	0	0
	Non-SWD	SWD	61	61	0	0	0	0
	Paper-and-pencil	Online	61	60	1	0	0	0
8	Male	Female	61	58	0	3	0	0
	White	Black or African American	61	61	0	0	0	0
	White	Hispanic	61	61	0	0	0	0
	White	Asian	61	58	1	2	0	0
	White	Two or More Races	61	61	0	0	0	0
	Non-ML	ML	61	61	0	0	0	0
	Non-SWD	SWD	61	61	0	0	0	0
	Paper-and-pencil	Online	61	59	1	1	0	0
All Grades	Male	Female	360	347	3	10	0	0
	White	Black or African American	360	358	1	1	0	0
	White	Hispanic	360	359	0	0	0	1
	White	Asian	360	341	3	11	2	3
	White	Two or More Races	360	360	0	0	0	0
	Non-ML	ML	360	358	0	1	0	1
	Non-SWD	SWD	360	359	0	1	0	0

Grade	Reference Group	Focal Group	Total *N* of Items	DIF Classification				
				A	B+	B-	C+	C-
	Paper	Online	360	353	3	4	0	0

Table 4.9. Operational DIF Summary, Mathematics

Grade	Reference Group	Focal Group	Total *N* of Items	DIF Classification				
				A	B+	B-	C+	C-
3	Male	Female	50	48	0	1	0	1
	White	Black or African American	50	49	1	0	0	0
	White	Hispanic	50	50	0	0	0	0
	White	Asian	50	45	3	2	0	0
	White	Two or More Races	50	50	0	0	0	0
	Non-ML	ML	50	50	0	0	0	0
	Non-SWD	SWD	50	50	0	0	0	0
	Paper-and-pencil	Online	50	50	0	0	0	0
4	Male	Female	56	53	1	2	0	0
	White	Black or African American	56	56	0	0	0	0
	White	Hispanic	56	56	0	0	0	0
	White	Asian	56	53	2	1	0	0
	White	Two or More Races	56	56	0	0	0	0
	Non-ML	ML	56	56	0	0	0	0
	Non-SWD	SWD	56	56	0	0	0	0
	Paper-and-pencil	Online	56	56	0	0	0	0
5	Male	Female	56	54	0	2	0	0
	White	Black or African American	56	54	1	1	0	0
	White	Hispanic	56	56	0	0	0	0
	White	Asian	56	54	2	0	0	0
	White	Two or More Races	56	56	0	0	0	0
	Non-ML	ML	56	56	0	0	0	0
	Non-SWD	SWD	56	56	0	0	0	0
	Paper-and-pencil	Online	56	56	0	0	0	0
6	Male	Female	64	62	0	2	0	0
	White	Black or African American	64	63	0	1	0	0
	White	Hispanic	64	64	0	0	0	0
	White	Asian	60	59	0	1	0	0
	White	Two or More Races	64	64	0	0	0	0
	Non-ML	ML	64	64	0	0	0	0
	Non-SWD	SWD	64	64	0	0	0	0
	Paper-and-pencil	Online	60	60	0	0	0	0
7	Male	Female	64	62	0	2	0	0
	White	Black or African American	64	62	1	1	0	0
	White	Hispanic	64	63	0	1	0	0
	White	Asian	60	58	1	1	0	0
	White	Two or More Races	64	64	0	0	0	0
	Non-ML	ML	64	63	0	1	0	0
	Non-SWD	SWD	64	64	0	0	0	0
	Paper-and-pencil	Online	60	58	2	0	0	0
8	Male	Female	67	67	0	0	0	0
	White	Black or African American	67	67	0	0	0	0

Grade	Reference Group	Focal Group	Total *N* of Items	DIF Classification				
				A	B+	B-	C+	C-
	White	Hispanic	67	67	0	0	0	0
	White	Asian	62	61	1	0	0	0
	White	Two or More Races	67	67	0	0	0	0
	Non-ML	ML	67	67	0	0	0	0
	Non-SWD	SWD	67	67	0	0	0	0
	Paper-and-pencil	Online	62	59	3	0	0	0
All Grades	Male	Female	357	346	1	9	0	1
	White	Black or African American	357	351	3	3	0	0
	White	Hispanic	357	356	0	1	0	0
	White	Asian	344	330	9	5	0	0
	White	Two or More Races	357	357	0	0	0	0
	Non-ML	ML	357	356	0	1	0	0
	Non-SWD	SWD	357	357	0	0	0	0
	Paper-and-pencil	Online	344	339	5	0	0	0

4.3.3 Evaluating Bias through Impact Analysis and Effect Size

The impact of achievement testing on minorities can be determined and reported in the form of average scores and in terms of test score reliability. Tables 4.10 and 4.11 present the numbers of students, scale score means, standard deviations (SDs), effect sizes (ESs) (Cohen's *d*), and test form reliability statistics (Coefficient Alpha) can be found in Tables 4.1 through 4.3 (see Section 4.1) for the various subgroups of interest.

One way to evaluate the magnitude of the differences is to calculate the effect size (ES). Cohen's *d* was used to calculate the ES. Cohen's *d* is given by the following formula:

$$d = \frac{\overline{x_a} - \overline{x_b}}{\sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a + n_b) - 2}}}, \qquad (4.16)$$

where $\overline{x_a}$ is the mean score of group A, $\overline{x_b}$ is the mean score of group B, s_a^2 is the variance of group A, s_b^2 is the variance of group B, n_a is the number of students in group A, and n_b is the number of students in group B.

Cohen's *d,* then, expresses the difference in group means in terms of the SD. For example, if *d* = 0.34 for two groups, then it may be interpreted that the mean difference between the two groups is 0.34 of the pooled standard deviation. Cohen (1988) offered guidelines for interpreting the meaning of the *d* statistic: *d* = 0.20 is a small ES, *d* = 0.50 is a medium ES, and *d* = 0.80 is a large ES.

Using Cohen's (1988) guidelines, certain trends become apparent in Tables 4.10 and 4.11. There is a small ES in mean for grades four, six, seven, and eight in ELA for females compared to males. Also, there is a small ES in mean in grades three, four, and five for ELA and grades five,

six, and eight for mathematics for limited English proficient students compared to English proficient students, with English proficient students outperforming limited English proficient students.

There is a medium ES in mean for ELA in all grades, and in mathematics for all grades except grade 8 for Hispanic or Latino students compared to white students, with white students outperforming Hispanic or Latino students. For ELA grades six, seven, and eight, and grade seven in mathematics, there is a medium ES in mean for mathematics for limited English proficient students compared to English proficient students, with English proficient students outperforming limited English proficient students. For grade 3 mathematics, there is a medium ES in mean for students with disabilities compared to students without disabilities, with students without disabilities outperforming students with disabilities.

There is a large ES in mean for all grades in ELA, and all but grade three in mathematics for students with disabilities compared to students without disabilities, with students without disabilities outperforming students with disabilities. Also, for all grades in ELA and mathematics there is a large ES in mean for Black or African American students compared to white students, with white students outperforming Black or African American students.

Table 4.10. Impact Analysis, ELA

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
3	Ethnicity	White	26,782	481.95	122.82	.
		Asian	1,019	518.96	125.61	0.30
		Black or African American	17,226	389.94	110.37	0.78
		Hispanic or Latino	6,632	409.21	122.01	0.59
		Two or More Races	3,317	445.10	123.66	0.30
	Gender	Male	28,337	430.62	126.55	.
		Female	27,310	454.39	125.69	0.19
	ML	English Proficient	53,538	443.95	127.04	.
		Limited English	2,367	398.55	111.30	0.36
	SWD	Without Disabilities	46,789	458.91	121.06	.
		With Disabilities	9,116	355.37	119.68	0.86
4	Ethnicity	White	26,891	553.57	121.58	.
		Asian	1,013	592.87	125.95	0.32
		Black or African American	17,434	455.44	111.57	0.83
		Hispanic or Latino	6,798	482.40	122.92	0.58
		Two or More Races	3,273	514.75	121.09	0.32
	Gender	Male	28,486	498.67	126.30	.
		Female	27,663	524.89	126.90	0.21
	ML	English Proficient	54,058	512.89	127.64	.

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
		Limited English	2,339	475.94	113.85	0.29
	SWD	Without Disabilities	46,940	529.27	122.01	.
		With Disabilities	9,457	422.45	115.19	0.88
5	Ethnicity	White	27,280	586.45	106.91	.
		Asian	956	627.62	107.47	0.39
		Black or African American	17,774	498.76	96.59	0.85
		Hispanic or Latino	6,932	519.77	106.50	0.62
		Two or More Races	3,159	550.27	107.78	0.34
	Gender	Male	29,077	539.14	112.59	.
		Female	27,753	558.35	109.90	0.17
	ML	English Proficient	54,757	550.21	112.07	.
		Limited English	2,308	504.58	92.72	0.41
	SWD	Without Disabilities	47,161	565.18	106.13	.
		With Disabilities	9,904	468.27	102.50	0.92
6	Ethnicity	White	27,101	597.28	112.29	.
		Asian	974	641.17	115.98	0.39
		Black or African American	18,291	504.47	103.19	0.85
		Hispanic or Latino	7,108	532.19	113.35	0.58
		Two or More Races	3,049	558.61	114.67	0.34
	Gender	Male	29,106	541.63	117.71	.
		Female	28,133	572.86	116.34	0.27
	ML	English Proficient	55,214	559.31	118.25	.
		Limited English	2,352	495.74	95.88	0.54
	SWD	Without Disabilities	47,649	575.00	112.34	.
		With Disabilities	9,917	468.84	104.83	0.96
7	Ethnicity	White	28,740	634.25	115.72	.
		Asian	985	688.03	113.88	0.46
		Black or African American	19,118	546.48	99.94	0.80
		Hispanic or Latino	7,054	569.40	108.51	0.57
		Two or More Races	3,100	600.51	114.47	0.29
	Gender	Male	30,266	580.42	117.02	.
		Female	29,528	613.14	115.66	0.28
	ML	English Proficient	58,103	598.88	117.66	.
		Limited English	2,006	521.54	83.74	0.66
	SWD	Without Disabilities	50,122	613.09	113.48	.
		With Disabilities	9,987	512.01	99.84	0.91
8	Ethnicity	White	28,762	657.43	111.45	.

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
		Asian	987	705.70	111.35	0.43
		Black or African American	19,842	577.24	105.24	0.74
		Hispanic or Latino	7,332	594.70	114.69	0.56
		Two or More Races	2,903	627.99	111.71	0.26
	Gender	Male	30,734	602.99	116.96	.
		Female	29,874	641.16	112.75	0.33
	ML	English Proficient	58,554	624.41	116.40	.
		Limited English	2,366	549.30	93.68	0.65
	SWD	Without Disabilities	50,802	638.90	111.21	.
		With Disabilities	10,118	534.08	102.36	0.95

Table 4.11. Impact Analysis, Mathematics

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
3	Ethnicity	White	26,732	493.62	122.22	.
		Asian	1,019	549.77	129.43	0.46
		Black or African American	17,211	393.91	97.68	0.88
		Hispanic or Latino	6,633	431.59	112.23	0.52
		Two or More Races	3,311	451.09	117.07	0.35
	Gender	Male	28,324	456.15	126.66	.
		Female	27,308	449.30	118.25	0.06
	ML	English Proficient	53,540	453.37	123.05	.
		Limited English	2,356	432.97	110.83	0.17
	SWD	Without Disabilities	46,801	465.07	120.28	.
		With Disabilities	9,095	387.86	113.89	0.65
4	Ethnicity	White	26,879	518.84	115.15	.
		Asian	1,013	579.46	125.56	0.52
		Black or African American	17,432	421.13	88.18	0.93
		Hispanic or Latino	6,799	459.29	103.65	0.53
		Two or More Races	3,274	477.84	109.42	0.36
	Gender	Male	28,495	483.87	120.05	.
		Female	27,663	473.42	109.38	0.09
	ML	English Proficient	54,066	479.26	115.52	.
		Limited English	2,338	459.99	101.00	0.17
	SWD	Without Disabilities	46,947	491.53	113.35	.
		With Disabilities	9,457	413.57	100.30	0.70
5	Ethnicity	White	27,268	563.04	117.14	.

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
		Asian	955	637.05	126.03	0.63
		Black or African American	17,773	474.80	88.77	0.83
		Hispanic or Latino	6,930	505.09	103.96	0.51
		Two or More Races	3,163	522.58	112.92	0.35
	Gender	Male	29,071	528.55	120.29	.
		Female	27,760	524.02	109.32	0.04
	ML	English Proficient	54,760	527.36	115.73	.
		Limited English	2,306	497.49	93.51	0.26
	SWD	Without Disabilities	47,160	540.82	112.79	.
		With Disabilities	9,906	456.36	99.09	0.76
6	Ethnicity	White	27,116	558.13	113.24	.
		Asian	974	626.75	130.27	0.60
		Black or African American	18,350	465.04	78.73	0.92
		Hispanic or Latino	7,117	501.12	97.03	0.52
		Two or More Races	3,052	513.30	105.47	0.40
	Gender	Male	29,142	518.31	114.06	.
		Female	28,188	519.49	105.77	0.01
	ML	English Proficient	55,306	520.47	110.79	.
		Limited English	2,356	475.17	77.48	0.41
	SWD	Without Disabilities	47,708	532.25	109.63	.
		With Disabilities	9,954	453.28	85.58	0.75
7	Ethnicity	White	28,743	576.83	108.30	.
		Asian	984	654.50	123.73	0.71
		Black or African American	19,152	492.84	74.52	0.87
		Hispanic or Latino	7,056	523.71	92.08	0.50
		Two or More Races	3,098	539.03	97.95	0.35
	Gender	Male	30,281	541.44	108.36	.
		Female	29,555	542.61	100.03	0.01
	ML	English Proficient	58,148	543.49	104.82	.
		Limited English	2,007	491.13	70.50	0.50
	SWD	Without Disabilities	50,152	554.08	103.87	.
		With Disabilities	10,003	479.89	82.00	0.74
8	Ethnicity	White	28,771	605.40	110.48	.
		Asian	988	683.64	124.82	0.70
		Black or African American	19,895	524.63	82.01	0.81
		Hispanic or Latino	7,341	554.93	96.44	0.47
		Two or More Races	2,896	573.47	104.96	0.29

Grade	Category	Subgroup	N	Mean	Std. Dev.	Effect Size
	Gender	Male	30,762	566.28	111.67	.
		Female	29,909	576.78	102.57	0.10
	ML	English Proficient	58,619	573.03	107.98	.
		Limited English	2,368	524.52	77.93	0.45
	SWD	Without Disabilities	50,849	583.73	106.95	.
		With Disabilities	10,138	508.00	84.91	0.73

4.4 Test Dimensionality

As another measure of the tests' internal structure, DRC examined the unidimensionality of each SC READY ELA and mathematics assessment. One of the underlying assumptions of the IRT models used to scale SC READY is that the tests being calibrated are unidimensional. That is, items composing the SC READY tests in each grade and content area measure a single content area domain. For example, ELA items should measure language ability and not mathematics skills. Standard 1.13 of the AERA, APA, & NCME (2014) *Standards* states the following:

> If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided. (pp. 26–27)

In this section, we examine the internal structure of the tests by evaluating the unidimensionality assumption through Principal Components Analysis (PCA) using WINSTEPS 4.2.0 (Linacre, 2018). This analysis seeks evidence that there exists a single primary factor, the first principal component, which accounts for much of the relationship between items. The presence of a single or dominant factor suggests that a test is sufficiently unidimensional (i.e., measures only one underlying construct).

A PCA was conducted on each test form in each grade and content area. A large first principal component is evident in each analysis. While data are generally considered to be unidimensional if the second eigenvalue is less than or equal to 1.0, it is common to have additional eigenvalues greater than 1.0, which may suggest the presence of other factors. The PCA results are presented in Tables 4.12 and 4.13 for ELA and mathematics, respectively. For ELA, the values of the first eigenvalue ranged from 21.73 to 49.37 and the values of the second eigenvalue ranged from 2.08 to 2.56 across all test forms. For mathematics, the values of the first eigenvalue ranged from 4.41 to 7.03 and the values of the second eigenvalue ranged from 1.70 to 2.46 across all test forms.

For all grades of ELA and mathematics, the ratio of the variance accounted for by the first factor to the second and third is sufficiently large to support the claim that these tests are unidimensional (Cattell, 1952). All tests exhibit first principal components accounting for at least 20.1% of the test variance for ELA and at least 6.1% of the test variance for mathematics. To further investigate the unidimensionality of the SC READY assessments, the ratio of the first

eigenvalue to the second eigenvalue was explored. Previous research shows that the examination of the ratio of the first two (i.e., the two largest) eigenvalues can be useful in determining the existence of dominant factors. Specifically, where large ratios exist between the first and second eigenvalues, a single dominant factor can be said to exist. Although the definition of "large" in the present context is somewhat subjective, the results in Tables 4.12 and 4.13 show that the eigenvalue of the first factor is at least 9.53 times as large as the eigenvalue of the second factor for all grades in ELA and at least 2.48 times as large as the eigenvalue of the second factor for all grades in mathematics. Additionally, for all grades in ELA and mathematics, the percentage of variance explained for the second factor is substantially less than the first factor. This substantial difference, in magnitude in Eigenvalues and the proportion of variance explained, for the two factors indicates that one factor appears to be dominant and that the ELA and mathematics tests are essentially unidimensional.

