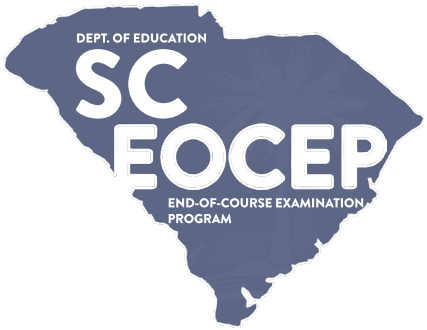
South Carolina

End-of-Course Examination Program

2023–2024 Operational Test Technical Report



Issued by the

South Carolina Department of Education

Office of Assessment and Standards

Division of College, Career, and Military Readiness

Ellen Weaver

State Superintendent of Education

Table of Contents

Table of Tables and Figures

Executive Summary

This report is a technical summary of the 2023–2024 administration of the End-of-Course Examination Program (EOCEP) tests. The EOCEP assessments are designed to measure students' knowledge of Algebra 1, Biology 1, English 2, and U.S. History and the Constitution (USHC). The assessments are aligned to the state's academic standards. There are five performance levels for each assessment: A, B, C, D, and F.

Test forms for the 2023–2024 administration year were developed by Data Recognition Corporation (DRC), under the supervision of the South Carolina Department of Education (SCDE), and were directly aligned to the appropriate standards. The EOCEP is administered online, but paper formats are available for students whose IEP, Section 504 Plan, or ILAP requires paper testing.

E.1 Overview of the EOCEP Assessments

To meet federal accountability requirements, the EOCEP exams in English/language arts, mathematics, and science must be administered to all public-school students, including those students as required by the federal Individuals with Disabilities Education Improvement Act (IDEA) and by Title 1 of the Elementary and Secondary Education Act (ESEA).

To receive a South Carolina high school diploma, students are required to pass a high school credit course in science and a high school credit course in United States history that include the state's end-of-course examinations in Biology 1 and USHC respectively.

E.2 Administration

The EOCEP test administration has three testing windows in a typical year: Fall/Winter, Spring, and Summer.

E.3 Student Performance

The EOCEP student assessment reports a scale score. Each scale score is assigned a letter grade equivalent (A, B, C, D, or F) in accordance with the South Carolina Uniform Grading Scale (UGS). Scale scores range from 0–100 for each EOCEP test. In addition to the total scale scores, students' performances in every reporting category for each EOCEP assessment are classified in one of four ordinal categories: *Does Not Meet*, *Minimally Meets*, *Meets*, and *Exceeds*.

Table E1. State-Level Percentages of Students in Each Performance Level, Algebra 1, Biology 1, English 2, and USHC

EOCEP	*N*	Does Not Meet- *Letter Grade F*	Minimally Meets- *Letter Grade D*	Meets - Letter *Grades C & B*	Exceeds - *Letter Grade A*
Algebra 1	67,719	27.6	21.8	37.4	13.3
Biology 1	62,784	37.1	15.6	28.1	19.1
English 2	64,660	14.2	16.3	43.7	25.8
USHC	58,699	41.3	15.0	25.3	18.5

Note. Data combines Fall/Winter, Spring, and Summer Administrations.

E.4 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 EOCEP test scores. The EOCEP 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 EOCEP assessments.

Evidence of validity based on test content area was supported by the test specifications, including the test design and test blueprint. The EOCEP assessments were developed in alignment with the South Carolina Academic Standards for each content area. 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 EOCEP assessments are primarily administered in an online format, but paper-and-pencil, Braille, and Large Print forms are available for students who require them. The EOCEP 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 required by their Individualized Education Programs (IEP), 504-Plans or Individualized Language Acquisition Plans (ILAP). 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) and constructed response item types, followed predefined scoring criteria. The selected-response, multi-select, and TE items were

machine-scored (details are included in Section 4.5). The English 2 assessments contained one handscored text-dependent analysis (TDA) item.

The test scaling and equating was conducted using item response theory (IRT) methodology. Students' scale scores were derived using a pre-equated item pool. The IRT models used for EOCEP 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 performance levels and associated performance level descriptors were established during the performance level settings for each EOCEP. Performance level setting for each EOCEP was performed in a collaborative and participatory process, further supporting the validity and interpretation of the EOCEP 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 EOCEP tests produce scores that would be relatively stable if the tests were administered repeatedly under similar conditions. The assumption that the content area EOCEP tests were unidimensional (that is, that each subject test measured one primary dimension) was confirmed through principal component analysis. The evidence of the validity of the intended interpretation of the EOCEP test scores based on the relationships with other variables was evaluated through the correlations computed between EOCEP content areas. The student scores were found to be moderately 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—Statewide System of Standards & Assessments

1.1 History and Overview

The South Carolina Education Accountability Act of 1998 requires the administration of the state's end-of-course examinations in gateway courses for which credit in English language arts, mathematics, science, and social studies is awarded. Students must take the appropriate End-of-Course Examination Program (EOCEP) tests if they are enrolled in courses in which the curriculum standards for Algebra 1, Biology 1, English 2, and U.S. History and the Constitution (USHC) are taught. In the beginning of the 2017–18 school year, to meet federal accountability requirements, the EOCEP tests in Algebra 1, English 2, and Biology 1 were required to be administered by the third year of high school to all public school students who did not qualify for South Carolina Alternate Assessments, including those students as required by the federal Individuals with Disabilities Education Improvement Act (IDEA) and by Title 1 of the Elementary and Secondary Education Act (ESEA).

As they are enunciated in State Board of Education Regulation 43-262, the purposes and uses of the EOCEP tests are as follows:

- *The examinations shall encourage instruction in the specific academic standards for the courses, encourage student achievement, and document the level of students' mastery of the academic standards.*
- *The examinations shall serve as indicators of program, school, and school district effectiveness in the manner prescribed by the Education Oversight Committee in accordance with the provisions of the Education Accountability Act of 1998 (EAA).*
- *The examinations shall be weighted 20 percent in the determination of students' final grades in the gateway courses.*

Since the beginning of the 2017–18 school year, EOCEP scores have been reported on the basis of the South Carolina Uniform Grading Policy (UGP) as revised in 2016 and updated in 2019 (see the South Carolina Uniform Grading Policy for more information). The score reported is a scale score and not the percentage of correct answers.