This evidence supports the claim that there is a dominant dimension underlying the items/tasks in each test and that scores from each test represent performance that is primarily determined by that ability. Construct-irrelevant variance, such as factual knowledge that is irrelevant to doing well in a subject, does not appear to create significant nuisance factors.

Table 4.12: Principal Component Analysis, ELA

Grade	Components/Factors	Eigenvalue	% Variance Explained	Cumulative % Variance Explained
3	First Component	49.37	36.0	36.0
	Second Component	2.17	1.6	37.6
	Ratio (First/Second)	22.78		
4	First Component	44.20	33.6	33.6
	Second Component	2.11	1.6	35.2
	Ratio (First/Second)	20.96		
5	First Component	26.52	24.5	24.5
	Second Component	2.51	2.3	26.8
	Ratio (First/Second)	10.55		
6	First Component	31.17	26.7	26.7
	Second Component	2.21	1.9	28.6
	Ratio (First/Second)	14.09		
7	First Component	21.73	20.1	20.1
	Second Component	2.08	1.9	22.0
	Ratio (First/Second)	10.47		
8	First Component	24.46	21.5	21.5
	Second Component	2.56	2.3	23.8
	Ratio (First/Second)	9.53		

Table 4.13: Principal Component Analysis, Mathematics

Grade	Components	Eigenvalue	% Variance Explained	Cumulative % Variance Explained
3	First Component	4.41	6.4	6.4
	Second Component	1.70	2.5	8.9
	Ratio (First/Second)	2.59		
4	First Component	6.76	8.7	8.7
	Second Component	1.89	2.4	11.1
	Ratio (First/Second)	3.58		
5	First Component	5.34	6.9	6.9
	Second Component	2.15	2.8	9.7
	Ratio (First/Second)	2.48		
6	First Component	5.23	6.1	6.1
	Second Component	1.99	2.3	8.4
	Ratio (First/Second)	2.63		
7	First Component	6.47	7.6	7.6
	Second Component	2.46	2.9	10.5
	Ratio (First/Second)	2.63		
8	First Component	7.03	7.7	7.7
	Second Component	2.31	2.5	10.2
	Ratio (First/Second)	3.04		

4.5 Item Scoring

This section will describe the scoring process used for the SC READY assessments. In particular, this section focuses on the PAS (Performance Assessment Services) process of handscoring ELA TDA items and on autoscoring multiple-choice, multi-select, technology-enhanced, evidence-based selected-response (EBSR), and short-answer items for all content areas. The end of this section describes and reports the results of the inter-rater reliability study conducted on the handscoring of the SC READY TDA items.

This section describes how the SC READY assessments adhere to AERA, APA, & NCME (2014) Standards 4.18, 4.20, 6.8, and 6.9. Each of these standards will be presented in the pertinent sections of this report. Standard 4.18 provides some general guidance for this section:

> Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays. (p. 91)

To preserve the integrity of the items for future use, the scoring criteria used for each item are not presented in this section. Procedures related to recruitment and training of human readers

and monitoring scoring processes contribute to the validity evidence based on response processes.

4.5.1 Handscoring Process

TDA items were scored at a DRC scoring site outside of South Carolina. SCDE personnel remained in contact, as needed, until scoring was complete. DRC staff conducted systematic reviews and analyses of student data on the TDA items to help ensure accurate scoring. Student responses were captured online for all students. Braille, Large Print, and paper-based non-accommodated form student responses were transcribed (entered) into the online system by a test examiner.

4.5.1.1 Selection of Readers

AERA, APA, & NCME (2014) Standard 4.20 specifies the following:

> The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training readers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring. (p. 92)

The following section explains how readers were selected and trained for the SC READY assessment handscoring process. Section 4.5.1.2 describes how the scorers were monitored throughout the SC READY assessment handscoring process.

DRC strived to develop a highly qualified, experienced core of readers so that the integrity of all projects was appropriately maintained. The SC READY 2022 scoring team was staffed with a large number of readers and team leaders who had previous experience with DRC PAS projects. DRC retains a number of raters from year to year. This pool of experienced raters was drawn from to staff the scoring of the 2022 SC READY assessments.

To complete the rater staffing, recruiting events were held and applications for rater positions were screened by DRC's recruiting staff. Candidates were personally interviewed by DRC staff. In addition, each candidate was required to provide an on-demand writing sample and proof of a four-year college degree. In this screening process, preference was given to candidates with previous experience scoring large-scale assessments and degrees emphasizing expertise in ELA. In some locations, staffing partners were used to augment hiring using the same practices as those employed by DRC. The rater pool consisted of educators and other professionals with content-specific backgrounds. These individuals were valued for their content-specific knowledge, but they were required to set aside their own biases about student performance and accept the scoring standards outlined in the SC READY materials.

4.5.1.2 Training Process

AERA, APA, & NCME (2014) Standard 6.9 specifies the following:

> 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. (p. 118)

All materials necessary for scoring were developed by DRC program management and scoring directors. These materials included the scoring guides and training papers used to complete the handscoring of TDA writing prompts.

4.5.1.2.1 Rangefinding Activities

Rangefinding for ELA TDA items took place at DRC's scoring facilities in Minnesota in April of 2022. DRC conducted an internal rangefinding session using scoring directors and project managers from PAS to review all writing prompts to ensure consistency between prompt scoring sessions. Sets of annotated student responses were presented to the committee one prompt at a time. Discussions of student responses were conducted in a manner that emphasized the use of rubric and scoring guideline language. DRC PAS staff recorded the score point decisions made by the rangefinding committee to include the information in final material preparation. The reasoning and scoring philosophies utilized in arriving at the final scores were also noted in order to provide this information during reader training and scoring. After all papers for a prompt were reviewed, the DRC rangefinding committee collaboratively identified responses that would be utilized as anchors during rater training and scoring. Anchor packets for each prompt consisted of fourteen or fifteen papers. All score points and examples of responses within each score point was represented in the anchor papers. The anchor papers were used in the training and qualifying of the readers.

4.5.1.2.2 Qualifying Procedures for Scorers

Each TDA item requires item-specific training materials including a scoring guide composed of a rubric, a passage, and three annotated anchor responses per score point. Following rangefinding, scoring directors composed anchor and training sets of committee-scored responses for each item to be scored. Notes generated during the rangefinding process remained with each response selected, either in the annotation (for anchor examples) or in the scoring director's notes (for training/qualifying set examples).

Anchor responses are selected to illustrate particular scoring concepts. These responses help ensure that scorers are able to make accurate and consistent scoring decisions for the response types they are likely to encounter. For each TDA item, DRC develops two training sets and two qualifying sets of 10 student responses each. The entire group of scorers works on one training set at a time. These responses for training and qualifying further hone each scorer's ability to discern the different score-point levels in an accurate and consistent manner.

Following completion of each training set, the scoring director reviews how each scorer performed and how each response within the set was scored by the group. Next, the scoring director and/or team leaders lead a thorough discussion of each set, answering questions to help ensure that scorers understand the proper way to apply the rubric to each of the training

responses. For operational assessments, after the scoring guide and all training sets have been discussed, scorers must demonstrate their ability to apply the scoring criteria by qualifying (i.e., scoring with an acceptable agreement rate) on at least one of the qualifying sets. For the SC READY exam, scorers are required to achieve 70 percent exact agreement on one of two qualifying sets. Any scorer who does not qualify by the end of the qualifying process will not be allowed to score actual student work.

4.5.1.2.3 Quality Control for Rater Accuracy

AERA, APA, & NCME (2014) Standard 6.8 states the following:

> Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment 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. (p. 118)

This section explains the monitoring procedures that DRC uses to ensure that readers follow established scoring criteria while items are being scored. Detailed scoring rubrics are available for all handscored items, which specify the criteria for scoring those items.

Throughout TDA item scoring, a rater must maintain at least 70 percent exact agreement on validity checks. Any rater who fell below the 70 percent rate could no longer score until they were retrained and re-qualified. All papers scored by that rater since the last acceptable validity check were re-scored.

Throughout handscoring, calculations of inter-rater agreement were provided to the SCDE. The minimum requirement for rater accuracy is an average inter-rater agreement of 70 percent. Overall inter-rater reliability must be maintained at 70 percent exact agreement. Scoring cannot be considered completed if the agreement rate is below this level.

One of DRC's quality control processes is the distribution of validity responses to readers. Validity responses are pre-scored responses that are "seeded" to readers during scoring. Readers cannot tell if a response is a validity response or a live response, making this a powerful measure of quality control. Validity reports compare the true scores of the validity responses to the scores given by each reader in order to monitor for scorer drift. If scoring trends are detected, the reader will be retrained before resuming scoring, and all responses scored by that reader since the last acceptable validity check will be rescored.

Another measure of quality control is inter-rater reliability. Throughout handscoring, daily and cumulative reports detail inter-rater results. DRC understands that inter-rater reliability must be at least 0.70.

DRC's imaging system allows supervisors to spot-check reader performance by reading behind each rater. To this end, the Image Handscoring System randomly selects responses that have been scored by each rater and forwards them to supervisors for review. DRC's imaging system

allows a scoring director to determine read-behind rates (frequency of monitoring) for each individual rater. DRC typically monitors one out of five readings, making adjustments to that ratio as needed. The imaging system randomly selects which images the team leader monitors.

If the supervisor disagrees with the reader's score, the supervisor typically corrects the score. Supervisors can then use the response to retrain the reader by explaining how the response should have been scored and providing a rationale consistent with other, similar training papers. This has proven to be a very effective form of feedback because it is implemented with items live-scored by individual readers.

4.5.1.2.4 Internal Quality Control Reports

A scoring summary report provides daily and cumulative inter-rater reliability results, score point distribution data, and production volumes for each reader and item. Inter-rater reliability monitors how often readers are in exact, adjacent, and nonadjacent agreement with each other, ensuring that an acceptable agreement rate is maintained. The calculations for this report are as follows:

- Percent Exact—total number of responses by scorer where scores are equal divided by the number of responses that were scored twice.
- Percent Adjacent—total number of responses by scorer where scores are one point apart divided by the number of responses that were scored twice.
- Percent Nonadjacent—total number of responses by scorer where scores are more than one score point apart divided by the number of responses that were scored twice.

The ELA TDA item rater results for spring 2022 SC READY are detailed in Table 4.15.

The **Score Point Distribution Report** provides the percentage of responses given each of the score points. For example, for items on a 1–4-point scale, this daily and cumulative report shows how many 1s, 2s, 3s, and 4s a reader has given to all the responses that have been scored at the time the report is produced. These percentages can be compared to room-wide percentages to detect individual scoring issues.

The **Production volumes** indicate the number of responses read by each reader each day so that production rates can be monitored. Additionally, this report includes totals for each item so that progress toward completion can be monitored.

The **Item Status Report** monitors the progress of handscoring. This report tracks each response and indicates the response's status (e.g., "needs a second reading," "complete"). This report ensures that all discrepancies are resolved by the end of the project.

The **Responses Read by Reader Report** identifies all responses scored by an individual reader. This report is useful if any responses need rescoring due to potential reader drift.

The **Read-Behind Log** is a tool used by team leaders/scoring directors to monitor reader reliability. Team Leaders randomly select and read scored responses from each team member daily. If the team leader disagrees with the reader's score, remediation occurs, either with the team leader or with the scoring director. This has proven to be a very effective form of feedback because it is implemented with items live-scored by individual readers.

The **Validity Reports** compare predetermined scores to readers' scores for validity responses. These reports can be run at the individual, team, or room level in order to detect individual, team, or room-wide reader drift.

4.5.1.3 Inter-rater Reliability

All TDA ELA items were scored independently by two readers. The statistics for the inter-rater reliability were calculated for all handscored items at all grades. To determine the reliability of scoring, the percentage of exact agreement and adjacent agreement between the scores from two readers was examined. Non-scorable responses were not included in the inter-rater agreement analysis.

For each item, a quadratic weighted kappa statistic was calculated to reflect the level of improvement beyond the chance level in the consistency of scoring. These quadratic weighted kappa values and the rater agreement statistics are presented in Table 4.15. To aid in the interpretation of the kappa statistic, Table 4.14 provides the suggested cutoffs (Landis & Koch, 1977; Altman, 1991).

Table 4.14: Kappa Statistic Cutoffs

Kappa Value	Strength of Agreement
0	None
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very Good

As shown in Table 4.15, raters demonstrated at least 99.0% exact and adjacent agreement for the writing prompt component scoring in all ELA grades and across all components. The exact agreement ranged from 80.3% to 85.5% for components scored using a 1–4-point rubric. The quadratic weighted kappa values ranged from 0.60 to 0.76 for components scored using a 1–4-point rubric, indicating generally good inter-rater agreement for these components.

Table 4.15. SC Ready TDA Reader Agreement

Grade	Reader Exact %	Reader Adjacent %	Reader Nonadjacent %	Quadratic Weighted Kappa
3	85.3	14.3	0.4	0.60
4	85.5	13.9	0.7	0.68
5	82.0	17.6	0.3	0.73
6	82.3	17.1	0.5	0.73
7	80.3	18.9	0.8	0.73
8	81.5	18.1	0.5	0.76

4.5.2 Technology-Enhanced Item Scoring Process

All technology-enhanced, EBSR, and short-answer items were processed through DRC's autoscoring engine and scored according to the assigned scoring rules. DRC ensured that all rubrics and scoring rules were verified for accuracy before scoring any of these items. DRC established an adjudication process for technology-enhanced, EBSR, and short-answer items to verify that correct answers were identified. DRC's autoscoring quality assurance process included the following:

- A scoring rubric was created for each autoscored item. It was as simple as describing the one and only correct answer for dichotomously scored items (scored as either right or wrong).
- The information from the scoring rubric was entered into the scoring system within the item banking system so that the information resided in one place, along with the item image and other metadata. This scoring information designated specific information that varied by item type. For example, for a drag-and-drop item, the information included which objects are to be placed in which drop region to receive credit.
- The information was then verified by another autoscoring expert.
- After testing started, reports were generated that showed every response, how many students gave that response, and the score the scoring system provided.
- The scoring was then checked against the scoring rubric.
- If any discrepancies were found, the scoring information was modified and verified again. Scoring was then rerun. This checking and modification process continued until no other issues were found.
- As a final check, a final report was run that showed all student responses, along with their frequencies and received scores.

4.5.3 Multiple-Choice and Multi-select Item Scoring Process

Responses to multiple-choice and multi-select items were captured during the online test administration. Responses to multiple-choice and multi-select items were scored using a predefined answer key.

4.5.4 Key Verification

AERA, APA, & NCME (2014) Standard 6.8 states the following:

> Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment 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. (p. 118)

DRC monitors item scoring through item analyses performed using early return data. The purpose of these analyses was to confirm the answer keys by using classical item analysis statistics. Item statistics were flagged using the following statistical criteria.

- *p*-value of keyed response <0.20
- *p*-value of keyed response >0.95
- item-total correlation of keyed response <0.20
- item-total correlation of a distractor >0.05

In addition to the criteria listed, DRC psychometric staff utilized a series of item analyses based on the ability levels of students taking the SC READY assessments to further screen for potential key errors. DRC test development and psychometric staff subsequently reviewed all flagged items.

4.6 Operational Data Analysis

This section of the technical report describes the analyses that occurred on the operational data. These analyses include a classical item analysis and an examination of the raw scores and an item response theory (IRT) analysis involving calibration and scaling.

This section presents the classical item statistics, including aggregate raw score statistics and individual item-level statistics. Next, this section discusses the IRT models used for calibrating the data and addresses the purpose of data calibration and scaling. The calibration samples are presented next, followed by the data calibration results, including the model-data fit for the SC READY ELA and mathematics data. If the IRT models fit the empirical item response distributions of the target population for which generalizations are to be made (i.e., South Carolina students), then the claim is strengthened that the scores are valid indicators of an underlying ability. The lowest obtainable scale score (LOSS) and highest obtainable scale score (HOSS) for the SC READY tests are presented.

This section demonstrates adherence in the SC READY ELA and mathematics tests to *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) Standards 1.8, 4.14, 5.2, 5.13, 5.15, and 7.2. Each standard will be explicated within the appropriate section. Standard 7.2 provides general guidance that is relevant to this section:

> The population for whom a test is intended and specifications for the test should be documented. (p. 126)

Section 4.6.2.1 discusses the calibration sample and compares it to the general population. Section 2 presents the test specifications. Information regarding reported data is discussed in detail in Section 6.3.

4.6.1 Classical Item Analyses

This section presents summary test statistics for each grade or course of SC READY ELA and mathematics. This section also presents item-level statistics for each content area and grade.

4.6.1.2 Test-Level Statistics

Table 4.16 presents the number of items and score points on each test, as well as the means and standard deviations of the raw scores, *p*-values, and item-total correlations (also known as item discrimination values) for each test form at each grade level of the SC READY tests.
The mean *p*-value is the average of all item *p*-values for a specific grade level. The mean item-total correlation (R_{it}) is the average of all item biserial correlations for a specific grade level. The *p*-value and item-total correlation are explained in the next section.

Table 4.16. SC READY Means and Standard Deviations for Raw Scores, p-values, and Item-Total Correlation

Content	Grade	*N* of Items	Mean RS (SD)	Mean *p*-value (SD)	*p*-value Range	R_{it} (SD)
ELA	3	59	39.75 (14.44)	0.65 (0.11)	0.33-0.84	0.47 (0.09)
	4	59	39.45 (14.70)	0.64 (0.11)	0.35-0.83	0.48 (0.09)
	5	59	37.97 (13.97)	0.61 (0.13)	0.34-0.85	0.45 (0.10)
	6	61	40.06 (14.57)	0.62 (0.13)	0.30-0.84	0.46 (0.09)
	7	61	39.21 (14.97)	0.61 (0.11)	0.33-0.78	0.45 (0.07)
	8	61	40.79 (14.95)	0.63 (0.12)	0.37-0.82	0.46 (0.10)
Mathematics	3	50	30.92 (11.71)	0.62 (0.10)	0.42-0.81	0.46 (0.07)
	4	56	32.07 (12.66)	0.57 (0.13)	0.29-0.88	0.44 (0.08)
	5	56	31.90 (12.97)	0.57 (0.11)	0.38-0.83	0.45 (0.08)
	6	60	32.22 (13.62)	0.53 (0.11)	0.34-0.80	0.44 (0.09)
	7	60	31.76 (12.86)	0.53 (0.13)	0.24-0.88	0.41 (0.10)
	8	62	32.24 (13.79)	0.51 (0.12)	0.24-0.80	0.43 (0.08)

4.6.1.3 Item-Level Statistics

Individual operational item level statistics on the test forms are provided in Appendix C.

p-value: The *p*-value is a measure of item difficulty. For a multiple-choice item, the *p*-value is calculated from the number of students who correctly responded to an item divided by the total number of students who attempted the item. The value is reported as a proportion.

As shown in Table 4.16 and Appendix C, the average *p*-values for the ELA assessments ranged from 0.61 to 0.65. The average *p*-values for the mathematics assessments ranged from 0.51 to 0.62. It is important that one examines the range of *p*-values and not just the average *p*-value to determine whether a test measures well. The range of p-values on each form is given in the next section. It is desirable for the test to measure well throughout the range of skills present at

a given grade. These ranges of *p*-values indicate that the items cover the targeted range of difficulty for each grade and content areas, which supports the accuracy of the SC READY ELA and mathematics test scores over the full test scale.

Item-Total Correlations (R_{it}): An item-total correlation is the correlation between an item and the total test score, where the item score is included in the total score. It indicates how well an item differentiates between low- and high-achieving students. In general, items with correlations below 0.20 are said to be poorly discriminating. Nearly all items on the SC READY ELA and mathematics tests had item-total correlations above this threshold. Any item with an item-total correlation below the 0.20 threshold was further analyzed to ensure that the item was correctly keyed.

Omit Rates: The omit rate for each item indicates the percentage of students who did not answer the item. Omit rates can be used to examine possible speededness issues on tests. A test may be speeded if students do not have adequate time to answer all questions on the test. As a rule, an item is said to have a high omit rate if more than 5% of students failed to respond to the item.

This examination of omit rates complies with Standard 4.14 of the AERA, APA, & NCME (2014) *Standards*. This standard is concerned with speededness of a test:

> For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure. (p. 90)

The results in this section will show that, overall, student test scores are not adversely affected by the rate at which students complete the test. In general, students have ample time to complete all sections of the test. This is supported by the omit rates presented in Appendix C, which show that all SC READY items had omit rates well below 5%, suggesting that the majority of students were not rushed.

4.6.2 Item Calibration Using Rasch Measurement Models

Scale scores for the SC READY assessments were developed using the family of Rasch (1960) measurement models for scaling and equating. The advantage of using Rasch models in scaling and equating is that all items measuring performance in a particular content area can be placed on a common difficulty scale, allowing the Rasch difficulty values for the individual items to be used in computing a Rasch logit describing student performance for any raw score point on any form constructed from scaled and equated items.

The Rasch model expresses item difficulty (and student proficiency), rather than percent correct, in units commonly referred to as logits. In the simplest case, a logit is a transformed *p*-value, with the average *p*-value represented by a logit of zero. The logit metric has several mathematical advantages over *p*-values. It is an interval scale, meaning two items with logits of 0 and +1 are the same distance apart as items with logits of +3 and +4. Estimates of item difficulty logits are independent of the ability distribution of the students taking a particular

test. A specific form will have a mean logit of zero, whether the average *p*-value of the test is 0.8 or 0.3. The Rasch model also allows person measures and item measures to be placed on a common metric. This allows the comparison of person proficiency and item difficulty to determine the probability that a person will respond correctly to any given test item. This comparison is not possible in the percentage correct metric used in the true-score model. It is impossible to predict how well a person who answered 80% of the items correctly will perform on an item answered correctly by 80% of the persons.

The standard Rasch calibration procedure sets the mean difficulty of the items on any unanchored calibration at zero. Any item with a *p*-value lower than the mean receives a positive logit, and any item with a *p*-value higher than the mean receives a negative logit. Consequently, the logits for any calibration, whether it is a third-grade reading test or a high school mathematics test, relate to an arbitrary origin defined by the average of item difficulties for that form. The average third-grade reading item will have a logit of zero; the average high school mathematics item will have a logit of zero in unanchored calibrations. This common logit scale describes both item difficulties and student abilities.

Since dichotomous items were part of the SC READY assessments, DRC utilized the Rasch model (Rasch, 1960) for calibrating within grades. The Rasch model predicts the probability of person *n* getting item *i* correct, as follows:

$$P_{ni}(X=1)=\frac{\exp\left(\theta_n - D_{ij}\right)}{1+\exp\left(\theta_n - D_{ij}\right)}. \qquad (4.17)$$

The Rasch model places both student ability and item difficulty (estimated in terms of log-odds or logits) on the same continuum. When the model assumptions are met, the Rasch model provides estimates of a person's ability that are independent of the items employed in the assessment and, conversely, estimates item difficulty independently of the sample of examinees.

Pre-equated item parameters were used in the scoring table generation for SC READY ELA and mathematics. Parameter estimation for items on the SC READY ELA and mathematics assessments using the Rasch model was implemented using WINSTEPS (Linacre, 2018). WINSTEPS uses unconditional joint maximum likelihood estimation as described by Wright and Masters (1982). This calibration software is commercially available and widely used in the testing industry and is considered the industry standard for Rasch calibration.

4.6.2.1 Calibration Sample

This section describes the calibration sample used in the post-equating checks and in the IRT (Rasch) based data analysis in adherence to Standard 1.8 of the AERA, APA, & NCME (2014) *Standards for Educational and Psychological Testing*. Standard 1.8 states the following:

> The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics. (p. 25)

As shown in Tables 4.17 and 4.18, calibration samples were based on census data or very close to census data for the SC READY ELA and mathematics tests. They are thus representative of the South Carolina student population in a given grade level in regard to gender, ethnicity, and accommodation status distribution.

Table 4.17. Summary of Calibration and State Data SC READY ELA

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
3	All Students	55,854		55,905		
	Female	27,225	48.74	27,310	48.85	-0.11
	Male	28,238	50.56	28,337	50.69	-0.13
	Hispanic or Latino	6,617	11.85	6,632	11.86	-0.02
	American Indian or Alaska Native	166	0.30	166	0.30	0.00
	Asian	1,018	1.82	1,019	1.82	0.00
	Black or African American	17,178	30.76	17,226	30.81	-0.06
	Native Hawaiian or Other Pacific Islander	67	0.12	67	0.12	0.00
	White	26,753	47.90	26,782	47.91	-0.01
	Two or More Races	3,310	5.93	3,317	5.93	-0.01
	ML	2,360	4.23	2,367	4.23	-0.01
	Accommodations	4,895	8.76	4,924	8.81	-0.04
4	All Students	56,305		56,397		
	Female	27,559	48.95	27,663	49.05	-0.10
	Male	28,343	50.34	28,486	50.51	-0.17
	Hispanic or Latino	6,769	12.02	6,798	12.05	-0.03
	American Indian or Alaska Native	153	0.27	153	0.27	0.00
	Asian	1,008	1.79	1,013	1.80	-0.01
	Black or African American	17,319	30.76	17,434	30.91	-0.15
	Native Hawaiian or Other Pacific Islander	77	0.14	77	0.14	0.00
	White	26,905	47.78	26,891	47.68	0.10
	Two or More Races	3,267	5.80	3,273	5.80	0.00
	ML	2,335	4.15	2,339	4.15	0.00
	Accommodations	7,032	12.49	7,134	12.65	-0.16
5	All Students	56,968		57,065		
	Female	27,647	48.53	27,753	48.63	-0.10
	Male	28,947	50.81	29,077	50.95	-0.14
	Hispanic or Latino	6,901	12.11	6,932	12.15	-0.03
	American Indian or Alaska Native	137	0.24	137	0.24	0.00
	Asian	954	1.67	956	1.68	0.00
	Black or African American	17,670	31.02	17,774	31.15	-0.13

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
	Native Hawaiian or Other Pacific Islander	72	0.13	72	0.13	0.00
	White	27,274	47.88	27,280	47.81	0.07
	Two or More Races	3,151	5.53	3,159	5.54	0.00
	ML	2,303	4.04	2,308	4.04	0.00
	Accommodations	6,844	12.01	6,934	12.15	-0.14
6	All Students	57,474		57,566		
	Female	28,018	48.75	28,133	48.87	-0.12
	Male	28,988	50.44	29,106	50.56	-0.12
	Hispanic or Latino	7,074	12.31	7,108	12.35	-0.04
	American Indian or Alaska Native	165	0.29	165	0.29	0.00
	Asian	973	1.69	974	1.69	0.00
	Black or African American	18,225	31.71	18,291	31.77	-0.06
	Native Hawaiian or Other Pacific Islander	58	0.10	58	0.10	0.00
	White	27,074	47.11	27,101	47.08	0.03
	Two or More Races	3,046	5.30	3,049	5.30	0.00
	ML	2,334	4.06	2,352	4.09	-0.02
	Accommodations	5,940	10.34	6,038	10.49	-0.15
7	All Students	60,001		60,109		
	Female	29,393	48.99	29,528	49.12	-0.14
	Male	30,121	50.20	30,266	50.35	-0.15
	Hispanic or Latino	7,027	11.71	7,054	11.74	-0.02
	American Indian or Alaska Native	177	0.29	177	0.29	0.00
	Asian	981	1.63	985	1.64	0.00
	Black or African American	19,035	31.72	19,118	31.81	-0.08
	Native Hawaiian or Other Pacific Islander	70	0.12	71	0.12	0.00
	White	28,713	47.85	28,740	47.81	0.04
	Two or More Races	3,099	5.16	3,100	5.16	0.01
	ML	1,987	3.31	2,006	3.34	-0.03
	Accommodations	5,628	9.38	5,721	9.52	-0.14
8	All Students	60,844		60,920		
	Female	29,784	48.95	29,874	49.04	-0.09
	Male	30,593	50.28	30,734	50.45	-0.17
	Hispanic or Latino	7,306	12.01	7,332	12.04	-0.03
	American Indian or Alaska Native	185	0.30	185	0.30	0.00
	Asian	988	1.62	987	1.62	0.00
	Black or African American	19,785	32.52	19,842	32.57	-0.05
	Native Hawaiian or Other Pacific Islander	82	0.13	82	0.13	0.00
	White	28,735	47.23	28,762	47.21	0.01

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
	Two or More Races	2,900	4.77	2,903	4.77	0.00
	ML	2,349	3.86	2,366	3.88	-0.02
	Accommodations	5,467	8.99	5,538	9.09	-0.11

Table 4.18. Summary of Calibration and State Data SC READY Mathematics

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
3	All Students	55,818		55,896		
	Female	27,217	48.76	27,308	48.86	-0.09
	Male	28,210	50.54	28,324	50.67	-0.13
	Hispanic or Latino	6,623	11.87	6,633	11.87	0.00
	American Indian or Alaska Native	165	0.30	165	0.30	0.00
	Asian	1,017	1.82	1,019	1.82	0.00
	Black or African American	17,126	30.68	17,211	30.79	-0.11
	Native Hawaiian or Other Pacific Islander	66	0.12	66	0.12	0.00
	White	26,710	47.85	26,732	47.82	0.03
	Two or More Races	3,308	5.93	3,311	5.92	0.00
	ML	2,349	4.21	2,356	4.21	-0.01
	Accommodations	7,635	13.68	7,723	13.82	-0.14
4	All Students	56,301		56,404		
	Female	27,552	48.94	27,663	49.04	-0.11
	Male	28,339	50.33	28,495	50.52	-0.18
	Hispanic or Latino	6,778	12.04	6,799	12.05	-0.02
	American Indian or Alaska Native	153	0.27	153	0.27	0.00
	Asian	1,007	1.79	1,013	1.80	-0.01
	Black or African American	17,324	30.77	17,432	30.91	-0.14
	Native Hawaiian or Other Pacific Islander	77	0.14	77	0.14	0.00
	White	26,865	47.72	26,879	47.65	0.06
	Two or More Races	3,269	5.81	3,274	5.80	0.00
	ML	2,334	4.15	2,338	4.15	0.00
	Accommodations	7,841	13.93	7,944	14.08	-0.16
5	All Students	56,990		57,066		
	Female	27,665	48.54	27,760	48.65	-0.10
	Male	28,951	50.80	29,071	50.94	-0.14
	Hispanic or Latino	6,914	12.13	6,930	12.14	-0.01
	American Indian or Alaska Native	137	0.24	137	0.24	0.00

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
	Asian	953	1.67	955	1.67	0.00
	Black or African American	17,688	31.04	17,773	31.14	-0.11
	Native Hawaiian or Other Pacific Islander	72	0.13	72	0.13	0.00
	White	27,252	47.82	27,268	47.78	0.04
	Two or More Races	3,156	5.54	3,163	5.54	0.00
	ML	2,303	4.04	2,306	4.04	0.00
	Accommodations	7,726	13.56	7,804	13.68	-0.12
6	All Students	57,421		57,662		
	Female	28,005	48.77	28,188	48.88	-0.11
	Male	28,941	50.40	29,142	50.54	-0.14
	Hispanic or Latino	7,067	12.31	7,117	12.34	-0.04
	American Indian or Alaska Native	165	0.29	165	0.29	0.00
	Asian	973	1.69	974	1.69	0.01
	Black or African American	18,185	31.67	18,350	31.82	-0.15
	Native Hawaiian or Other Pacific Islander	58	0.10	58	0.10	0.00
	White	27,067	47.14	27,116	47.03	0.11
	Two or More Races	3,039	5.29	3,052	5.29	0.00
	ML	2,334	4.06	2,356	4.09	-0.02
	Accommodations	6,599	11.49	6,725	11.66	-0.17
7	All Students	59,973		60,155		
	Female	29,393	49.01	29,555	49.13	-0.12
	Male	30,095	50.18	30,281	50.34	-0.16
	Hispanic or Latino	7,028	11.72	7,056	11.73	-0.01
	American Indian or Alaska Native	178	0.30	179	0.30	0.00
	Asian	982	1.64	984	1.64	0.00
	Black or African American	19,025	31.72	19,152	31.84	-0.12
	Native Hawaiian or Other Pacific Islander	71	0.12	71	0.12	0.00
	White	28,691	47.84	28,743	47.78	0.06
	Two or More Races	3,097	5.16	3,098	5.15	0.01
	ML	1,990	3.32	2,007	3.34	-0.02
	Accommodations	6,121	10.21	6,252	10.39	-0.19
8	All Students	60,823		60,987		
	Female	29,792	48.98	29,909	49.04	-0.06
	Male	30,566	50.25	30,762	50.44	-0.19
	Hispanic or Latino	7,305	12.01	7,341	12.04	-0.03
	American Indian or Alaska Native	186	0.31	186	0.30	0.00
	Asian	987	1.62	988	1.62	0.00
	Black or African American	19,802	32.56	19,895	32.62	-0.06

Grade	Subgroup	Calibration Sample		State Data		Diff
		N	%	N	%	(Calib % - State %)
	Native Hawaiian or Other Pacific Islander	81	0.13	82	0.13	0.00
	White	28,714	47.21	28,771	47.18	0.03
	Two or More Races	2,891	4.75	2,896	4.75	0.00
	ML	2,352	3.87	2,368	3.88	-0.02
	Accommodations	5,941	9.77	6,029	9.89	-0.12

4.6.2.2 Rasch Item Fit

The Rasch fit statistics are used to determine how well items conform to the requirements of the Rasch measurement model. WINSTEPS provides item fit statistics (i.e., infit and outfit) for evaluating the degree to which the Rasch model predicts the observed item responses. Each fit statistic can be expressed as a mean square (MnSq) statistic or on a standardized metric (i.e., Zstd, with mean = 0 and variance = 1). MnSq values are more oriented toward practical significance, while Zstd values are more oriented toward statistical significance. Though both are informative, the Zstd values are very likely too sensitive to the large sample sizes observed on the SC READY ELA and mathematics assessments. In this situation, it is recommended that the Zstd values be ignored if the MnSq values are acceptable (Linacre, 2018).

Both infit and outfit MnSq statistics are the average of standardized residual variance (i.e., the difference between the observed score and the Rasch estimated score divided by the square root of the Rasch model variance). The outfit statistic, however, gives all examinees equal weight in computing the fit and tends to be affected more by unexpected responses far from the person, item, or rating scale category measure. The infit statistic is weighted by the examinee locations relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure. Some consider that extreme infit values are a greater threat to the measurement process than extreme outfit values, since most tests are designed to measure the on-target population rather than extreme outliers.