The Algebra 1/Mathematics for the Technologies 2 end-of-course examination was implemented in the baseline year 2002–03 and was operational for the first time in 2003–04. The English 1, Physical Science, and Biology 1/Applied Biology 2 examinations were field-tested in May 2003 and implemented for the baseline year in 2003–04. These subject-area EOCEP examinations became operational in 2004–05. The Biology 1/Applied Biology 2 examination was discontinued after the 2005–06 school year. The State Board of Education reinstated the Biology 1 test with a field test

in 2008. Additional field-testing was conducted in Spring 2009. The 2009–10 school year was an implementation year for Biology 1. The first operational administration for Biology 1 was Fall/Winter 2010–11. The last administration of Physical Science was in Spring 2011. The U.S. History and the Constitution examination was field-tested in 2005–06, with baseline implementation in 2006–07 and a second implementation in 2007–08. The first operational administration was in 2008–09. The South Carolina State Board of Education later adopted the South Carolina Social Studies College- and Career-Ready Standards in 2019, and the first operational administration of the new USHC test occurred during Fall/Winter 2021–2022.

Beginning with the Fall/Winter 2017–18 administration, the English 1 examination was administered in two separate sections, English 1-Reading and English 1-Writing. The English 1-Writing section includes a text-dependent analysis item in addition to multiple-choice writing items. Field-testing for English 1-Writing was conducted during the Spring, prior to the implementation of the operational English 1 assessment. In 2020–21, English 1 was only administered to specific students as needed for previous graduation requirements and accountability.

Beginning with the Fall/Winter 2020–21 administration, the English 2 examination was administered in two separate sections, English 2-Reading and English 2-Writing. The English 2-Writing section includes a text-dependent analysis item in addition to multiple-choice writing items. Field-testing for English 2-Writing was conducted during Spring 2019. Additionally, the English 2 test did not count toward the student's course grade in 2019–20, but it counted starting in the 2020–2021 school year.

EOCEP exams are delivered primarily in an online format. The first opportunity for online testing was for Adult Education students in Fall/Winter 2004–05. The opportunity for online testing was expanded to include most other students in Spring 2005. The proportion of EOCEP exams administered online has increased steadily. Beginning in 2017–18, all students test online except for those students whose Individualized Educational Program (IEP), Section 504 Plan, or ILAP requires a paper test.

The South Carolina Department of Education (SCDE) awarded the contract for the development and scoring of the EOCEP tests in October 2001 to American Institutes for Research (AIR) and its partners Insite, Inc., and Pearson Educational Measurement (PEM). In Spring 2007, PEM became the sole contractor. In fall 2008, Data Recognition Corporation (DRC) took over administration while PEM remained the development contractor. DRC became the sole contractor in mid-2013.

Until 2014–2015, all EOCEP exams contained only selected-response operational items. The online versions of the Fall/Winter 2014–2015 and Spring 2015 English 1 EOCEP exams included a small number of technology-enhanced operational items. Currently, all subjects include technology-enhanced, constructed response operational items.

In this report, all data are based on the students in public middle and high schools or adult education programs only. Data on students in district-approved homeschools have been excluded.

1.2 Groups Involved with the EOCEP Assessments

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

1.2.1 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.

1.2.2 Technical Advisory Committee

The Technical Advisory Committee (TAC) makes recommendations to the SCDE on issues regarding field test design, item analysis, linking, 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.

1.2.3 Contractors and Other Groups

In addition to SCDE staff members, contractors and SC educators were involved in EOCEP development and administration. DRC was contracted to provide test administration, scoring, and reporting services. Under the current contract, DRC participates in a number of item and test development, review, implementation, and data analysis activities.

1.3 State Adoption of Academic Content Standards for All Students

South Carolina has adopted challenging academic content standards for all students in English language arts, mathematics, science, and US history and the constitution. The 2014 South Carolina Academic Standards and Performance Indicators for Science were, with one exception, adopted by the State Board of Education (SBE) on January 8, 2014, and by the Education Oversight Committee (EOC) on February 10, 2014. The exception consisted of a single high school Biology 1 standard (H.B.5). The existing (2005) version of this standard remained in effect until the 2021 Biology 1 standards were approved and implemented. Full implementation and assessment of the 2014

Biology 1 standards began with the 2016–17 school year. The 2015 South Carolina College- and Career-Ready Standards for English Language Arts and South Carolina College- and Career-Ready Standards for Mathematics were approved by the EOC on March 9, 2015, and received final approval by the SBE on March 11, 2015. The assessment of these standards for Algebra 1 and English 1 began with the 2015–16 school year, and English 2 began in the 2020–21 school year, replacing English 1. The South Carolina State Board of Education adopted the current South Carolina Social Studies College- and Career-Ready Standards in 2019. The South Carolina College- and Career-Ready Science Standards 2021 were adopted in 2021. The assessment of these standards for Biology 1 began with the 2023–2024 school year.

1.4 Coherent & Rigorous Academic Content Standards

The South Carolina College- and Career-Ready Standards for English Language Arts and Mathematics are aligned with the entrance requirements for credit-bearing coursework in the system of public higher education in South Carolina and relevant career and technical education standards as evidenced by the certification letters from several South Carolina universities. The South Carolina Academic Standards and Performance Indicators for Science and the South Carolina College- and Career-Ready Standards for English Language Arts and Mathematics were linked to the statewide ACT® assessment. The technical link to the College- and Career-Ready Standards and the ACT® college readiness benchmark is discussed in more detail in Section 6.2.

EOCEP assessment items are aligned to the South Carolina Academic Standards for each content area. Standards describe what schools are expected to teach and what students are expected to learn. Academic standards are statements of the specific cognitive processes and the content knowledge and skills that students must demonstrate to meet the grade-level standards. EOCEP test items are written to assess the content knowledge and skills described in the academic standards. The academic standards and supporting documents are available on the South Carolina Department of Education website.

1.5 Required Assessments

To meet federal accountability requirements, the EOCEP in English language arts, mathematics, and science must be administered to all public school students, including those students as required by the federal Individuals with Disabilities Education Improvement Act (IDEA) and by Title 1 of the Elementary and Secondary Education Act (ESEA).