The expected MnSq value is 1.0 and can range from zero to infinity. Deviation in excess of the expected value can be interpreted as noise or as lack of fit between the items and the model. Values lower than the expected value can be interpreted as item redundancy or overfitting items (i.e., too predictable, too much redundancy), and values greater than the expected value indicate underfitting items (i.e., too unpredictable, too much noise). Rules of thumb regarding “practically significant” MnSq values vary. More conservative practitioners might prefer items with MnSq values that range from 0.8 to 1.2. Others believe reasonable test results can be achieved with values that range from 0.5 to 1.5. In the results below, values outside the range of 0.7 to 1.3 are used to define thresholds for potential significant misfit.

Tables 4.19 and 4.20 present the summary statistics of infit and outfit MnSq statistics for the SC READY ELA and mathematics forms, including the mean, the SD, the minimum values, the

maximum values, and several percentiles (i.e., P10, P25, P50, P75, and P90). As can be seen, the mean values for both fit statistics were close to 1.00 for all content areas and grades. Almost all items had infit values falling in the range of 0.7 to 1.3. Though more outfit values fell outside this range than did infit values, relatively few items fell outside this range. All items flagged for potential misfit were reviewed by DRC psychometric staff. Overall, these results indicate that the Rasch model fits the SC READY ELA and mathematics item data. The model-data fit suggests that the use of the Rasch model provides an appropriate and coherent framework for all scaling and score reporting activities.

Table 4.19. Infit and Outfit Mean Square Statistics: SC READY, ELA

Statistic	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
N	59	59	59	59	59	59	61	61	61	61	61	61
Mean	0.99	0.96	1.01	1.00	0.97	0.95	0.98	0.97	0.99	0.96	0.99	0.98
SD	0.12	0.20	0.14	0.25	0.16	0.23	0.14	0.21	0.11	0.16	0.16	0.24
Minimum	0.72	0.56	0.75	0.62	0.70	0.51	0.66	0.53	0.69	0.59	0.65	0.48
P_{10}	0.83	0.68	0.83	0.73	0.78	0.67	0.79	0.65	0.83	0.78	0.77	0.68
P_{25}	0.90	0.82	0.93	0.82	0.84	0.76	0.94	0.82	0.91	0.84	0.87	0.79
P_{50}	1.00	0.98	0.99	0.97	0.97	0.97	1.00	0.99	0.99	0.94	0.99	0.96
P_{75}	1.08	1.12	1.08	1.14	1.11	1.16	1.07	1.13	1.07	1.09	1.11	1.12
P_{90}	1.13	1.20	1.21	1.36	1.14	1.22	1.15	1.21	1.13	1.17	1.21	1.36
Maximum	1.29	1.41	1.45	1.62	1.33	1.41	1.31	1.37	1.22	1.25	1.45	1.47

Table 4.20. Infit and Outfit Mean Square Statistics: SC READY, Mathematics

Statistic	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
N	50	50	56	56	56	56	64	64	64	64	67	67
Mean	1.02	1.02	1.01	1.03	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.01
SD	0.12	0.20	0.11	0.18	0.11	0.18	0.12	0.17	0.14	0.23	0.11	0.19
Minimum	0.82	0.67	0.81	0.74	0.75	0.63	0.77	0.70	0.77	0.65	0.80	0.73
P_{10}	0.90	0.83	0.85	0.79	0.84	0.76	0.86	0.81	0.86	0.75	0.88	0.80
P_{25}	0.94	0.86	0.94	0.90	0.92	0.89	0.92	0.87	0.91	0.86	0.92	0.87
P_{50}	0.99	0.97	1.03	1.05	0.99	0.96	0.98	0.97	0.99	0.98	0.98	0.99
P_{75}	1.08	1.11	1.07	1.14	1.07	1.10	1.08	1.12	1.08	1.11	1.09	1.14
P_{90}	1.21	1.24	1.14	1.19	1.16	1.25	1.18	1.25	1.19	1.25	1.16	1.25
Maximum	1.32	1.82	1.35	1.69	1.23	1.41	1.24	1.40	1.57	2.12	1.38	1.63

4.6.2.3 Local Independence

Local independence (LI) is a fundamental assumption of IRT. No relationship should exist between examinees' responses to different items after accounting for the abilities measured by a test. In formal statistical terms, a test X that comprises items X_1, X_2, … X_n has LI with respect to the latent variable ϑ if, for all $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ and ϑ,

$$P(\mathbf{X} = \mathbf{x} \mid \theta) = \prod_{i=1}^{I} P(X_i = x_i \mid \theta). \tag{4.18}$$

This formula essentially states that the probability of any pattern of responses across all items (**x**), after conditioning on the abilities (θ) measured by the test, should be equal to the product of the conditional probabilities across each item (cf. the multiplication rule for independent events where the joint probabilities are equal to the product of the associated marginal probabilities).

The equation above shows the condition after satisfying the strong form of LI. A weak form of local independence (WLI) is proposed by McDonald (1979). The distinction is important because many indicators of local dependency are actually framed by WLI. The requirement would be for the conditional covariances of all pairs of item responses, conditioned on the abilities, to be equal to zero. When this assumption is met, the joint probability of responses to an item pair, conditioned on abilities, is the product of the probabilities of responses to these two items, as shown below. (This is a weaker form because higher-order dependencies among items are allowed.) Based on the WLI, the following equation can be derived:

$$P(X_i = x_i, X_j = x_j \mid \theta) = P(X_i = x_i \mid \theta) P(X_j = x_j \mid \theta). \tag{4.19}$$

Marais and Andrich (2008) point out that local item dependence in the Rasch model can occur in two ways that some may not distinguish. The first way occurs when the assumption of unidimensionality is violated. Here, other nuisance dimensions besides a dominant dimension determine student performance. This can be called *trait dependence*. The second violation occurs when responses to one item depend on responses to another item. This is a violation of statistical independence and can be called *response dependence*. Many people treat the assumptions of unidimensionality and LI as one phenomenon and believe that once unidimensionality holds, LI also holds. By distinguishing the two sources of local dependence, one can see that while LI can be related to unidimensionality, the two are different assumptions and, therefore, require different tests.

Residual item correlations provided in WINSTEPS for each item pair were used to assess the local dependence among SC READY ELA and mathematics items. In general, these residuals are computed as follows. First, expected item performance based on the Rasch model is determined using ability and item parameter estimates. Next, the deviations (residuals) between the examinees' expected and observed performances are determined for each item. Finally, for each item pair, a correlation between the respective deviations is computed.

Three types of residual correlations are available in WINSTEPS: raw, standardized, and logit. It should be noted that the raw score residual correlation essentially corresponds to Yen's *Q3* index, a popular LI statistic. The expected value for the *Q3* statistic is approximately $-1/(k - 1)$ when no local dependence exists, where k is test length (Yen, 1993). Thus, the expected *Q3* values should be approximately -0.04 for the SC READY ELA and mathematics assessments.

Absolute index values that are greater than 0.20 indicate a degree of local dependence that probably should be examined by test developers (Chen & Thissen, 1997).

Since the three residual correlations are very similar, the default "standardized residual correlation" in WINSTEPS is used for these analyses. Tables 4.21 and 4.22 show the summary statistics—mean, SD, minimum, maximum, and several percentiles (P_{10}, P_{25}, P_{50}, P_{75}, and P_{90})—for all the residual correlations for each ELA and mathematics test. The total number of item pairs and the number of pairs with residual correlations greater than 0.20 are also reported in these tables. There were no item pairs in ELA found to have a residual correlation greater than 0.20. For mathematics, there was one item pair with residual correlations greater than 0.20 in grade three, six in grade five, and three in grade eight. The mean residual correlations are slightly negative, and the values range from -0.02 to -0.01 for ELA and have a consistent mean of -0.02 for mathematics. Most of the correlations are very small (<0.01), suggesting local item independence generally holds for the SC READY ELA and mathematics tests.

Table 4.21. Summary of Residual Correlations for SC READY ELA

Statistic	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
N pairs of items	1711	1711	1711	1830	1830	1830
Mean	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01
SD	0.03	0.03	0.03	0.03	0.03	0.03
Minimum	-0.11	-0.11	-0.10	-0.12	-0.14	-0.12
P_{10}	-0.04	-0.04	-0.05	-0.04	-0.04	-0.05
P_{25}	-0.03	-0.03	-0.04	-0.03	-0.03	-0.03
P_{50}	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
P_{75}	0.00	0.00	0.00	0.00	0.00	0.00
P_{90}	0.01	0.02	0.02	0.02	0.01	0.02
Maximum	0.11	0.17	0.13	0.13	0.19	0.18
>|0.20|	0	0	0	0	0	0

Table 4.22. Summary of Residual Correlation for SC READY Mathematics

Statistic	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
N pairs of items	1225	1540	1540	2016	2016	2211
Mean	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01
SD	0.03	0.02	0.04	0.02	0.03	0.03
Minimum	-0.10	-0.10	-0.10	-0.09	-0.12	-0.10
P_{10}	-0.04	-0.04	-0.05	-0.04	-0.05	-0.04
P_{25}	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
P_{50}	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
P_{75}	-0.01	-0.01	-0.01	0.00	0.00	0.00
P_{90}	0.01	0.01	0.01	0.01	0.02	0.01
Maximum	0.26	0.12	0.59	0.15	0.20	0.52
>|0.20|	1	0	6	0	0	3

Table 4.23 lists all item pairs with residual correlations greater than 0.20. The item sequence shown in the table is the same item sequence as in the administration. In the ELA assessment, items with large residual correlations typically refer to the same passage or prompt. In the

mathematics assessment, large residual correlations tend to be between items that reflect a very specific subject matter within a given domain. Future form construction activities will look to further reduce the residual correlations where possible.

Table 4.23. SC READY Item Pairs with Large Residuals

Content Area	Grade	Item 1	Item 2	Correlation
Mathematics	3	17	46	0.26
Mathematics	5	7	22	0.59
Mathematics	5	6	13	0.42
Mathematics	5	22	34	0.30
Mathematics	5	20	33	0.27
Mathematics	5	7	34	0.26
Mathematics	5	15	26	0.21
Mathematics	8	58	67	0.52
Mathematics	8	47	49	0.36
Mathematics	8	11	36	0.20

4.6.2.4 Linking and Equating

Test equating is a statistical process of placing scores from two or more parallel assessments onto a common scale, resulting in direct comparability of scores from two different test forms. A common-item design was used to link the SC READY assessments and any field test items to create SC READY item banks to establish the SC READY scale. When new items are field tested, DRC used Rasch equating to link each new SC READY test items to the existing scoring scales, which were established following the 2016 SC READY administration.

Standard 5.13 of the AERA, APA, & NCME (2014) *Standards* states the following:

> 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. (p. 105)

Standard 5.15 of the AERA, APA, & NCME (2014) *Standards* states the following:

> 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 area specifications 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. (p. 106)

During each administration, DRC conducts post-equating checks to ensure that the banked Rasch difficulty parameters are still appropriate for the given administration. After compiling a sufficient number of student responses, DRC conducted an unanchored (free) calibration of the items on the current form. Using the existing and current item difficulties, DRC equated the current form to the existing SC READY scale, based on the following guidelines using the Robust *z* method (Huynh & Meyer, 2010).

Guidelines for a successful equating:

- The correlation of existing and current Rasch difficulties should be equal to or greater than 0.95.
- The ratio of the standard deviations of existing and current Rasch difficulties should be within the range of 0.90 to 1.10.
- The distribution of students scoring in each achievement level should not vary unusually from year to year.
- The mean SC READY scale score should not vary unusually from year to year.
- If more than one potential linking item is deleted in Step 9 below, the items should not come from a single content area standard and should vary in difficulty.

Steps in equating checks:
The following steps were used to perform the Rasch equating**:**

1. Calculate the mean and standard deviation of the linking pool's existing item difficulties.
2. Calculate the mean and standard deviation of the linking pool's current (unanchored) item difficulties.
3. Calculate the ratio of the two standard deviations (from Steps 1 and 2).
4. Calculate the correlation between the existing and current item difficulties for the items in the linking pool.
5. Calculate the difference between the existing and current item difficulties for each item in the linking pool.
6. Calculate the mean of the differences determined above.

If the set of linking items meets the above guidelines, go to Step 10. Otherwise, determine robust *Z* statistics as follows:

7. Calculate the median of the differences ($m_{\mathbf{diff}}$).
8. Calculate the interquartile range of the differences ($r_{\mathbf{iq}}$).
9. Calculate the robust *Z* statistic for each item in the linking pool, where the robust *Z* is defined as the difference between the item's existing ($b_{\mathbf{e}}$) and current ($b_{\mathbf{c}}$) item difficulties minus the median of the differences, that quantity divided by the quantity the interquartile range multiplied by 0.74:

$$Z = [(b_{\mathbf{e}} - b_{\mathbf{c}}) - m_{\mathbf{diff}}] / (r_{\mathbf{iq}} * 0.74). \quad (4.20)$$

Once the above calculations have been made, the following procedure will be used in determining the set of linking items to be used for the Rasch equating:

10. Remove any items with absolute values of the robust *Z* statistic greater than 1.645 from the pool of potential linkers, unless this would result in more than 20% of potential linking items being deleted. In that case, remove the items with the largest absolute values of *Z* up to 20% of the items.
11. Repeat Steps 1 through 6.

12. The mean difference of the difficulties of the items currently in the linking pool (from Step 6, above) is the additive constant used for equating the current scale to the existing scale.

DRC provided SCDE with documentation of the above process and its results. Note that SCDE may choose to accept an equating which fails to meet one or more of the above guidelines. DRC and SCDE will keep track of deleted potential linking items across administrations to ensure that deleted items are not selected from one or two specific strands or narrow ranges of difficulty.

Table 4.24 provides equating results for the robust *Z* method. This table summarizes the following information for each grade level and content area: grade level, number of anchors, number of iterations, correlation and ratio of the standard deviations between the anchored and unanchored difficulty parameters, and the unweighted link constant. It is not unusual for the Robust Z procedure to flag items to be freely calibrated. For both ELA grades three and five and all grades of mathematics, all anchor sets remained intact, because they met the criteria established for the equating guidelines: the anchored and unanchored difficulties had high correlation (>0.95) and ratio of the standard deviations between the anchored and unanchored, the standard deviation ratios were within the 0.90 to 1.10 range , except grade 4 ELA at 0.80. For ELA grades four, six, seven, and eight, where items were flagged for removal from the anchor set, the scoring tables for the pre-and post-equated results were examined closely by SCDE and DRC staff. Student level results were examined by SCDE and DRC and the pre-equated and post-equated solutions were similar, indicating that no scoring table adjustments were needed.

Table 4.24. Equating Summary for SC READY

Content Area	Grade	Number of Anchors	Number of Iterations	Correlation	Standard Deviation Ratio	Unweighted Link Constant
ELA	3	58	2	0.89	1.1	-1.49
	4	58	4	0.89	0.8	-0.72
	5	58	4	0.92	1.0	-0.11
	6	60	4	0.90	0.9	-0.14
	7	60	7	0.90	0.9	0.34
	8	60	6	0.96	1.1	0.57
Mathematics	3	50	0	0.98	1.0	-1.28
	4	56	0	0.97	1.0	-0.72
	5	56	0	0.95	1.0	-0.15
	6	64	0	0.96	1.0	-0.03
	7	64	0	0.97	1.0	0.27
	8	67	0	0.89	1.1	0.67

4.6.3 Analyses by Reporting Categories

Three sets of analyses were conducted at the reporting category level for ELA and mathematics in another attempt to assess the internal structure of the SC READY assessments. The reporting categories are content area categories and consist of items measuring similar sets of skills or knowledge. Each category was measured by at least eight items and was worth at least eight raw score points.

In order to assess the internal structure of the SC READY assessments, correlation coefficients that measure the relationship between the reporting category scores within a grade and content area were first computed. Second, the reliability of each category was computed. Finally, the SEM was computed for each reporting category.

4.6.3.1 Correlations among Reporting Category Scores

In this section, we report the strength of the interrelationships among the reporting categories by computing correlations between them. Tables 4.25 through 4.27 report the uncorrected Pearson product-moment (PPM) correlation coefficients and the PPM corrected for attenuation (CAPPM) for the two content area tests. The PPM among the reporting category subscores is presented below the diagonal portion of the matrix, and the CAPPM is presented above the diagonal portion of the matrix. It should be noted that Table 4.25 shows ELA at the reporting category level and Table 4.27 provides results for ELA at the standard level.

The uncorrected PPM in Tables 4.25 through 4.27 should be interpreted in the context of the reliability coefficient. In general, we expect to see lower PPM coefficients between variables that are less reliable. In most cases, the PPM coefficients show that performance on one reporting category is moderately related to performance on another reporting category within the same grade and content area. For ELA at the reporting category level, the correlations ranged from 0.65 to 0.84, for ELA at the standard level, the correlations ranged between 0.50 and 0.82, and for mathematics, the correlations ranged from 0.62 to 0.78 for all pairs of reporting categories across all grades. It should be noted that, in general, the value of the correlation coefficients was affected by the number of items measuring each reporting category in all content areas. We expect to see a more modest relationship reported between the reporting categories because of the lower number of items measuring each of the reporting categories. The PPM between two reporting category subscores may be artificially low because of measurement error.

AERA, APA, & NCME (2014) Standard 1.21 states the following:

> When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates. (p. 29)

The attenuation of the PPM can be corrected for statistically using Spearman's formula:

$$CAPPM = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}, \quad (4.21)$$

where r_{xy} is the PPM between two content area strands, r_{xx} is the reliability of one of those content area strands, and r_{yy} is the reliability of the other content area strand.

In Tables 4.25 through 4.27, the CAPPMs indicate strong relationships between the content area strands. In some cases, the CAPPM is 1.00. "Disattenuated values of or greater than 1.00 indicate that measurement error is not randomly distributed" (Schumacker & Muchinsky, 1996). The strong relationships suggested by the CAPPM in Tables 4.25 through 4.27 are further evidence of the validity based on the tests' internal structure. Since the overall content area comprises the content area strand subscores and the content area is expected to measure a single dimension, we would expect that these subscores are also highly related.

Table 4.25. Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Categories, ELA Reporting Category Level

Grade	No.	Reporting Category	*N* Items	1	2	3
3	1	Reading	38		0.97	0.93
	2	Writing	15	0.84		0.96
	3	Inquiry	6	0.69	0.66	
4	1	Reading	38		0.96	0.92
	2	Writing	15	0.81		1.00
	3	Inquiry	6	0.70	0.69	
5	1	Reading	37		0.96	0.95
	2	Writing	16	0.80		0.99
	3	Inquiry	6	0.69	0.65	
6	1	Reading	38		0.96	0.93
	2	Writing	15	0.81		1.00
	3	Inquiry	8	0.75	0.74	
7	1	Reading	38		0.96	0.95
	2	Writing	16	0.82		0.99
	3	Inquiry	7	0.74	0.72	
8	1	Reading	38		0.98	0.91
	2	Writing	15	0.80		1.00
	3	Inquiry	8	0.71	0.70	

Table 4.26. Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Categories, ELA Standard Level

Grade	No.	Standard	*N* Items	1	2	3	4	5	6	7
3	1	Reading Literary—Meaning & Context	10		0.95	0.98	0.94	0.91	0.88	0.91

Grade	No.	Standard	*N* Items	1	2	3	4	5	6	7
	2	Reading Literary—Language, Craft, & Structure	9	0.74		0.97	0.99	0.95	0.92	0.90
	3	Reading Informational—Meaning & Context	9	0.74	0.75		1.00	0.96	0.91	0.94
	4	Reading Informational—Language, Craft, & Structure	10	0.73	0.78	0.77		0.95	0.93	0.93
	5	Writing—Meaning, Context, & Craft	7	0.65	0.69	0.68	0.70		0.98	0.96
	6	Writing—Language	7	0.62	0.66	0.64	0.66	0.65		0.90
	7	Inquiry	6	0.62	0.62	0.63	0.64	0.61	0.56	
4	1	Reading Literary—Meaning & Context	11		0.99	0.97	0.97	0.89	0.84	0.89
	2	Reading Literary—Language, Craft, & Structure	9	0.82		0.99	1.00	0.93	0.88	0.91
	3	Reading Informational—Meaning & Context	9	0.76	0.75		1.00	0.93	0.88	0.93
	4	Reading Informational—Language, Craft, & Structure	9	0.77	0.78	0.74		0.94	0.91	0.93
	5	Writing—Meaning, Context, & Craft	8	0.67	0.68	0.64	0.66		0.94	0.98
	6	Writing—Language	6	0.62	0.63	0.59	0.63	0.61		0.91
	7	Inquiry	6	0.65	0.65	0.62	0.64	0.63	0.58	
5	1	Reading Literary—Meaning & Context	9		1.00	0.94	0.99	0.95	0.86	0.93
	2	Reading Literary—Language, Craft, & Structure	9	0.76		0.93	0.98	0.96	0.86	0.94
	3	Reading Informational—Meaning & Context	9	0.72	0.70		0.95	0.95	0.89	0.93
	4	Reading Informational—Language, Craft, & Structure	10	0.76	0.75	0.73		0.95	0.86	0.93
	5	Writing—Meaning, Context, & Craft	8	0.67	0.67	0.67	0.68		0.95	1.00
	6	Writing—Language	7	0.55	0.54	0.56	0.56	0.56		0.89
	7	Inquiry	6	0.62	0.62	0.62	0.62	0.62	0.50	
6	1	Reading Literary—Meaning & Context	10		0.99	0.97	0.98	0.91	0.91	0.91
	2	Reading Literary—Language, Craft, & Structure	8	0.73		0.99	0.99	0.90	0.91	0.93
	3	Reading Informational—Meaning & Context	10	0.74	0.72		1.00	0.88	0.91	0.91
	4	Reading Informational—Language, Craft, & Structure	10	0.75	0.72	0.75		0.92	0.92	0.94
	5	Writing—Meaning, Context, & Craft	8	0.67	0.63	0.64	0.67		0.94	0.99
	6	Writing—Language	6	0.64	0.61	0.62	0.64	0.62		0.95
	7	Inquiry	8	0.68	0.65	0.66	0.69	0.69	0.63	
7	1	Reading Literary—Meaning & Context	10		1.00	1.00	1.00	0.94	0.86	0.94
	2	Reading Literary—Language, Craft, & Structure	8	0.75		1.00	1.00	0.94	0.87	0.93
	3	Reading Informational—Meaning & Context	9	0.73	0.71		1.00	0.95	0.87	0.95
	4	Reading Informational—Language, Craft, & Structure	11	0.76	0.73	0.74		0.96	0.87	0.96
	5	Writing—Meaning, Context, & Craft	8	0.68	0.66	0.66	0.69		0.93	0.99

Grade	No.	Standard	*N* Items	1	2	3	4	5	6	7
	6	Writing—Language	7	0.63	0.61	0.61	0.63	0.64		0.89
	7	Inquiry	7	0.68	0.64	0.65	0.68	0.67	0.61	
8	1	Reading Literary—Meaning & Context	10		1.00	0.96	0.98	0.90	0.89	0.89
	2	Reading Literary—Language, Craft, & Structure	9	0.81		0.97	0.99	0.92	0.91	0.90
	3	Reading Informational—Meaning & Context	11	0.77	0.75		1.00	0.93	0.90	0.91
	4	Reading Informational—Language, Craft, & Structure	8	0.76	0.73	0.75		0.94	0.92	0.94
	5	Writing—Meaning, Context, & Craft	8	0.64	0.63	0.65	0.63		0.97	1.00
	6	Writing—Language	6	0.60	0.58	0.59	0.58	0.57		0.95
	7	Inquiry	8	0.65	0.63	0.65	0.64	0.63	0.56	

Table 4.27. Uncorrected Correlation Coefficient (below Diagonal) and Corrected Correlation Coefficient (above Diagonal) among Reporting Categories, Mathematics

Grade	No.	Reporting Category	*N* Items	1	2	3	4	5
3	1	Number Sense and Base Ten	9		0.93	0.97	0.87	0.97
	2	Number Sense—Fractions	8	0.69		0.94	0.92	0.96
	3	Algebraic Thinking & Operations	12	0.74	0.72		0.90	0.96
	4	Geometry	8	0.62	0.66	0.66		0.91
	5	Measurement & Data Analysis	13	0.74	0.74	0.77	0.67	
4	1	Number Sense and Base Ten	12		0.97	0.98	0.89	0.95
	2	Number Sense—Operations & Fractions	12	0.75		0.97	0.90	0.96
	3	Algebraic Thinking & Operations	12	0.77	0.74		0.87	0.98
	4	Geometry	8	0.65	0.65	0.63		0.91
	5	Measurement & Data Analysis	12	0.74	0.73	0.75	0.65	
5	1	Number Sense and Base Ten	12		0.96	0.96	0.89	0.94
	2	Number Sense—Operations & Fractions	11	0.74		0.95	0.87	0.95
	3	Algebraic Thinking & Operations	12	0.75	0.74		0.94	0.95
	4	Geometry	10	0.66	0.64	0.71		0.90
	5	Measurement & Data Analysis	11	0.73	0.73	0.75	0.67	
6	1	The Number System	15		0.99	0.99	1.00	0.95
	2	Ratios & Proportional Relationships	10	0.75		0.97	0.96	0.92
	3	Expressions, Equations, & Inequalities	15	0.78	0.73		0.98	0.97
	4	Geometry & Measurement	9	0.76	0.70	0.74		0.96
	5	Data Analysis & Statistics	11	0.71	0.66	0.72	0.69	
7	1	The Number System	14		0.98	0.98	0.99	0.96
	2	Ratios & Proportional Relationships	9	0.73		0.94	0.97	0.97
	3	Expressions, Equations, & Inequalities	12	0.73	0.70		0.96	0.93
	4	Geometry & Measurement	13	0.71	0.69	0.68		0.97
	5	Data Analysis, Statistics, & Probability	12	0.72	0.72	0.69	0.69	
8	1	The Number System	8		0.88	0.88	0.88	0.84
	2	Functions	14	0.66		0.99	1.00	1.00

Grade	No.	Reporting Category	*N* Items	1	2	3	4	5
	3	Expressions, Equations, & Inequalities	14	0.66	0.74		0.99	0.91
	4	Geometry & Measurement	14	0.63	0.75	0.71		1.00
	5	Data Analysis, Statistics, & Probability	12	0.63	0.75	0.68	0.76	

4.6.3.2 Reliability and Standard Error of Measurement of Reporting Categories

Raw score summary statistics (mean and standard deviation), Cronbach's (1951) coefficient alpha, and SEM were computed for each of the content area strands by grade and content area using the calibration sample. These statistics are presented in Tables 4.28 through 4.30 for the SC READY assessments. Reliability indices, such as Cronbach's coefficient alpha (and resulting SEM), are a function of the number of test items. It is expected that coefficient alpha would be lower for a content area strand assessed by a small number of items than for a content area strand assessed by a larger number of items.

Table 4.28. Reliability and Standard Error of Measurement of Reporting Categories
SC READY, ELA Reporting Category Level

Grade	No	Reporting Category	N Items & Points	N Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	SEM
3	1	Reading	38	55,905	24.53	9.41	0.93	2.49
	2	Writing	15 (22 points)	55,905	11.62	4.32	0.79	1.96
	3	Inquiry	6	55,905	3.60	1.65	0.59	1.05
4	1	Reading	38	56,397	23.58	9.79	0.93	2.50
	2	Writing	15 (22 points)	56,397	11.92	4.27	0.76	2.09
	3	Inquiry	6	56,397	3.96	1.62	0.63	0.99
5	1	Reading	37	57,065	23.06	9.03	0.93	2.46
	2	Writing	16 (23 points)	57,065	11.16	4.37	0.75	2.20
	3	Inquiry	6	57,065	3.75	1.59	0.57	1.04
6	1	Reading	38	57,566	22.35	9.18	0.92	2.56
	2	Writing	15 (22 points)	57,566	12.00	4.32	0.77	2.06
	3	Inquiry	8	57,566	5.71	2.05	0.71	1.11
7	1	Reading	38	60,109	23.16	9.04	0.92	2.57
	2	Writing	16 (23 points)	60,109	11.85	4.95	0.79	2.26
	3	Inquiry	7	60,109	4.20	1.96	0.67	1.13
8	1	Reading	38	60,920	23.88	9.49	0.93	2.47
	2	Writing	15 (22 points)	60,920	11.52	4.57	0.73	2.38
	3	Inquiry	8	60,920	5.39	1.97	0.64	1.18

Note: The number of items and points differ for writing because the four-point TDA item is included.

Table 4.29. Reliability and Standard Error of Measurement of Reporting Categories
SC READY, ELA Standard Level

Grade	No	Standard	N Items & Points	N Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	SEM
3	1	Reading Literary—Meaning & Context	10	55,905	5.99	2.73	0.77	1.33
	2	Reading Literary—Language, Craft, & Structure	9	55,905	6.09	2.57	0.79	1.17
	3	Reading Informational—Meaning & Context	9	55,905	5.84	2.42	0.75	1.20
	4	Reading Informational—Language, Craft, & Structure	10	55,905	6.62	2.71	0.79	1.25
	5	Writing—Meaning, Context, & Craft	7	55,905	4.88	1.84	0.68	1.04
	6	Writing—Language	7	55,905	4.85	1.78	0.65	1.04
	7	Inquiry	6	55,905	3.60	1.65	0.59	1.05
4	1	Reading Literary—Meaning & Context	11	56,397	6.60	3.32	0.85	1.31
	2	Reading Literary—Language, Craft, & Structure	9	56,397	5.90	2.61	0.80	1.16
	3	Reading Informational—Meaning & Context	9	56,361	5.13	2.41	0.71	1.30
	4	Reading Informational—Language, Craft, & Structure	9	56,397	5.94	2.40	0.75	1.21
	5	Writing—Meaning, Context, & Craft	8	56,397	5.60	1.99	0.66	1.15
	6	Writing—Language	6	56,397	4.32	1.58	0.64	0.95
	7	Inquiry	6	56,397	3.96	1.62	0.63	0.99
5	1	Reading Literary—Meaning & Context	9	57,065	5.68	2.50	0.76	1.22
	2	Reading Literary—Language, Craft, & Structure	9	57,065	6.27	2.29	0.75	1.14
	3	Reading Informational—Meaning & Context	9	57,065	4.89	2.52	0.76	1.23
	4	Reading Informational—Language, Craft, & Structure	10	56,997	6.22	2.76	0.78	1.30
	5	Writing—Meaning, Context, & Craft	8	57,065	4.97	2.03	0.65	1.20
	6	Writing—Language	7	57,065	3.62	1.77	0.53	1.21

Grade	No	Standard	N Items & Points	N Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	SEM
	7	Inquiry	6	57,065	3.75	1.59	0.57	1.04
6	1	Reading Literary—Meaning & Context	10	57,566	5.93	2.75	0.78	1.30
	2	Reading Literary—Language, Craft, & Structure	8	57,566	4.86	2.20	0.70	1.20
	3	Reading Informational—Meaning & Context	10	57,566	5.36	2.64	0.75	1.32
	4	Reading Informational—Language, Craft, & Structure	10	57,566	6.21	2.65	0.76	1.29
	5	Writing—Meaning, Context, & Craft	8	57,566	5.67	2.03	0.70	1.11
	6	Writing—Language	6	57,566	3.80	1.60	0.63	0.97
	7	Inquiry	8	57,566	5.71	2.05	0.71	1.11
7	1	Reading Literary—Meaning & Context	10	60,109	5.89	2.77	0.78	1.31
	2	Reading Literary—Language, Craft, & Structure	8	60,109	4.92	2.19	0.72	1.17
	3	Reading Informational—Meaning & Context	9	60,109	5.61	2.32	0.70	1.27
	4	Reading Informational—Language, Craft, & Structure	11	60,109	6.74	2.80	0.76	1.38
	5	Writing—Meaning, Context, & Craft	8	60,109	4.56	2.13	0.68	1.20
	6	Writing—Language	7	60,109	4.56	1.97	0.69	1.09
	7	Inquiry	7	60,109	4.20	1.96	0.67	1.13
8	1	Reading Literary—Meaning & Context	10	60,920	6.52	2.84	0.83	1.19
	2	Reading Literary—Language, Craft, & Structure	9	60,920	5.93	2.44	0.76	1.20
	3	Reading Informational—Meaning & Context	11	60,920	6.42	2.98	0.79	1.38
	4	Reading Informational—Language, Craft, & Structure	8	60,920	5.00	2.19	0.72	1.16
	5	Writing—Meaning, Context, & Craft	8	60,920	4.96	1.96	0.62	1.20
	6	Writing—Language	6	60,920	3.75	1.59	0.55	1.07
	7	Inquiry	8	60,920	5.39	1.97	0.64	1.18

Table 4.30. Reliability and Standard Error of Measurement of Reporting Categories, Mathematics

Grade	No.	Reporting Category	N Items & Points	N Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	SEM
3	1	Number Sense and Base Ten	9	55,896	5.27	2.46	0.73	1.28
	2	Number Sense—Fractions	8	55,896	4.92	2.27	0.74	1.16
	3	Algebraic Thinking & Operations	12	55,896	7.50	3.14	0.80	1.42
	4	Geometry	8	55,896	4.89	2.14	0.69	1.19
	5	Measurement & Data Analysis	13	55,896	8.33	3.33	0.80	1.49
4	1	Number Sense and Base Ten	12	56,404	7.36	3.15	0.79	1.44
	2	Number Sense—Operations & Fractions	12	56,404	6.81	3.01	0.76	1.47
	3	Algebraic Thinking & Operations	12	56,404	6.51	3.12	0.78	1.47
	4	Geometry	8	56,404	5.33	2.06	0.68	1.16
	5	Measurement & Data Analysis	12	56,404	6.05	3.05	0.76	1.50
5	1	Number Sense and Base Ten	12	57,066	6.82	3.13	0.77	1.49
	2	Number Sense—Operations & Fractions	11	57,066	5.76	2.95	0.76	1.43
	3	Algebraic Thinking & Operations	12	57,066	7.14	3.20	0.80	1.44
	4	Geometry	10	57,066	6.16	2.44	0.71	1.32
	5	Measurement & Data Analysis	11	57,066	6.02	3.01	0.79	1.39
6	1	The Number System	15	57,662	8.40	3.83	0.80	1.72
	2	Ratios & Proportional Relationships	10	57,662	5.55	2.67	0.73	1.39
	3	Expressions, Equations, & Inequalities	15	57,662	7.99	3.65	0.78	1.71
	4	Geometry & Measurement	9	57,662	4.68	2.50	0.73	1.29
	5	Data Analysis & Statistics	11	57,662	5.59	2.70	0.70	1.47
7	1	The Number System	14	60,155	7.44	3.30	0.75	1.65
	2	Ratios & Proportional Relationships	9	60,155	4.96	2.46	0.74	1.26
	3	Expressions, Equations, & Inequalities	12	60,155	5.81	3.02	0.75	1.51
	4	Geometry & Measurement	13	60,155	6.43	2.91	0.68	1.63
	5	Data Analysis, Statistics, & Probability	12	60,155	7.12	3.02	0.74	1.52
8	1	The Number System	8	60,987	3.91	2.14	0.66	1.25
	2	Functions	14	60,987	6.99	3.58	0.79	1.63
	3	Expressions, Equations, & Inequalities	14	60,987	6.43	3.28	0.74	1.68
	4	Geometry & Measurement	14	60,987	7.31	3.62	0.81	1.59

Grade	No.	Reporting Category	N Items & Points	N Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	SEM
	5	Data Analysis, Statistics, & Probability	12	60,987	7.60	3.12	0.79	1.43

4.7 Scaling and Scale Evaluation

The purpose of scaling a test is to enhance its validity by increasing the comparability of test takers' scores. This section explicates the way in which the SC READY ELA and mathematics scales are produced to comply with Standard 5.2 of the AERA, APA, & NCME (2014) *Standards*, which states the following:

> The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly. (p. 102)

In this section, the results of the test scaling of the SC READY tests are described and evaluated.