To receive a South Carolina high school diploma, students are required to pass a high school credit course in science and a high school credit course in United States history. In these courses, the state's end-of-course examinations in Biology 1 and in USHC are administered.

Gateway courses in English language arts, mathematics, science, and social studies will be defined by the State Board of Education. EOCEP examination scores count 20 percent of a student's final grade in gateway courses. Defined gateway courses currently include Algebra 1, Biology 1, English 2, and USHC or other courses in which the academic standards corresponding to these subjects are taught. The USHC test administered during the 2021–2022 school year was newly developed in response to the state's adoption of new US History and the Constitution Standards in 2019. The state's requirement to use the USHC scores in course grade calculations was waived because the USHC 2019 academic standards differed substantially from the standards they replaced, so that scoring the test could not occur until after standard setting.

1.6 Data Reporting

The EOCEP student assessment reports a scale score. Each scale score is assigned a letter grade equivalent (A, B, C, D, or F) in accordance with the South Carolina Uniform Grading Scale (UGS). Possible scale scores range from 0–100 for each EOCEP. In addition to the total scale scores, students' performances in every reporting category for each EOCEP assessment are classified in one of three ordinal categories: Low, Middle, and High. This classification is based on the subset of items that assess the reporting category. Section 6.3 includes a full explanation.

1.7 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 (EAA) of 1998. Gateway courses in English/language arts, mathematics, science, and social studies are required per Section 59-18-310 of the EAA.

All public middle school, high school, alternative school, virtual school, and adult education students enrolled in courses in which the academic standards corresponding to the EOCEP subjects are taught, regardless of course name or number, must take the appropriate end-of-course test. This testing policy includes all students with IEPs or Section 504 Plans, suspended students, home school students who are registered through the district or local school board, homebound students, and home-based students. Also included are Multilingual Learner (ML) students, charter school students (including virtual charter schools), and incarcerated students.

For students with documented disabilities, the decision about a student's participation in the EOCEP or alternate assessment is made by the student's IEP team and documented in the IEP for the student. All students with documented disabilities with IEPs or Section 504 Plans must have the necessary accommodations documented.

1.7.1 Special Groups of Students

1.7.1.1 Students with Disabilities—Students with disabilities must participate in EOCEP in accordance with their Individualized Education Program (IEP) or Section 504 Plan. (See Appendix C of the TAM for guidelines on administering the test to students with disabilities.)

Adult Education Students with Disabilities—Students with disabilities in adult education centers who are 21 years of age or younger and do not have diplomas may be served under the provisions of IDEA or Section 504. The IEP or Section 504 Plan must state any accommodations to be used. Students older than 21 cannot be served under the provisions of IDEA but may be served under Section 504. Students who are older than 21 must prove they are disabled (e.g., provide documentation that they were served under an IEP or a Section 504 Plan in high school) prior to taking the test.

Suspended and Expelled Students (with or without disabilities)—Students who are suspended must be tested. The district may consider delaying the suspension dates, bringing students into school during suspension for testing purposes only, or testing students in alternative locations. The district is not required to test expelled students who do not have IEPs. When a student with an IEP has been expelled, a new IEP must be written that outlines the services to be provided during the expulsion period and the manner in which the student will be tested.

Home School Students—Home school students are defined as those students whose parents or guardians teach their children at home. Students whose home school instruction is approved by the district board of trustees of the district in which the student resides must be tested according to S.C. Code Ann. §59-65-40 (A)(6)(2004): "The tests must be administered by a certified school district employee either with public school students or by special arrangement at the student's place of instruction, at the parent's option. The parent is responsible for paying the test administrator if the test is administered at the student's home." It is recommended, but not required, that a monitor accompanies the TA if the parent chooses to have the student tested at home. Parents, guardians, or other relatives may not be present in the room with the student during testing. Home school students will receive individual student results but will not be included in the district or school data.

Homebound Students—Homebound students (with or without disabilities) must be tested. These students receive instruction at home or in a hospital because they cannot attend school due to illness, accident, or pregnancy, even with the aid of transportation [2 S.C. Code Ann. Regs. 43-241 (2011)]. The district must administer the required tests to a student who is homebound, except in individual cases in which it is documented that the student is not physically and/or mentally able to take the test. It is the district's decision to choose whether to have a monitor present when testing homebound students. Homebound students may be tested online with a district-owned laptop or with paper/pencil by request.

Homebased Students—Students who receive homebased instruction must be tested. Homebased students normally receive instruction at a place other than school because the student's IEP team has determined this placement to be the most appropriate, least restrictive environment for the administration of the student's educational program. The district must send a TA to the place of instruction.

Multilingual Learner (ML) Students—ESOL/EL students enrolled in courses in which the curriculum standards corresponding to EOCEP subjects are taught must take the appropriate tests.

Foreign Exchange Students—Foreign exchange students who meet the EOCEP eligibility criteria must participate in EOCEP testing.

Students with Disabilities Who Have Been Placed by Districts and Public Agencies in Private or Nonpublic Schools—Students with disabilities who are placed by districts or other public agencies in private or nonpublic schools or state-operated programs must participate in statewide and districtwide assessments and must be tested by the home school district. The home school district is the district that carries the student on enrollment and receives state or federal funding for educating the student.

Students who are placed by other public agencies through the foster home/group home proviso, General Appropriations Act, 2003 S.C. Acts 91 Proviso 1.9, must be tested by the district in which the alternate residence (such as a foster home, group home, orphanage, or state-operated health care facility including a facility for treatment of mental health or chemical dependence) is located.

1.8 Participation Data

All schools administered EOCEP tests to students who completed courses in which the standards for Algebra 1, Biology 1, English 2, and U.S History and the Constitution were taught. Summary data are reported for operational tests only. For federal accountability, the Algebra 1, English 2, and Biology 1 EOCEP tests were administered to all public school students except those who qualify for alternate assessments, including those who are not enrolled in the traditional gateway credit-bearing courses.

Demographic data were collected for each student. These data included the categories of sex, test mode, grade, race/ethnicity, English language proficiency (limited English proficiency, or LEP), Individualized Education Plan (IEP), migrant status, gifted/talented, Section 504 Plan status, and accommodations. Table 1.1 presents the combined student participation in the two EOCEP administrations (Fall/Winter 2023, Spring 2024, and Summer 2024) sorted by demographic variables.