4.7.1 Vertical Scale of SC READY

The structure of the SC READY scale score metric was determined by SCDE staff. In consultation with the TAC, it was decided that SC READY scores would be reported for 2017 onward on across-grade (vertical) scales. The range of scale scores was set to have a minimum score of 100 for all grades and a maximum score varying by grade, starting with 825 in third grade and increasing by 25 points for each subsequent grade, ending at 950 in eighth grade; any scale score that exceeded these limits would be truncated at the limiting value.

To develop the common vertical scale for the SC READY ELA and mathematics tests, five vertical scaling equating blocks were developed. These equating blocks were selected from the lower grade of each pair of adjacent grades (three and four, four and five, five and six, six and seven, and seven and eight). The number of items in each block ranged from 15 to 18 depending on the subject and grade.

These items were selected by DRC test development to meet the following criteria based on the 2016 operational administration:

1) The content of the item must be shared between the lower and upper grade standards and test blueprints, i.e., the students would have had the opportunity to review and practice the skills at the higher grade.
2) The item difficulty of the item at the lower grade cannot be extreme, i.e., the *p*-values for the item must be as close as possible to the mean *p*-value for the lower grade form.
3) The point biserial for the item must have been .25 or higher in the 2016 operational administration.

4) The equating block must be representative of the approximate content area representation on the lower and higher test blueprints.

Exceptions to these rules were made on an item by item basis with psychometric services reviewing the decisions.

The vertical logit scale was developed by equating the ELA and mathematics scales for grades five and six. The link constant was established using Huynh (1994) Robust *Z* methodology. The required arbitrary zero for the grades five and six scale was set at the midpoint between grades five and six. With grades five and six on a scale with a common origin, grades four and seven were separately linked to the common grade five and six scale. This placed grades four, five, six, and seven on a scale with the same origin. With grades four, five, six, and seven scale set, grades three and eight were then linked to the grades four, five, six, and seven scale. The resulting scale linked grades three through eight on a common logit scale, with the origin being the midpoint between the grades five and six scale.

4.7.2 SC READY Scale Scores

Once the common scale of measurement was established using the calibration methods described above, a linear transformation was used to define the reporting metric that would be used to support the testing program.

Calibration of SC READY test forms yielded a value of the Rasch ability, *theta* (ϑ), corresponding to every possible raw score. Scale scores were calculated for every raw score for each grade and subject using the following formula:

$$SS_{unrounded} = 500 + (\vartheta_{RS} / \sigma_{\vartheta}) * 100, \quad (4.22)$$

where ϑ_{RS} is the value of theta corresponding to that raw score and σ_{ϑ} is the initial observed pooled standard deviation of theta for all grades in a subject.

Table 4.31 contains the cut scores (both in the scale score and theta metrics), the pooled σ_{ϑ}, and the LOSS and HOSS for every grade in ELA and mathematics. Cut score values were obtained from the SC READY standard settings.

Final adjustments were made at the LOSS and HOSS for the raw scores of zero, for perfect scores, or for any scale scores that fell outside the LOSS or HOSS. Once these final adjustments were made on the LOSS and HOSS, the scale score CSEM associated with the LOSS and HOSS were computed. For these adjustments, the LOSS or HOSS were first converted to a theta estimate. Then, the CSEM for this theta estimate was obtained as the reciprocal of the square root of the test information function. This CSEM was then expressed on the SC READY ELA and mathematics reporting scale by applying the linear transformation.

Table 4.31. SC READY ELA Scale Score Cuts (Rasch Ability Cuts), Scaling Constants, and LOSS and HOSS

		Achievement Level			Pooled σ_ϑ	LOSS	HOSS
	Grade	Approaches	Meets	Exceeds			
ELA	3	359 (-1.4641)	452 (-0.4966)	540 (0.4156)	1.0387	100	825
	4	419 (-0.8381)	509 (0.0950)	592 (0.9658)	1.0387	100	850
	5	449 (-0.5207)	557 (0.6017)	653 (1.5897)	1.0387	100	875
	6	455 (-0.4632)	575 (0.7876)	667 (1.7443)	1.0387	100	900
	7	512 (0.1265)	615 (1.1969)	704 (2.1242)	1.0387	100	925
	8	537 (0.3938)	642 (1.4838)	737 (2.4673)	1.0387	100	950
Mathematics	3	360 (-1.6080)	438 (-0.7104)	543 (0.5011)	1.1495	100	825
	4	401 (-1.1303)	482 (-0.2112)	563 (0.7269)	1.1495	100	850
	5	448 (-0.5956)	535 (0.4122)	622 (1.4057)	1.1495	100	875
	6	453 (-0.5295)	543 (0.4967)	627 (1.4664)	1.1495	100	900
	7	488 (-0.1347)	577 (0.8935)	649 (1.7219)	1.1495	100	925
	8	527 (0.3113)	615 (1.3267)	683 (2.1136)	1.1495	100	950

4.7.3 Scoring Table Production

WINSTEPS provides a conversion table that maps raw scores to logits (i.e., Rasch model ability estimates) for a given set of item parameters. Score pre-equated conversion tables were produced for each operational form of the SC READY ELA and mathematics assessments administered during spring 2022 school year. The maximum likelihood ability estimates provided in these tables were transformed to scale scores using the linear transformation defined above. Additionally, domain level scoring tables were also generated using WINSTEPS. Score tables can be found in Appendix B.

4.7.4 SC READY Scale Evaluation

The SC READY ELA and mathematics assessments were established on a vertical scale. The scales being established in such a way means that the scale score means for higher grades are higher than the scale score means for lower grades.

The scale score distribution results are discussed in more detail in Section 6.3. However, to summarize the scale score distribution and behavior, as expected the scale scores increase as

the percentile rank increases, showing increasing student ability along the scale for all grades in both content areas.

The Test Characteristic Curve (TCC) and Test Information Function (TIF) curves for ELA forms in grades three through eight are presented in Figures 4.1 and 4.2. The TCC and TIF curves for mathematics forms in grades three through eight are presented in Figure 4.3 and 4.4. The TCCs are *S* shaped, indicating increasing probability of a higher test score as a student's ability increases in both grades.

The TIF curves for SC READY assessment forms are an inverse *U* shape, indicating larger information or smaller errors around ability estimates approximately in the middle of the scale score distribution. The TIF is expected to be lower at the top and bottom ends of the ability scale, where fewer items measuring these students are found. Overall, the test information around the scale score was found to be reasonable for the SC READY assessments.

Figure 4.1: Test Characteristic Curves, Grades 3 through 8 ELA

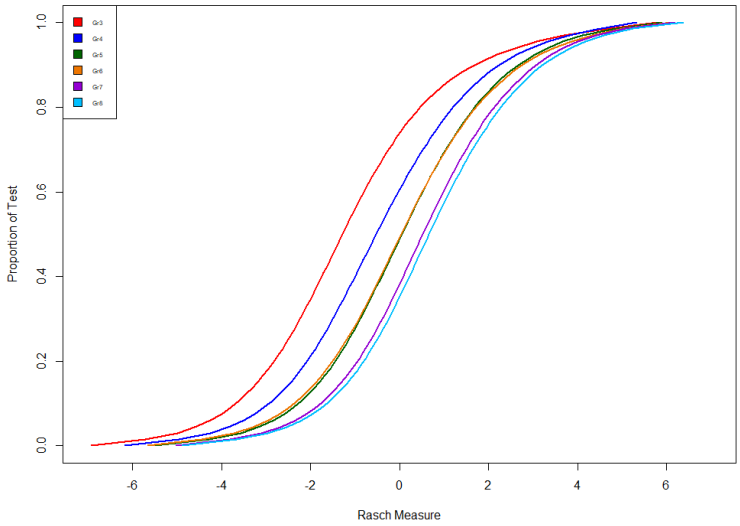


Figure 4.2: Test Information Functions, Grades 3 through 8 ELA

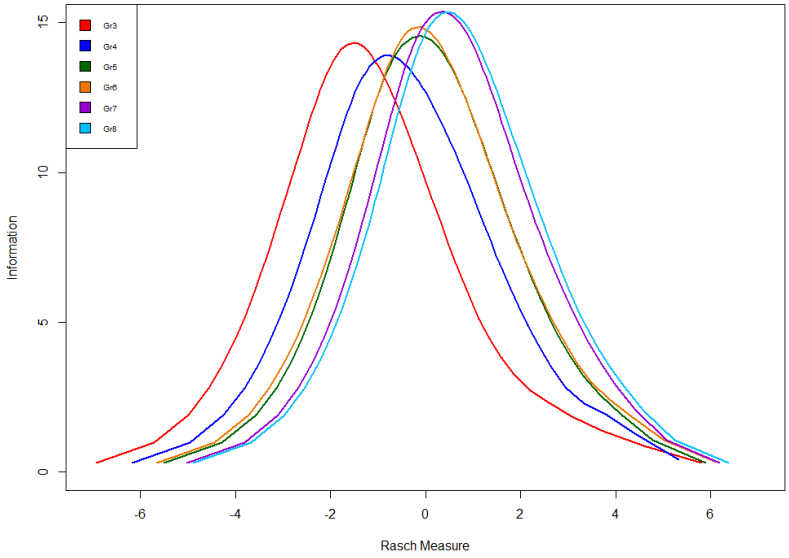


Figure 4.3: Test Characteristic Curve, Grades 3 through 8 Mathematics

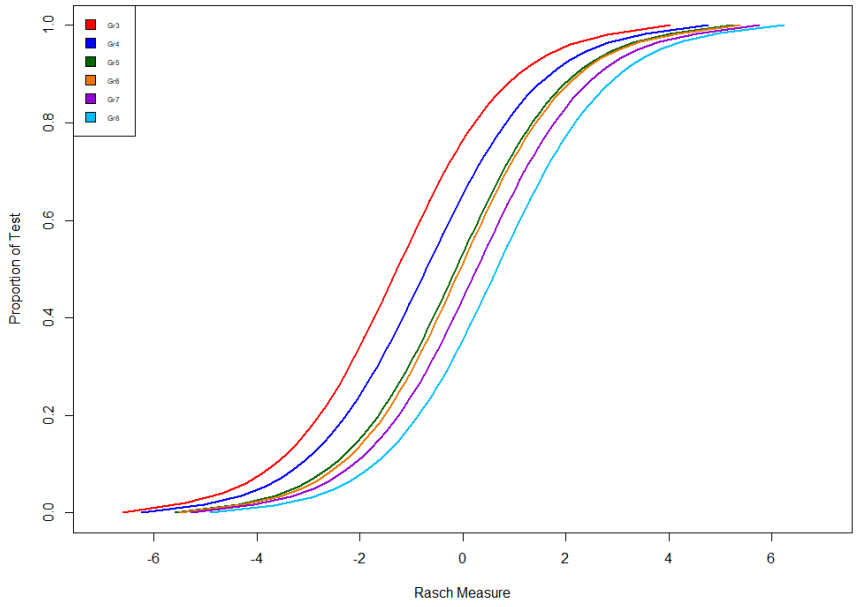
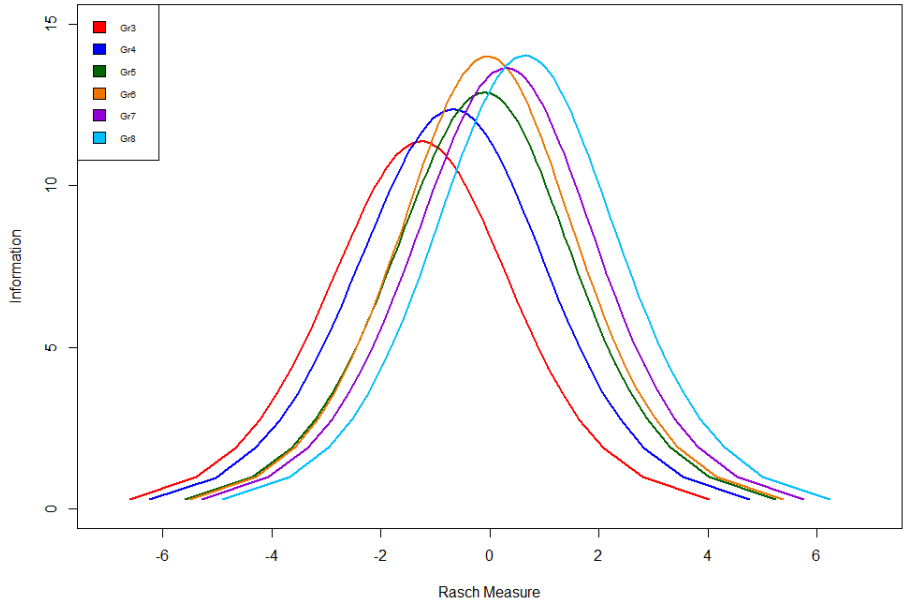


Figure 4.4: Test Information Functions, Grades 3 through 8 Mathematics



4.8 Technical Analyses and Ongoing Maintenance

The appropriate item banks were updated with the operational and field test statistics from the spring 2022 SC READY administration. The item statistics were uploaded to the DRC proprietary

item banking program, IDEAS. The statistics and item cards for the administration were also provided to SCDE.

4.9 Summary

In summary, the information presented in this section summarizes the scoring procedures for different types of items and steps taken by DRC to ensure accuracy in the handscoring and autoscoring processes. The inter-rater reliability statistics presented in Section 4.5.1.3 demonstrate that the handscored items are scored reliably. Additionally, this section reinforces that the overall purpose of the operational data analyses is to ensure that the test items, as well as the overall test, are functioning appropriately. It also helps maintain the test scale across years so that test results may be appropriately compared across years. The data analyses undertaken by DRC are in alignment with multiple best practices of the testing industry and support the following AERA, APA, & NCME (2014) Standards:

- Standard 1.8—The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics.
- Standard 1.13—If the rationale for a test score interpretation for a given use depends on premises about the relationships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.
- Standard 1.21—When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.
- Standard 2.0—Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.
- 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.
- Standard 2.11—Test publishers should provide estimates of reliability/precision as soon as feasible for each relevant subgroup for which the test is recommended.
- Standard 2.13—The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score.
- Standard 2.14—When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.
- Standard 2.16—When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of test takers who would be classified in the same way on two replications of the procedure.

    - Standard 2.19—Each method of quantifying the reliability/precision of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select test takers for reliability/precision analyses and the descriptive statistics on these samples, subject to privacy obligations where applicable, should be reported.
    - Standard 3.1—Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
    - 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.
    - 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.
    - Standard 3.4—Test takers should receive comparable treatment during the test administration and scoring process.
- Standard 3.5—Test developers should specify and document provisions that have been made to test administration and scoring procedures to remove construct-irrelevant barriers for all relevant subgroups in the test-taker population.
- Standard 3.6—Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.
- Standard 4.14—For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure.
- Standard 4.18—Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.
- Standard 4.20—The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring.

- Standard 5.2—The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly.
- 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.
- 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 area specifications 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.
- Standard 6.8—Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment 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.
- 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.
- Standard 7.2— The population for whom a test is intended and specifications for the test should be documented.

Section 5 – Inclusion of All Students

5.1 Procedures for Including Students with Disabilities

All students, including those with a current IEP or 504 Accommodation Plan, must participate in the SC READY ELA and mathematics assessments. A student's IEP team determines whether the student will participate in the assessment in the same manner as other students, with accommodations, or in the alternate assessment, if the student meets alternate assessment eligibility criteria. This complies with AERA, APA, & NCME (2014) Standard 3.9, which states the following:

> Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs. (p. 67)

Guidance for IEP Teams and IEP templates for students in tested grades can be found at https://ed.sc.gov/tests/assessment-information/testing-swd/. The website includes information for testing students with IEPs including the testing process guide, allowable accommodations, training information, testing materials, frequently asked questions, and the *South Carolina Accessibility Support Document*.

5.2 Procedures for Including Multilingual Learners

The SC READY assessments are not available in alternate language formats; all multilingual learner (ML) students must take these tests in English. TAs may not translate any part of the SC READY tests except the test directions. Accommodations should be used only as appropriate for individual students and should not be applied to all ML students indiscriminately. Appropriate accommodations should be based on the English fluency levels of individual students, teacher judgments, and other evidence, including the accommodations used in the classroom for individual students.

Documentation of procedures for determining student eligibility for accommodations and guidance on selection of appropriate accommodations for ML students in tested grades can be found at https://ed.sc.gov/policy/federal-education-programs/esea-title-iii1/ and includes information on enrollment, additional guidance for oral administration procedures for ML students, and the *South Carolina Accessibility Support Document*.

5.3 Accommodations

Students with disabilities or ML students may be provided with test administration accommodations based on their IEPs. Accommodation code definitions can be found in the TAM.

Braille and Large Print test versions were constructed for each grade and content area to enable students who are blind or visually impaired to participate in the SC READY testing.

Universal tools and accommodations are permitted on the SC READY tests. These types of student aids are described below.

- Universal tools are available to all students based on student preference and selection. Some tools, such as a ruler and a digital notepad, are embedded in the online system, while others, such as a physical thesaurus and scratch paper, are not embedded in the system. The availability of universal tools varies by item.
- Accommodations are changes in procedures or materials that increase equitable access during the SC READY assessments. Assessment accommodations allow students to access assessment content area to show what they know and can do. Accommodations are available for students with documented IEPs or 504 Accommodation Plans and for students with limited English proficiency.

Accommodations may be used by students who qualify under the Individuals with Disabilities Education Act (IDEA) and have an IEP, who qualify under Section 504 of the Americans with Disabilities Act and have a 504 Accommodation Plan, or who are identified as ML students. Accommodations must be specified in the qualifying student's individual plan and must be consistent with accommodations used during daily classroom instruction and testing. The use of any accommodation must be indicated on the student information sheet at the time of test administration. AERA, APA, & NCME (2014) Standard 6.2 states the following:

> When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing. (p. 115)

In compliance with this standard, Appendix C of the TAM contains the list of universal tools and accommodations permissible for the SC READY tests. Braille and Large Print forms are provided for blind or visually impaired students.

5.4 Customized Materials

Customized materials include Braille and Large Print materials. Customized test booklets are ordered through precode in the customized test booklet fields.

Accommodations include presentation, scheduling, setting, and timing accommodations. Specific tables, lists, and administration procedures for allowable accommodations can be found in Appendix C of the TAM.

Testing accommodations are documented by test in the IEP under the testing accommodations section (section IX of the IEP). Furthermore, schools and districts input accommodations in PowerSchool, the state's education management system, during precode so that the proper materials may be ordered. There are two means to monitor accommodations use: through

Enrich and through PowerSchool. The state runs annual reports at a state and district level to examine the percentage of students receiving specific accommodations on the statewide assessments. Each district is sent a copy of their report, which compares their district accommodations use to the state data. When the same test has been used for multiple years, the report shows trend data. This trend data allows the state and district to be aware of significant changes in the number of students using an accommodation.

Based on this data, the Office of Assessment and Standards and the Office of Special Education Services, along with the district, can identify areas where additional training on appropriate selection of accommodations is needed. In addition to the data sent to districts, the state data information is shared with stakeholder groups, including DTCs, special education directors, and the Testing and Accountability Roundtable.

South Carolina has a process for reviewing and approving requests for assessment accommodations beyond those routinely allowed. The procedure for special requests can be found in Appendix C of the TAM. There is a committee that reviews special requests. Tables 5.1 and 5.2 present the percentages of students using accommodations in the 2022 testing.

Table 5.1. SC READY Percentage of Students Using Accommodations, ELA

Accommodations	Grade 3 (*N* = 55,905)	Grade 4 (*N* = 56,397)	Grade 5 (*N* = 57,065)	Grade 6 (*N* = 57,566)	Grade 7 (*N* = 60,109)	Grade 8 (*N* = 60,920)
Setting	8.65	8.74	8.23	6.77	6.19	5.95
Timing	1.21	1.20	1.19	0.76	0.60	0.58
Scheduling	0.11	0.09	0.11	0.04	0.05	0.06
Response Options	0.18	0.18	0.17	0.07	0.07	0.07
Presentation	0.19	8.26	7.84	6.68	5.70	5.27
Supplemental Materials	0.04	0.07	0.05	0.05	0.05	0.03

Table 5.2. SC READY Percentage of Students Using Accommodations, Mathematics

Accommodations	Grade 3 (*N* = 55,896)	Grade 4 (*N* = 56,404)	Grade 5 (*N* = 57,066)	Grade 6 (*N* = 57,662)	Grade 7 (*N* = 60,155)	Grade 8 (*N* = 60,987)
Setting	8.22	8.48	8.18	6.71	6.13	5.87
Timing	1.13	1.18	1.18	0.76	0.61	0.57
Scheduling	0.09	0.09	0.08	0.05	0.05	0.06
Response Options	0.09	0.09	0.08	0.04	0.04	0.05
Presentation	11.13	11.17	10.48	8.42	6.96	6.38
Supplemental Materials	0.03	0.07	0.05	0.05	0.05	0.03

5.5 Monitoring Test Administration for Special Populations

The state has a monitoring process for reviewing IEPs. The procedure includes a monitoring overview and rubric for IEP development that is used during onsite monitoring. Results of the

onsite monitoring of IEP development are entered by monitors online. The results of onsite monitoring of IEP implementation are then inputted by monitors online.

According to 2 S.C. Code Ann. Regs. (2015), it is a test security violation to test a student without the accommodations or customized materials specified in the student's IEP or 504 Accommodation Plan (e.g., not providing an oral administration specified in the IEP) or with accommodations or customized materials not specified in the student's IEP or 504 Accommodation Plan. See the TAM for procedures that must be followed to report these security violations.

5.6 Summary

In summary, the information presented in this section is related to allowing access to the assessments for special populations by clearly delineating appropriate universal tools or accommodations and monitoring test administration for special populations. These communication and monitoring efforts by SCDE and the ancillary information developed by DRC are in alignment with multiple best practices of the testing industry and support the following AERA, APA, & NCME (2014) Standards:

- Standard 3.9—Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs.

- 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.

Section 6 – Academic Performance Standards & Reporting

6.1 State Adoption of Academic Performance Standards for All Students

This section briefly describes the SC READY ELA and mathematics standard settings and presents the cut scores derived from the standard setting.

The AERA, APA, & NCME (2014) *Standards for Educational and Psychological Testing* addressed in Sections 6.2 through 6.4 are 5.21 and 5.22, which will each be presented in the pertinent sections.

A brief overview of the standard setting procedures during which the cut scores were derived is presented in Section 6.1.3 of this report, and a detailed discussion and the results of the standard setting may be found in the *2016 SC READY Standard Setting Report* (Data Recognition Corporation, 2016).

The process of the performance level settings for the SC READY ELA and mathematics tests adhered to AERA, APA, & NCME (2014) Standard 5.21, which states the following:

> When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (p. 107)

Standard 5.22 is also relevant and states the following:

> 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 judgments can bring their knowledge and experience to bear in a reasonable way. (p. 108)

In terms of the validity of the SC READY ELA and mathematics cut scores, it is essential to understand that performance level descriptions and cut scores are established in a collaborative and participatory process. The performance level descriptions clearly establish, in plain language, the proper frame of reference for understanding how to interpret test scores, particularly cut scores.

6.1.1 Performance Level Setting

On June 5 through 9, 2016, the SCDE and DRC conducted a standard setting for the SC READY ELA and mathematics tests. The purpose of this workshop was to develop performance standards for both content areas, including the development of cut points, which divide students into four performance levels: *Does Not Meet Expectations*, *Approaches Expectations*, *Meets Expectations*, and *Exceeds Expectations*.

6.1.2 Methodology

For SC READY ELA and mathematics standard settings, South Carolina educators from across the state participated in the Bookmark Standard Setting Procedure (BSSP) (Lewis, Mitzel, & Green,

1996; Lewis, Mitzel, Mercado, & Schulz, 2012) for each SC READY ELA and mathematics grade. In the BSSP, these educators recommended cut scores for the SC READY ELA and mathematics assessments.

6.1.3. Workshop

During the performance level setting, participants studied the South Carolina performance level descriptors (PLDs) and *SC CCR Standards for ELA and Mathematics* to review the knowledge, skills, and abilities expected of students in each performance level. Each performance level was associated with a level of mastery of *SC CCR Standards for ELA and Mathematics.* Participants then discussed the content-based expectations for students at the threshold of each performance level (e.g., a student who is just *"Meets Expectations"*). Participants studied ordered item booklets (OIBs) that comprised collections of operational test items that were ordered by difficulty. A separate OIB was created for each test, and items' difficulty values were based on students' performance on the test items. Participants studied the OIBs to understand the knowledge and skills measured by the tests.

Participants engaged in three rounds of individual judgments and group discussions. In each round, participants recommended cut scores by considering the content-based expectations for students in each performance level and then identifying the sets of items in their OIBs that best represented these expectations. By placing bookmarks, participants recommended cut scores on the test scale. Between rounds, participants were shown feedback (e.g., median bookmarks, impact data). The committees' median judgments were taken as their recommendations.

Following the initial committee meetings, a vertical articulation panel was selected from the various grade-level committees to examine proposed standards across grades and recommend possible adjustments. After the committee meetings, SCDE considered the grade-level committee recommendations, accompanied by their associated standard errors, and suggested modifications by the vertical articulation panel. SCDE also considered results of other assessments and policy implications before editing the final achievement standards.

For the SC READY ELA and mathematics standard settings, DRC computed standard errors around the panelists' cut score recommendations to aid SCDE in making policy adjustments prior to finalizing the cut scores. Once the policy adjustments were made, the cut scores were presented to the State Superintendent of Education for approval. During standard setting (and as a standard process in the Bookmark procedure), panelists were able to articulate the cut scores with the PLDs if they were judged to have been modified during the standard setting procedure.

6.1.4 Performance Levels

Performance standards were set during the Bookmark procedure (explained under Section 6.1.3 Workshop).

The following general verbal descriptions of the SC READY ELA and mathematics performance levels were given to the standard setting committees:

Does Not Meet Expectations – The student does not meet expectations as defined by the grade-level content area standards.

Approaches Expectations – The student approaches expectations as defined by the grade-level content standards.

Meets Expectations – The student meets expectations as defined by the grade-level content standards.

Exceeds Expectations – The student exceeds expectations as defined by the grade-level content standards.

More detailed descriptions of the specific concepts and skills are provided for each grade level in the PLDs. PLDs are narrative descriptions of the knowledge and skills expected at each of the four performance levels. The PLDs are based on the state-adopted content area standards. In some cases, the standard setting committees made revisions to the PLDs. A copy of the detailed PLDs can be found in Appendix B of DRC's 2016 SC READY standard setting report (Data Recognition Corporation, 2016).

6.1.5 Cut Scores

Table 6.1 provides the ELA and mathematics cut scores in terms of the scale scores.

Table 6.1. SC READY ELA and Mathematics Scale Score Ranges

Content Area	Grade	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*
ELA	3	100 - 358	359 - 451	452 - 539	540 - 825
	4	100 - 418	419 - 508	509 - 591	592 - 850
	5	100 - 448	449 - 556	557 - 652	653 - 875
	6	100 - 454	455 - 574	575 - 666	667 - 900
	7	100 - 511	512 - 614	615 - 703	704 - 925
	8	100 - 536	537 - 641	642 - 736	737 - 950
Mathematics	3	100 - 359	360 - 437	438 - 542	543 - 825
	4	100 - 400	401 - 481	482 - 562	563 - 850
	5	100 - 447	448 - 534	535 - 621	622 - 875
	6	100 - 452	453 - 542	543 - 626	627 - 900
	7	100 - 487	488 - 576	577 - 648	649 - 925
	8	100 - 526	527 - 614	615 - 682	683 - 950

6.2 Reporting

This section contains information on the results of the spring 2022 administration of the SC READY ELA and mathematics assessments. The scale score and performance level summaries for the total population of South Carolina students are presented here. Presenting the results by performance level translates the quantitative scale provided through scale scores into a

qualitative description of student performance, using the following terms: *Does Not Meet Expectations*, *Approaches Expectations***,** *Meets Expectations*, and *Exceeds Expectations*.

While the scale score provides an essential quantitative reference for student performance, the performance level information plainly outlines the meaning of the scores to parents, students, and educators. When combined, scale scores and performance levels provide a comprehensive set of tools to assess South Carolina student performance by content area and grade level.

All results presented in this section are based on South Carolina student census data from DRC. The results presented here may differ slightly from the official state summary report of all student populations due to ongoing resolution of test materials and student information. The results in the tables in this section are presented as evidence of the reliability and validity of the intended interpretation of scores from the SC READY ELA and mathematics assessments and should not be used for state accountability purposes.

6.2.1 Reports

Score reports are the primary means of communicating test scores to relevant district personnel (i.e., Test Coordinators or superintendents), teachers, and parents. AERA, APA, & NCME (2014) Standard 6.10 states the following:

> 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 scores are intended to be used. (p. 119)

Standard 5.1 is related in that it states the following:

> Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretation of scale scores, as well as their limitations. (p. 102)

Interpretations related to the test scores are disseminated in two ways: the individual score report and the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021).

In addition to providing interpretation, it is important that the information related to the test scores is understandable by the target audience. Standard 7.0 of the AERA, APA, & NCME (2014) *Standards* states the following:

> Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores. (p. 125)

In support of Standard 7.0, the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021) is accessible to parents, teachers, and laypeople alike.

The individual student report is the primary means for sharing student test results with parents. As such, it should be a stand-alone document, giving parents relevant information so they

understand their child's test score. In the 2021-2022 administration year, DRC reported the SC READY ELA and mathematics scores using paper and/or electronic reports in the DRC INSIGHT Portal, which is a browser-based system designed to deliver online interactive reporting to authorized users at the state and district levels for South Carolina public schools.

6.2.2 Description of Each Type of Report

In this section, descriptions of the following reports are provided: Student Roster, Individual Student Report, Student Score Label, Item Summary Report, and Results Summary Report.

In compliance with AERA, APA, & NCME (2014) Standard 12.18, the SC READY score reports provide clear information about the achievements of individual students and groups of students. Standard 12.18 states the following:

> In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (p. 200)

6.2.2.1 Student Roster

Available from the Portal is a Student Roster that displays a list of students based on the specific report filter options selected, such as test administration, grade, school, district, gender, race/ethnicity, and examiner. Total test scale scores and performance level indicators, as well as the reporting category scale scores, are displayed in a table format for the content area chosen. A sample Student Roster is provided in the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021).

6.2.2.2 Individual Student Report

The Individual Student Report (ISR) is another type of report available as a paper material and through the Portal. The two-page ISRs are provided to schools to be sent home to parents for each content area. On the top of the page, the student's identifying information is provided. Scale scores are provided at the top of the page along with the student's performance level. In the middle of the page, a bar graph and the student's scale score for a given content area are shown, along with the performance level associated with that scale score. This information is provided with a brief explanation of what the performance level means.

For ELA, the ISR shows both an overall ELA total test score and a reading score. Both are shown at the top of the page with the scale score and performance level information and on the bar graph. Additionally, for ELA, the student's Lexile range is provided, which describes the student's performance for the reading test. For mathematics, the student's quantile range is provided, which describes the student's level of mathematics achievement.

The SEM is indicated in a range format around the total test scale scores and the reporting category scale scores on the ISRs. The SEM represents the amount of variability that can be expected in a student's test or reporting category score due to the inherent imprecision of the

test. In other words, the SEM represents a range of scale scores in which the student's score would likely fall if the student took the same test again.

On page two, the ISR provides the student's score history, percentile rank comparisons, and student performance by reporting category. The percentile rank is the percentage of students in the comparison group who scored the same as or below a student's score. The ISR shows the student's percentile ranks for two comparison groups: students in South Carolina and students in other states with comparable standards. Also on page two is the student's performance for each reporting category; reporting category performance is classified as "Low," "Middle," or "High." This classification is based on the subset of items that assess the reporting category. For ELA, the results from the text-dependent analysis (TDA) item are provided on page two.

Sample ISRs for both ELA and mathematics are provided in the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021).

6.2.2.3 Student Score Label

The Student Score Label is designed so that each student's test results can be placed in the student's permanent record. A label is provided for every student who participates in the spring administration of the SC READY tests. Each label has a self-adhesive backing so that it can be peeled from the sheet and placed in the student's cumulative school record. The label presents a snapshot of the student's results on the SC READY tests. The label lists the student's scale score, performance level, and percentile ranks for each content area. DRC provided multiple labels per student submitted for scoring. The labels are provided in print only. A sample Student Score Label report is provided in the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021).

6.2.2.4 The DRC INSIGHT Portal

Schools and districts can access summary level reports through the Portal, DRC's online assessment management and reporting system. The Portal allows school district personnel with appropriate permissions to access SC READY data at a variety of levels and to request customized reports that are configured and disaggregated in ways that best meet their needs for such activities as evaluating programs, revising curricula, and improving teaching and learning. Users access the Portal from https://www.drcedirect.com. Each school and/or district is assigned a username and password to access the site.

6.3 Interpreting Test Results

The student's correct responses to the assessment questions are used to derive that student's SC READY ELA and mathematics scale scores. The scale score describes performance on a continuum that ranges from 100 to 950 for ELA and mathematics. Since the ELA and mathematics assessments are on a vertical scale, the scale scores can be compared across grades. Scale scores cannot be compared across content areas. For example, it is not appropriate to compare a student's ELA and mathematics scores as they do not represent comparable achievement.

The SC READY scale scores determine a student's performance level. Student performance is reported in terms of performance levels, and each performance level represents standards of performance for each assessed content area. Performance level scores provide a description of what students can do in terms of the content area and skills assessed, as described in the *SC CCR Standards for ELA and Mathematics*.

In addition to the total test score, students receive information on their performance in each reporting category of the test taken. The reporting category performance level information is classified as *Low*, *Middle*, or *High*. This classification is based on the subset of items that assess the standard.