Table 1.1. Summary of Student Demographics, Fall/Winter, Spring, and Summer Forms Combined

Demographic	Group	Algebra 1		Biology 1		English 2		USHC	
		N	%	*N*	%	*N*	%	*N*	%
All Students	**All**	67,719	100.00	62,784	100.00	65,064	100.00	58,699	100.00
Sex	Female	32,305	47.70	30,713	48.92	31,896	49.02	29,308	49.93
	Male	34,531	50.99	31,516	50.20	32,488	49.93	28,811	49.08
	Missing	883	1.30	555	0.88	680	1.05	580	0.99
Ethnicity	Hispanic or Latino	8,950	13.22	8,189	13.04	8,625	13.26	7,580	12.91
	American Indian or Alaska Native	179	0.26	197	0.31	205	0.32	157	0.27
	Asian	1,110	1.64	1,117	1.78	1,151	1.77	1,076	1.83
	Black or African American	20,170	29.78	19,105	30.43	19,775	30.39	17,687	30.13
	Native Hawaiian or Other Pacific Islander	68	0.10	55	0.09	64	0.10	72	0.12
	White	30,403	44.90	28,439	45.30	29,143	44.79	27,272	46.46
	Two or More Races	3,403	5.03	3,101	4.94	3,104	4.77	2,587	4.41
	Missing	3,436	5.07	2,581	4.11	2,997	4.61	2,268	3.86
IEP Status	Yes	7,308	10.79	6,320	10.07	6,384	9.81	5,128	8.74
	No or missing	60,411	89.21	56,464	89.93	58,680	90.19	53,571	91.26
Gifted Status	Academic only	11,327	16.73	9,823	15.65	10,402	15.99	7,718	13.15
	Artistic only	1,149	1.70	1,027	1.64	1,062	1.63	1,103	1.88
	Both	1,263	1.87	854	1.36	873	1.34	749	1.28
	No or unknown	53,980	79.71	51,080	81.36	52,727	81.04	49,129	83.70
Section 504 Plan Status	Yes	3,268	4.83	3,043	4.85	3,175	4.88	3,069	5.23
	No or missing	64,451	95.17	59,741	95.15	61,889	95.12	55,630	94.77
English Proficiency Status	Active Multilingual Learners	4,496	6.64	3,962	6.31	4,306	6.62	3,448	5.87
	Met ELP Exit Criteria in Monitoring Period	1,376	2.03	842	1.34	791	1.22	665	1.13
	Title III Exited	2,406	3.55	2,661	4.24	2,816	4.33	2,660	4.53
	English Speaker II	59,376	87.68	55,291	88.07	57,107	87.77	51,908	88.43
	All others	65	0.10	28	0.04	44	0.07	18	0.03
Migrant Status	Yes	30	0.04	22	0.04	26	0.04	22	0.04
	No or missing	67,689	99.96	62,762	99.96	65,038	99.96	58,677	99.96

Demographic	Group	Algebra 1		Biology 1		English 2		USHC	
		N	**%**	***N***	**%**	***N***	**%**	***N***	**%**
Customized Material	Braille	<10	0.01	<10	0.01	<10	0.01	<10	0.01
	Sign Language signed administration	14	0.02	11	0.02	13	0.02	10	0.02
	Large print	11	0.02	12	0.02	12	0.02	15	0.03
	Oral administration	3,391	5.01	2,945	4.69	2,879	4.42	1,865	3.18

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

Section 2—Assessment System Operations

2.1 Test Design & 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 EOCEP tests. Content-related evidence of the validity of the intended score interpretations in EOCEP 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 EOCEP assessment's consistent adherence to the assessment blueprints, which were constructed by South Carolina educators based on the South Carolina Academic Standards for each content area.

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 EOCEP assessments is well documented through the item and test development process, including the review of the assessment items for alignment to the South Carolina Academic Standards for each content area that each of the EOCEP tests measure.

The information found in this section provides an overview of the South Carolina Academic Standards and the process used for the development of the blueprints for EOCEP assessments. 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 EOCEP tests. Test development and psychometric staff worked extensively with the SCDE in the construction of all test forms to support pre-administration equating. All form construction activities focused on ensuring that test blueprint requirements were met and 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 test is selected to match the test blueprint and the psychometric specifications of previous operational administrations.

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. 85)*

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 EOCEP assessments is the assessment blueprint, which specifies the target score points for each content area standard. The overall structure of the EOCEP design is found in Table 2.1 for the Fall/Winter and Summer administrations. Since the Spring administration contains field test items, the overall structure of the Spring assessments is found in Table 2.2.

The 2023–2024 EOCEP operational forms matched the test blueprints found in Tables 2.3 through Table 2.6, including the actual point distributions. Actual point distributions on the 2023–2024 EOCEP operational forms matched blueprint targets. Additional blueprint information can be found on the SCDE's website. Note that reporting category results were not reported for the 2023–2024 Biology 1 tests.

Table 2.1. EOCEP Test Design, Fall/Winter and Summer

EOCEP	No. of OP Items	No. of OP Points
Algebra 1	50	50
Biology 1	50	50
USHC	55	55
English 2	55	70

Note. English 2 includes a 16-point TDA item.

Table 2.2. EOCEP Test Design, Spring

EOCEP	No. of OP Items	No. of Embedded FT Items	Total No. of Items	No. of OP Points
Algebra 1	50	8	58	50
Biology 1	50	10	60	50
USHC	55	8	63	55
English 2	55	10	65	70

Note. English 2 includes a 16-point TDA item.

Table 2.3. Algebra 1 Blueprint

Reporting Categories	No. of Standards	No. of Items per Reporting Category
Algebra	20	21–25
Functions	18	18–22
Number and Quantity Interpreting Data	9	8–11

Table 2.4. Biology 1 Blueprint

Reporting Categories	No. of Indicators	No. of Items per Reporting Category
Structure and Processes	4	10–15
Ecosystems	3	10–14
DNA and Heredity	3	11–14
Biological Evolution	4	10–15

Note: Reporting category results were not reported for Biology 1 for the 2023–2024 school year.