Additional information on score interpretation is included in the *SC READY Score Report User's Guide* (Data Recognition Corporation, 2021), which was developed collaboratively by DRC and SCDE staff.

6.4 Current Administration Results

Results for the spring 2022 SC READY administration can be found in Tables 6.2 and 6.3, which provide summaries of the total test scale scores based on the state population for the 2022 administrations of the ELA and mathematics assessments, respectively.

Table 6.2. State-Level SC READY Scale Score Summary Statistics, ELA

Grade	*N*	Mean SS	SD SS	Percentile				
				10th	25th	50th	75th	99th
3	55,905	442.03	126.75	270	342	449	535	711
4	56,397	511.36	127.31	341	407	515	618	783
5	57,065	548.36	111.71	393	458	552	633	792
6	57,566	556.71	118.09	401	464	557	650	806
7	60,109	596.30	117.51	445	499	589	681	856
8	60,920	621.49	116.51	458	526	623	715	871

Table 6.3. State-Level SC READY Scale Score Summary Statistics, Mathematics

Grade	*N*	Mean SS	SD SS	Percentile				
				10th	25th	50th	75th	99th
3	55,896	452.51	122.62	298	357	443	531	825
4	56,404	478.46	115.01	339	388	459	554	808
5	57,066	526.16	115.06	392	439	508	599	849
6	57,662	518.62	110.00	401	431	495	589	861
7	60,155	541.74	104.29	426	464	523	603	832
8	60,987	571.14	107.38	448	486	557	635	872

Tables 6.4 and 6.5 present the percentages of students in each performance level in the spring 2022 test administration. The percentage of students classified as *Meets Expectations* or *Exceeds Expectations* in ELA was found to be slightly higher at grade four and lower at grade

five but was otherwise consistent and ranged from approximately 39% in grade five to approximately 46% in grade four. The percentage of students classified as *Meets Expectations* or *Exceeds Expectations* in mathematics was found to be higher in lower grade levels compared to higher grade levels and ranged from approximately 30% in grades seven and eight to approximately 47% in grade three.

Table 6.4. State-Level Percentages of Students in Each Performance Level, ELA

Grade	N	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	55,905	29.0	23.0	23.6	24.4	48.0
4	56,397	28.0	21.6	19.8	30.6	50.4
5	57,065	22.1	29.7	27.9	20.4	48.3
6	57,566	23.1	31.9	25.4	19.7	45.1
7	60,109	26.9	29.8	23.1	20.1	43.2
8	60,920	27.2	26.9	29.2	16.7	45.9

Table 6.5. State-Level Percentages of Students in Each Performance Level, Mathematics

Grade	N	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	55,896	26.8	22.2	26.6	24.4	51.0
4	56,404	28.7	27.9	19.6	23.7	43.4
5	57,066	29.5	27.2	23.6	19.7	43.3
6	57,662	33.8	30.5	19.3	16.5	35.7
7	60,155	35.8	33.3	15.3	15.5	30.9
8	60,987	40.8	29.0	14.1	16.0	30.2

6.5 Longitudinal Comparison of Test Results

It is often desirable to examine the scores of students across time and monitor group performance. This is possible if the test content area and the construct measured by the test are comparable from year to year and if the scores are reported on the same scale in multiple years. Tables 6.6 and 6.7 show the across-year scale score summary statistics for ELA and mathematics, respectively. These tables show that the overall student performance in ELA and mathematics shows increasing trends for most grades. Note that due to a federal waiver, there was no spring 2020 testing.

For ELA, when comparing the 2017 through 2019 scale score means, the largest gain occurred in grade four and the smallest occurred in grade six. Comparing 2019 to 2021, all grades except six and seven for ELA, and all grades in mathematics showed mean scale score decreases.

When comparing 2021 to 2022 for ELA, grade seven showed a slight decrease in mean scale score, where grades three through six and grade 8 showed increases in mean scale score. For mathematics, when comparing the 2017 through 2019 scale score means, the largest gain occurred in grade three and the smallest occurred in grade six. When comparing 2021 to 2022 for mathematics, all grades showed an increase in scale score mean.

Table 6.6. State Level Scale Score Means 2017—2022, ELA

Grade	Year	Student Count	Scale Score Mean	Scale Score Standard Deviation
3	2017	59,740	432.80	98.50
	2018	59,902	439.63	101.48
	2019	57,236	445.84	120.45
	2021	51,313	427.26	122.59
	2022	55,905	442.03	126.75
4	2017	60,052	481.81	102.35
	2018	60,319	489.26	108.48
	2019	60,227	505.20	126.74
	2021	51,060	495.81	130.49
	2022	56,397	511.36	127.31
5	2017	57,739	522.69	104.24
	2018	60,829	526.35	104.89
	2019	61,030	529.09	109.07
	2021	51,435	527.12	109.51
	2022	57,065	548.36	111.71
6	2017	56,413	543.96	105.66
	2018	58,402	544.76	114.01
	2019	61,413	544.78	116.08
	2021	50,681	549.11	117.08
	2022	57,566	556.71	118.09
7	2017	55,849	580.01	101.29
	2018	56,935	586.26	108.75
	2019	58,969	595.63	116.77
	2021	50,972	597.16	119.66
	2022	60,109	596.30	117.51
8	2017	55,049	613.39	107.34

Grade	Year	Student Count	Scale Score Mean	Scale Score Standard Deviation
	2018	55,967	612.81	113.60
	2019	57,055	620.76	116.09
	2021	50,359	615.47	116.58
	2022	60,920	621.49	116.51

Table 6.7. State Level Scale Score Means 2017—2022, Mathematics

Grade	Year	Student Count	Scale Score Mean	Scale Score Standard Deviation
3	2017	59,887	453.94	113.22
	2018	59,909	462.21	118.28
	2019	57,238	465.77	122.38
	2021	51,268	439.67	119.64
	2022	55,896	452.51	122.62
4	2017	60,287	483.62	103.76
	2018	60,342	491.08	110.26
	2019	60,262	491.47	108.55
	2021	51,119	468.64	109.06
	2022	56,404	478.46	115.01
5	2017	57,908	524.66	107.55
	2018	60,845	532.88	113.63
	2019	61,039	534.59	114.21
	2021	51,489	516.75	116.81
	2022	57,066	526.16	115.06
6	2017	56,549	534.72	110.78
	2018	58,418	535.69	118.36
	2019	61,452	535.87	116.16
	2021	50,963	513.23	108.87
	2022	57,662	518.62	110.00
7	2017	56,055	548.24	97.00
	2018	56,964	552.62	105.51
	2019	59,009	551.03	103.34
	2021	51,230	536.38	100.95
	2022	60,155	541.74	104.29
8	2017	55,203	584.50	99.4
	2018	55,982	591.00	106.3

Grade	Year	Student Count	Scale Score Mean	Scale Score Standard Deviation
	2019	57,077	592.00	111.2
	2021	50,480	570.98	108.85
	2022	60,987	571.14	107.38

In addition to the scale score longitudinal data for the SC READY ELA and mathematics tests, the across-year performance level data were examined. Tables 6.8 and 6.9 show the percentages of students in each performance level in the spring 2017, spring 2018, spring 2019, spring 2021 and spring 2022 test administrations for ELA and mathematics, respectively. In these four years, South Carolina students were classified into performance levels based on the same set of cut scores set during the standard setting for each content area (see Section 6.1).

As seen in Table 6.8, for 2017 through 2019, the percentage of students classified as *Meets Expectations* or *Exceeds Expectations* for ELA tended to increase or remain the same between adjacent administrations. When comparing 2019 to 2021, for ELA, there were slight decreases in the percentage of students classified as *Meets Expectations* or *Exceeds Expectations*, except for grade six. When comparing 2021 to 2022, for ELA, there were increase in the percentage of students classified as *Meets Expectations* or *Exceeds Expectations* for all grades.

Table 6.9 shows, for 2017 through 2019, that the percentage of students classified as *Meets Expectations* or *Exceeds Expectations* for mathematics tended to increase or remain the same between adjacent administrations. When comparing 2019 to 2021, for mathematics, there were slight decreases in the percentage of students classified as *Meets Expectations* or *Exceeds Expectations*. When comparing 2021 to 2022, for mathematics, there were increases in the percentage of students classified as *Meets Expectations* or *Exceeds Expectations*.

Table 6.8. Comparison of SC READY Performance Level Distribution 2017—2022, ELA

Grade	Year	Student Count	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	2017	59,740	26.1	31.8	26.5	15.6	42.1
	2018	59,902	23.2	31.7	28.3	16.8	45.2
	2019	57,236	25.5	24.7	26.6	23.1	49.7
	2021	51,313	31.8	24.9	23.9	19.3	43.3
	2022	55,905	29.0	23.0	23.6	24.4	48.0
4	2017	60,052	29.5	29.5	26.8	14.1	40.9
	2018	60,319	28.2	28.0	24.6	19.2	43.9
	2019	60,227	27.8	21.0	22.7	28.6	51.2
	2021	51,060	32.5	21.4	20.3	25.7	46.1

Grade	Year	Student Count	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
	2022	56,397	28.0	21.6	19.8	30.6	50.4
5	2017	57,739	28.0	33.7	27.2	11.1	38.3
	2018	60,829	27.5	33.6	26.0	12.9	38.9
	2019	61,030	28.2	30.9	26.3	14.7	41.0
	2021	51,435	27.9	33.2	24.2	14.7	38.9
	2022	57,065	22.1	29.7	27.9	20.4	48.3
6	2017	56,413	23.7	36.6	26.0	13.6	39.7
	2018	58,402	25.5	34.6	24.6	15.3	39.9
	2019	61,413	26.1	32.8	24.7	16.4	41.0
	2021	50,681	25.5	32.7	24.9	16.9	41.8
	2022	57,566	23.1	31.9	25.4	19.7	45.1
7	2017	55,849	28.4	35.2	23.3	13.0	36.4
	2018	56,935	28.9	31.0	23.9	16.3	40.1
	2019	58,969	27.0	29.0	25.6	18.4	44.0
	2021	50,972	28.4	29.1	21.6	20.9	42.5
	2022	60,109	26.9	29.8	23.1	20.1	43.2
8	2017	55,049	28.0	31.9	26.9	13.2	40.1
	2018	55,967	29.1	31.6	23.1	16.2	39.2
	2019	57,055	27.2	28.2	27.8	16.7	44.6
	2021	50,359	30.1	28.0	26.1	15.9	41.9
	2022	60,920	27.2	26.9	29.2	16.7	45.9

Table 6.9. Comparison of SC READY Performance Level Distribution 2017—2022, Mathematics

Grade	Year	Student Count	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
3	2017	59,887	22.2	25.2	30.9	21.6	52.5
	2018	59,909	21.5	22.8	29.9	25.8	55.7
	2019	57,238	21.3	21.0	30.7	27.0	57.7
	2021	51,268	31.0	22.1	26.5	20.3	46.9
	2022	55,896	26.8	22.2	26.6	24.4	51.0
4	2017	60,287	24.1	29.5	25.4	21.0	46.4
	2018	60,342	25.1	26.8	22.4	25.7	48.1
	2019	60,262	24.3	25.2	24.6	25.9	50.5
	2021	51,119	32.5	25.6	21.8	20.1	42.0
	2022	56,404	28.7	27.9	19.6	23.7	43.4
5	2017	57,908	27.9	32.1	20.8	19.2	40.0
	2018	60,845	27.0	27.8	23.1	22.1	45.2

Grade	Year	Student Count	*Does Not Meet*	*Approaches*	*Meets*	*Exceeds*	*Meets + Exceeds*
	2019	61,039	24.5	30.2	22.3	23.1	45.4
	2021	51,489	33.4	28.5	19.1	19.0	38.1
	2022	57,066	29.5	27.2	23.6	19.7	43.3
6	2017	56,549	25.6	32.9	21.7	19.8	41.5
	2018	58,418	28.8	28.6	20.4	22.2	42.6
	2019	61,452	27.5	28.7	22.7	21.2	43.9
	2021	50,963	35.5	30.6	18.0	15.8	33.9
	2022	57,662	33.8	30.5	19.3	16.5	35.7
7	2017	56,055	31.0	35.8	17.1	16.2	33.3
	2018	56,964	32.6	32.5	16.8	18.2	34.9
	2019	59,009	32.7	32.0	17.2	18.1	35.3
	2021	51,230	36.5	33.1	15.8	14.6	30.4
	2022	60,155	35.8	33.3	15.3	15.5	30.9
8	2017	55,203	31.5	34.0	18.6	15.9	34.5
	2018	55,982	32.2	31.2	18.5	18.2	36.6
	2019	57,077	32.1	31.3	16.9	19.7	36.6
	2021	50,480	40.2	29.1	15.6	15.2	30.7
	2022	60,987	40.8	29.0	14.1	16.0	30.2

6.6 Summary

In summary, the overall purpose of reporting test results is to communicate information on student performance to stakeholders. These results are presented in the context of score reports that aid the user in understanding the meaning of the test scores. The reports and ancillary information developed by DRC are in alignment with multiple best practices of the testing industry and support the following AERA, APA, & NCME (2014) Standards:

- 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.
- 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.
- 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 judgments can bring their knowledge and experience to bear in a reasonable way.
- 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 scores are intended to be used.

- Standard 7.0—Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.
- Standard 12.18—In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and by supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

References

ACT. (n.d.). ACT College and career readiness standards. https://www.act.org/content/act/en/college-and-career-readiness/standards.html

Altman, D. G. (1991). Practical statistics for medical research. London: Chapman and Hall.

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.

Anderson, Lorin W., and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook 1: Cognitive domain*. New York, NY: McKay.

Brennan, R. L. (2004). BB-Class (Version 1.0) [Computer software]. Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement & Assessment.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.

Cattell, R. B. (1952). *Factor analysis*. New York, NY: Harper.

Center for Universal Design. (1997). *What is universal design?* Raleigh, NC: Center for Universal Design, North Carolina State University.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.2307/1165285

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

College Board (2009). College Board Standards for College Success. http://programs.ccsso.org/content/pdfs/State%20level_christine%20cassidy.pdf

College Board (2015). Test specifications for the redesigned SAT. https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group/Thomson Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Data Recognition Corporation. (2016). *June 2016 SC Ready Standard Setting Technical Report*. Columbia: South Carolina Education Oversight Committee.

Dorans, N. J., & Schmitt, M. P. (1991). *Constructed response and differential item functioning: A pragmatic approach*. Princeton, NJ: Educational Testing Service.

Green, D. R. (1975). *Procedures for assessing bias in achievement tests* [Paper presentation]. National Institute of Education Conference on Test Bias, Annapolis, MD.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Publishing.

Holland, P., & Thayer, D. (1986, April). *Differential item performance and the Mantel-Haenszel procedure* [Paper presentation]. American Educational Research Association annual meeting, San Francisco, CA.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*(4), 1–8.

Huynh, H. (1994). "Some Technical Aspects in Standard Setting." In Proceedings of the Joint Conference on Standard Setting for Large Scale Assessment Programs, 75–91. Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Huynh, H., & Meyer, P. (2010). Use of robust *z* in detecting unstable items in item response theory models. *Practical Assessment, Research & Evaluation*, *15*(2), 253–64.

Indiana Department of Education. (2014). Mathematics. https://www.doe.in.gov/standards/mathematics#MathAcademic

Indiana Department of Education. (2020). English/language arts standards. https://www.doe.in.gov/standards/englishlanguage-arts

International Reading Association & National Council of Teachers of English. (1996). *Standards for the English language arts* https://secure.ncte.org/library/NCTEFiles/Resources/Books/Sample/StandardsDoc.pdf?_ga=2.203002457.933648301.1586210788-836936228.1586210788

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lewis D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring* [Symposium]. Council of Chief State School Officers 1996 National Conference on Large-Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 225–254). New York, NY: Routledge.

Linacre, J. M. (2018). WINSTEPS Rasch measurement (Version 4.2.0) [Computer software]. Beaverton, OR: Winsteps.com.

Linn, R., & Gronlund, N. (1995). *Measurement in assessment and teaching* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*(2), 179–197.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local

independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

Massachusetts Department of Elementary and Secondary Education. (2001). 2001 Massachusetts English language arts curriculum framework. **Error! Hyperlink reference not valid.**

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of a general theory. *Multivariate Behavioral Research*, *14*, 21–38. https://doi.org/10.1207/s15327906mbr1401_2

Minnesota Department of Education. (2007). *Minnesota academic standards: mathematics K–12*. file:///C:/Users/skendallen/Downloads/MathStandards2007%20005247%20reposted%2092815.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *The common core state standards for mathematics*. Washington, DC.

National High School Center. (2012). College and career development organizer. **Error! Hyperlink reference not valid.**

Nebraska Department of Education. (2014). Nebraska English language arts standards. **Error! Hyperlink reference not valid.**

Nebraska Department of Education. (2015). Nebraska's college and career ready standards for mathematics. https://www.education.ne.gov/math/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Schumacker, R. E., & Muchinsky, P. M. (1996). Disattenuating correlation coefficients. *Rasch Measurement Transactions*, *10*(1), 479.

Texas Higher Education Coordinating Board & Texas Education Agency. (2009). *Texas college and career readiness standards*. Austin, TX.

Thompson, S., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (National Center on Educational Outcomes Synthesis Report 44). Minneapolis, MN: University of Minnesota.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233–251.

Zwick, R., & Thayer, D. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*(3), 187–201.