Table 2.5. USHC Blueprint

Reporting Categories	No. of Skills	No. of Items per Reporting Category
Foundations of American Republicanism	6	10–12
Expansion and Union	6	10–12
Capitalism and Reform	6	10–12
Modernism and Interventionism	6	10–12
Legacy of the Cold War	6	10–12

Table 2.6. English 2 Blueprint

Reporting Categories or Standards/Indicators	No. of Indicators	No. of Items per Reporting Category
Reading Literary Text - RL	8	16–26
Reading Informational - RI	8	18–25
Writing	4	6–12
Communication	2	2–6
Inquiry	3	4–8
TDA Item	1	1*

Note: For reporting, the Writing and Communications Strands are combined into one category. Inquiry is included in the total ELA score. TDA Item totals 16 points.

2.1.2 Item Types

The EOCEP tests include a variety of item types. Selected-response items ask students to select the best answer from four answer options. Multi-select items require students to select more than one answer from five or six options. There are also a variety of technology-enhanced constructed response item types, which involve different ways of responding on the computer (e.g., drag and drop, matching, drop-down list, etc.).

EOCEP English 2 consist of two separate sections (Writing and Reading). The Writing section includes a text-dependent analysis (TDA) item. The TDA item requires students to read a passage and write an essay using information from the passage to support their answer. The Reading section includes evidence-based selected-response (EBSR) items. For these two-part items, students first read a piece of text or a passage and choose the best answer from the choices. Students are then asked to support their response with evidence from the text. In order to receive credit for an EBSR item,

students must answer both parts correctly. EOCEP USHC field test items also include the EBSR item type.

2.1.3 Universal Design

The EOCEP 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 assessments may reduce the need for accommodations by reducing or eliminating access barriers associated with the tests themselves. Table 2.7 presents the elements of universal design that were implemented on the EOCEP assessments (Center for Universal Design, 1997; Thompson & Thurlow, 2002).

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 2023–2024 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 EOCEP 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 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 2.7. 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. On all test forms, items that include pictures or diagrams have descriptive captions that are accessible for students with visual disabilities.
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.

2.1.4 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)*

All EOCEP items were developed with reference to the South Carolina Academic Standards for each content area and measurement guidelines. All newly developed items were reviewed by committees of South Carolina educators for content area and fairness 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 committees evaluate 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, appropriate for the grade level, 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 unfairness 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 EOCEP assessments include several types of items including selected-response, technology-enhanced, EBSR, multi-select, and TDA. DRC content specialists used the following steps in the preparation of items for the EOCEP program:

1. Establish item/scenario development specifications and style guides and prepare item writing training manuals.
2. Determine item development plans.
3. Train item writers and/or scenario developers in the project requirements and specifications.
4. Develop passages and write items.
5. Review, edit, code, and track items and produce graphics.

6. Produce review forms for content and fairness/sensitivity reviews by external reviewers.
7. Modify items based on external reviewers' recommendations.
8. Review and approve field test–ready items and scenarios.
9. Develop field test forms and administer field test.
10. Internally review field test item data.
11. Approve items to be included in the item bank.

2.1.5 Content and Fairness Reviews

All newly developed items are put through a rigorous review process within DRC, and fairness and sensitivity issues are addressed by DRC content area specialists and a DRC fairness 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 that meet the following criteria:

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

Fairness and sensitivity reviews occurred with a panel of education professionals. The separate Fairness and Sensitivity Committee received training in addition to the general training given to content area reviewers. This training focused on fairness and sensitivity issues, how to identify them, and why it was important to avoid them in the interest of all students who would be responding to an item. Reviewers were introduced to their task through a welcome letter directing them how to access the items they'd be reviewing, as well as support materials to guide their fairness and sensitivity reviews.

During the Fairness and Sensitivity meeting with South Carolina educators, DRC also provided training on the procedures and forms used for item content and fairness review. The content and fairness 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 EOCEP assessments constituted the pools of items from which subsequent test forms for future administrations were created. The number of EOCEP items developed is summarized in Table 2.8. Note that the Text Dependent Writing (TDW) items for English 2 were developed for use beginning with the 2024–2025 assessments. These items were administered as part of a stand-alone field test in spring 2024 rather than embedded within the operational English 2 assessments.

Table 2.8. EOCEP Item Development

Item Type	Number of Items Developed for Each EOCEP			
	Algebra 1	**Biology 1**	**USHC**	**English 2**
SR	234	65	127	136
EBSR	0	0	11	12
MS	0	6	5	0
TE	26	73	16	12
Stimuli Set	NA	12	8	0
TDW				12

Note: SR = Selected Response, EBSR = Evidence-Based Selected Response, MS =- Multi-Select, TE = Technology Enhanced, TDW = Text Dependent Writing Prompt

2.1.6 Data Review

Spring field test item data were reviewed internally by SCDE and DRC 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 unfairness, 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.

After internal review, field test items that were flagged for any of the criteria listed in Section 2.1.7 went through an external review process with South Carolina educators. During this external data review meeting, a DRC content area test development specialist led a committee of South Carolina educators through review of each flagged item, focusing on the nature of the statistical flag, to determine whether the item was flawed. DRC psychometricians and representatives from the SCDE also attended the meeting to help the committee interpret statistical information or to help answer questions. The educators were asked to come to a consensus decision about whether to reject or accept each flagged item. Accepted items were added to the pool of items available for subsequent operational use in the EOCEP assessments. Rejected items were designated for revision and re-field testing on a future form or removed from the item pool entirely.

2.1.7 Field-Testing

This section of the report describes the timeline and process of field-testing EOCEP items for future operational use on an EOCEP assessment. For each content area, items accepted during the content and fairness reviews were then field-tested prior to the operational assessment. The online test forms were spiraled at the student level.

Once field test data are available, field test item analyses will include:

- classical item analysis
- item *p*-values (difficulty)
- item-total test correlation
- percentage of students selecting incorrect responses
- point-biserial correlation for incorrect responses on the selected-response (SR) items
- score point distribution for items worth more than one point
- omit rates for all items
- differential item functioning (DIF) analysis
- Rasch analysis

More details on classical item analysis methodology are provided in Section 4.6.1 of this report.

DIF analysis examines potential item unfairness and determines 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 sex, race/ethnicity, and accommodation use. More details on the DIF methodology are provided in Section 4.3.2 of this report.

After completion of the field test item analyses discussed above, item statistics are then reviewed 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 would be reviewed by the DRC content area specialists, South Carolina educators, and the SCDE prior to their selection as operational items on any future EOCEP 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.

2.1.8 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 guidelines for excellence, they must also meet technical guidelines 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, which includes reviewing whether each item matches the given standard. The entire pool of items is available 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 check to see that each item meets technical guidelines for well-crafted items, including having only one clearly correct answer for selected-response 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 unfair to demographic subgroups.

The construction of the test forms themselves is 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 unfair content; and,
- met the Principles of Universal Design and maximum accessibility.

Recommended psychometric properties of the items included:

- a *p*-value between 0.30 and 0.85;

- an item-total test correlation > 0.20;
- omit rates $\leq 5\%$;
- an acceptable item fit (no misfit flag); and,
- no DIF large (“C”) 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 unfairness but no content reason can be determined for the unfairness.

Scoring tables are generated for each selection, and it is recommended that the raw score cuts at each score point do not deviate within plus or minus 2 inclusive from the six previous forms’ median raw score cuts at each performance level.

In addition to the core online operational test forms, emergency forms and paper-and-pencil accommodated forms, including Braille and Large Print forms, were available for the 2023–2024 administrations. SCDE reviewed the items placed on the operational test forms during the form construction meeting. To mitigate the impact of potential test security breaches, the SCDE and DRC also strive to minimize or eliminate the amount of shared items across different forms of the assessment.

New field test items were embedded in the Spring test forms only. The field-test items embedded in the field-test positions were accepted for field-testing during the content and fairness 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 were involved in the review and approval of the forms. SCDE approval for 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 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.

2.1.9 Multiple Assessment Forms

The majority of EOCEP tests are computer-based assessments, but paper-based tests are available. 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 EOCEP 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 EOCEP 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.

2.2 Test Administration

The next section examines how test administration procedures implemented for the EOCEP 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 EOCEP 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 explained 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 EOCEP assessments. Homeschool students can participate in EOCEP testing, but their records are not included in the summary. The characteristics of the students who were eligible to participate in the administered 2023–2024 assessments are presented in Table 1.1. The number of students ranged from approximately 59,000 to approximately 68,000 per EOCEP assessment.

During the Fall 2023, Spring 2024, and Summer 2024 administrations, approximately 50% of South Carolina students who participated were identified as female and about 50% of students were identified as male. In terms of student ethnicity, approximately 45–46% of students were identified as White, approximately 30% of students were identified as Black or African American, approximately 13% of students were identified as Hispanic or Latino, less than 2% of students were identified as Asian, less than 1% were identified as American Indian or Alaska Native, and less than 1% of students were identified as Native Hawaiian or Other Pacific Islander. In addition, approximately 4–5% of students were identified as having Two or More Races. Approximately 9–11% of students were identified as having an IEP, 16–20% of students were identified as having a gifted status, and around 5% were identified as having a Section 504 Plan. Approximately 6–7% of students were identified as active multilingual learners. Less

than 1% of students were identified as migrants. Lastly, approximately 3–5% of students used customized materials.

2.2.2 EOCEP Testing Window

The test administration dates for the current year are given in Table 2.9. For Algebra 1, Biology 1, English 2 Reading, and USHC, the district testing window must not begin earlier than the last 15 instructional days of the semester or school year. For English 2 Writing, the district testing window must not begin earlier than the last 20 instructional days of the semester or school-year. English Reading and Writing sections must not be administered on the same day. Makeup testing was provided for students who missed the originally scheduled EOCEP test due to a death in the family, illness, or another situation deemed valid by the state. It was recommended that a single makeup test be given per day, but two could have been given per day if necessary. For all EOCEP administration windows, district test coordinators (DTCs) were responsible for providing the testing schedule to all school test coordinators (STCs) in their districts.

Table 2.9. EOCEP Test Administration Windows

Administration	State Testing Windows
Fall/Winter	November 27–January 26
Spring	April 26–June 12
Summer	June 26–August 2

2.2.3 Test Time

The EOCEP tests were not timed; however, each session had to be administered during a single day (unless a student's IEP or Section 504 Plan specifically stated that she or he 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. 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. The majority of students finished the test within 2 hours, as Tables 2.10 and 2.11 show, with the exception of students in English 2, which are 2-part tests administered over 2 days.

Table 2.10. EOCEP Test Time Distribution (In Minutes), Fall/Winter Testing Window

Subject	*N* Items	25th Percentile	Median	75th Percentile
Algebra 1	50	67	87	111
Biology 1	50	60	77	99
English 2 – Reading	41	58	80	106
English 2 – Writing	14	61	81	104
USHC	55	58	75	95

Table 2.11. EOCEP Test Time Distribution (In Minutes) Spring Testing Window

Subject	*N* Items	25th Percentile	Median	75th Percentile
Algebra 1	58	79	101	128
Biology 1	60	69	90	114
English 2 – Reading	44	59	80	106
English 2 – Writing	21	58	76	99
USHC	63	60	77	99

2.2.4 Test Administration Manuals

Test administration procedures and guidelines for the EOCEP 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) Standards 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 (TAMs), exemplified in Appendix A, were created for each Fall/Winter and Spring administration; the Spring TAM is also used for referencing each Summer. 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 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 subject specific. 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 carefully follow 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 EOCEP 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 without modifying 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 for All Subjects” (for more information, see the TAM).

Online Tools Training (OTT) tutorials and practice tools are provided, in advance of the EOCEP assessment windows, in all content areas to familiarize students/users with the navigation of the online systems, the functionality of the testing environment, and the different item types. Districts have the following options for training students on interacting with the DRC 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 number of times a student/user can access the OTT.
- Online Tutorials give students/users the ability to watch narrated 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 EOCEP assessments are administered according to the prescribed test schedule.

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 administration of the EOCEP assessments. 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 EOCEP assessments, examiners are instructed to transcribe students' responses from the Large Print test or Braille test book into the 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 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 that 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 Section 504 Plan team determines how, not if, a student with disabilities participates in the EOCEP 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.

The Standards explain that 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 returning physical materials has been greatly reduced. 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.

Materials in customized formats were sent only to the schools and only in the quantities ordered. Due to the small quantity of paper testing, overage materials are no longer shipped automatically. Schools may order additional materials in DRC's INSIGHT portal.

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.

2.2.7 Administrative Support and Training

To ensure that the EOCEP 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, SCDE contracted with DRC to provide training for the DTCs. DRC provided Technology Coordinator training on September 13, 2023.

The DTC training was held via a webinar for Fall/Winter administration on October 18, 2023, and for Spring 2024 on March 6, 2024. An STC/TA Training PowerPoint was

posted for districts and schools on November 1, 2023 for EOCEP Fall/Winter administration and March 13, 2024, for the Spring administration.

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 EOCEP test administration are outlined in the TAMs. The manuals include 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 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:*
>
> - *Give examinees access to test questions prior to testing;*
> - *Copy, reproduce, or use in any manner inconsistent with test security regulations all or any portion of any secure test booklet;*
> - *Coach examinees during testing or alter or interfere with examinees' responses in any way;*
> - *Make answer keys available to examinees;*
> - *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;*
> - *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 generates a missing document report listing the identification numbers of unreturned secure materials. The report is used to notify districts of missing materials. A toll-free telephone line is provided to answer questions regarding missing documents, and follow-up procedures are employed until all materials are accounted for. Subsequently, the districts locate and return the materials or send signed statements indicating that all secure materials have been returned.

2.3.1 Secure Materials

Secure materials—each assigned a human- and machine-readable security identification number—are test booklets, customized test materials, and administration scripts. For online testing, secure materials consist of 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 and Standards staff conducts on-site monitoring of test administrations to verify district and school compliance with policies and procedures as outlined in the TAMs. Monitoring is defined as an announced or 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 Section 504 Plan.

This section describes how the EOCEP 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 Test Security Committee develops a list of school sites to be monitored. The list of prioritized sites is constructed based on data forensics with additional consideration of the other named sources of information. Qualitative and quantitative data are examined and compared. Data is used to determine improbable gains in test scores as well as an unusual number of answer changes resulting in correct answers. Surveys, test security violations, training attendance records, calls, and emails are also considered in site selection. Individual members of the Test Security Committee are called upon to share their knowledge of the events that suggest a site's addition to the monitoring list.

Selected SCDE monitors are encouraged to identify additional schools near the selected sites for routine visits if it is deemed feasible. Visits to those additional schools

are approached as an opportunity for school staff to clarify test practices, reinforce the culture of good test administration procedures, and provide feedback on their school's testing program to the SCDE Office of Assessment and Standards.

As SCDE monitors observe, they are guided by the Monitor Observation Checklist that is completed online in a Google Form. If a test security violation is observed by the SCDE monitor, the SCDE monitor will request that the STC submit a Test Security Violation to their District Test Coordinator along with appropriate documentation. The SCDE monitor will inform the Test Security Program Manager of the violation observation immediately upon conclusion of the monitor visit. At the end of the visit, the SCDE monitor may discuss their findings with the STC and the school principal if time and circumstances allow.

Monitoring teams make visits to school systems as assigned. The teams not only make visits to individual schools but also sometimes visit the district office to review school system testing plans procedures. Visits may occur at any time during the testing window, and more than one visit may be made to a specific school if needed. Additional test observation or additional documentation review and interviews with personnel may be needed. Each individual monitor or monitoring team completes a checklist to document observations and make general notes related to test administration policy and procedures that are observed during the monitored test administration.

Following the on-site monitoring visit, checklists and notes of the school's testing processes and procedures are reviewed. Any strengths and weaknesses in the school's testing procedures are noted along with suggestions on follow-up action, if needed. This information is summarized in a letter that is emailed to the District Test Coordinator of the school district visited along with a copy sent to the school principal. If, during a monitoring visit, a potential test security violation is observed, the SCDE monitor will discuss the concern with the Test Security Program Manager who will then contact the District Test Coordinator to discuss any information that may impact the TAM or protocol or procedures for that particular test.

Upon completion of the monitor visit, the SCDE monitor enters the information from the Monitor Observation Form into the Monitoring Database. Finally, when the test window ends, the monitoring team debriefs to compare finds. The Test Security Program Manager then debriefs with the Testing Program Manager to discuss any information that may impact the TAM, protocol or procedures for that particular assessment.

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 website.

The website contains documents covering security policies and standards for different areas of operation, including Data Protection and Privacy, IT Compliance, and Threat and 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 explain the procedures used in the development and administration of the EOCEP assessments. The efforts by SCDE and DRC in developing the EOCEP assessments are in alignment with multiple best practices of the assessment industry and specifically 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 carefully follow 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 EOCEP 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 EOCEP 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 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 this 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 is timed but another isn't 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 the 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 EOCEP 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 EOCEP assessments.

3.4 Validity Based on Cognitive Processes

The Standards state that 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 in the intended test-taking population describing their performance strategies or their 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, EOCEP 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
- 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

Five indicators of construct validity for the EOCEP assessments 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 indicator has in supporting the validity of the EOCEP assessments.

3.5.1 Item-Total Correlations

An item-total correlation is the correlation between an item score 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 Fall/Winter 2023 and Spring 2024 EOCEP administrations can be found under Section 4.6. 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 EOCEP Fall/Winter and Spring Assessments

Administration	Subject	No. of Items	No. of Items $R_{it} > 0.30$	% Items $R_{it} > 0.30$
Fall/Winter	Algebra 1	50	38	76.00
	Biology 1	50	42	84.00
	English 2	55	51	92.73
	USHC	55	45	81.82
Spring	Algebra 1	50	45	90.00
	Biology 1	50	48	96.00
	English 2	55	54	98.18
	USHC	55	50	90.91

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 EOCEP assessments can be found in Section 4.6.2.2. The majority of items on the EOCEP 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 is 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 EOCEP items within each assessment. Most of the correlations are very small, suggesting local item independence generally holds for each EOCEP assessment. The assumption of local independence being met is additional evidence of the construct validity of the EOCEP 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 EOCEP assessments.

3.5.4 Item Distribution Across Content Domains

The EOCEP operational and implementation test forms were constructed according to the test specifications and the test blueprints. The 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

The primary evidence for the validity of the EOCEP 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 EOCEP 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 EOCEP 2023–2024 administrations is presented in Table 1.1.

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 EOCEP 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 sex, ethnic, or cultural unfairness. Section 2.1.4 contains discussion of how potential item unfairness is identified.

3.5.6 Dimensionality Assessment

Evidence presented in Section 4.4 assesses the dimensionality of the EOCEP 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 EOCEP test score relationship with other variables was examined to further support the validity of the intended score interpretation. This was done by examining the correlations between the EOCEP content area scores.

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 EOCEP content area scores, the correlations between EOCEP scale scores were calculated. 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 0.49 to 0.82 between the EOCEP content areas.

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 EOCEP Assessment Subject Areas, Combined Fall/Winter, Spring, and Summer Administrations

Subject	Inter-correlations			*N*		
	Algebra 1	Biology 1	US History	Biology 1	US History	English 2
Algebra 1	N/A	N/A	N/A	19,110	1,580	12,945
Biology 1	0.72	N/A	N/A	N/A	4,537	40,533
US History	0.49	0.73	N/A	N/A	N/A	3,718
English 2	0.70	0.82	0.75	N/A	N/A	N/A

Additionally, and for the purpose of establishing a college-readiness benchmark on the EOCEP tests, a cut score of ‘B’ was set by setting the percentage of students at this level to approximate the percentage of students statewide who scored at the corresponding ACT® college-readiness benchmark on the ACT® assessment. Further information regarding scaling and standard setting can be found in each subject’s standard setting document (DRC, 2016; DRC, 2017; DRC, 2022), and in Section 6 of this technical report.

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 EOCEP assessments, the purposes of the assessments must be specified. SCDE has carefully articulated the intended purposes of EOCEP as driving features of the development of the assessments and the implementation of the testing program. The specific purposes associated with the EOCEP assessments include the following:

- EOCEP assessments accurately describe 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 meet the South Carolina content area standards.
- EOCEP assessments inform state and federal accountability.
- EOCEP assessments are fair for all students, including those with disabilities or who are multilingual learners, 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 EOCEP test scores. EOCEP test 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 EOCEP assessments' 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 EOCEP 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 to develop and maintain tests of the highest quality, DRC has calculated the reliability of each EOCEP

assessment 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 SEM based on raw score 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

Classical test 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. Stated explicitly, the relationship between observed and true scores can be shown as follows:

$X = T + E$, 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

$$\rho_{XX} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}$$

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

$$\Sigma p_i \left(1 - p_i\right),$$

where p_i is the proportion correct for each item.

The reliability index used for the 2019 administration of the EOCEP assessments was the Coefficient Alpha (Cronbach, 1951):

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

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 4.4 show the test form reliability coefficients and standard errors of measurement (SEM) for the core online forms by EOCEP course for student race/ethnicity, student sex, student English proficiency status, student disability status, and whether a student used any testing accommodations. The overall reliability coefficients for the EOCEP assessments are reported in Tables 4.1 through 4.4 and ranged from 0.91 to 0.93. These results indicate acceptable reliability coefficients for the EOCEP assessments. Within each EOCEP course, the forms have very similar reliability and SEM ranges.

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 4.4, ranged from 0.88 to 0.93 for Algebra 1, from 0.84 to 0.93 for Biology 1, from 0.90 to 0.94 for English 2, and from 0.83 to 0.93 for USHC across reportable administrations, accommodation

statuses, and subgroups. The analysis of the test reliability by subgroup shows that the test reliability is acceptable 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

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

where σ_X is the standard deviation of the total test (observed measure scores) and is the reliability estimate (Coefficient Alpha) for the test. The classical 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 Algebra 1; Fall/Winter and Spring Core Online Forms

Administration	Subgroup	N Count	Cronbach's Alpha	SEM
Fall/Winter	**All Students**	**17,043**	**0.91**	**3.12**
	Female	8,069	0.90	3.10
	Male	8,791	0.91	3.13
	Asian	184	0.90	2.91
	Black or African American	5,828	0.88	3.18
	White	7,044	0.91	3.07
	Hispanic or Latino	2,519	0.90	3.13
	Two or More Races	905	0.90	3.13
	ML	1,380	0.89	3.18
	SWD	3,044	0.88	3.19
Spring	**All Students**	**50,422**	**0.93**	**3.01**
	Female	24,145	0.92	3.00
	Male	25,641	0.93	3.01
	Asian	923	0.93	2.59
	Black or African American	14,308	0.89	3.13
	White	23,303	0.92	2.90
	Hispanic or Latino	6,421	0.92	3.05
	Two or More Races	2,493	0.92	3.01
	ML	3,107	0.89	3.14
	SWD	7,421	0.91	3.12

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 Biology 1; Fall/Winter and Spring Core Online Forms

Administration	Subgroup	*N* Count	Cronbach's Alpha	*SEM*
Fall/Winter	**All Students**	**26,104**	**0.92**	**3.12**
	Female	12,885	0.91	3.12
	Male	13,046	0.92	3.10
	Asian	463	0.93	2.72
	Black or African American	8,022	0.88	3.23
	White	12,039	0.92	3.02
	Hispanic or Latino	3,560	0.90	3.18
	Two or More Races	1,397	0.91	3.14
	ML	1,682	0.84	3.23
	SWD	3,852	0.89	3.20
Spring	**All Students**	**36,511**	**0.92**	**3.08**
	Female	17,754	0.92	3.08
	Male	18,395	0.93	3.07
	Asian	651	0.93	2.72
	Black or African American	11,050	0.89	3.21
	White	16,342	0.92	2.97
	Hispanic or Latino	4,619	0.92	3.13
	Two or More Races	1,701	0.92	3.08
	ML	2,273	0.87	3.20
	SWD	5,415	0.91	3.15

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 English 2 Fall/Winter and Spring Core Online Forms

Administration	Subgroup	*N* Count	Cronbach's Alpha	*SEM*
Fall/Winter	**All Students**	**26,682**	**0.93**	**3.36**
	Female	13,145	0.93	3.29
	Male	13,339	0.93	3.40
	Asian	455	0.93	3.07
	Black or African American	8,376	0.92	3.47
	White	12,083	0.92	3.20
	Hispanic or Latino	3,756	0.93	3.48
	Two or More Races	1,323	0.92	3.31
	ML	1,835	0.90	3.61
	SWD	3,967	0.92	3.50
Spring	**All Students**	**37,736**	**0.93**	**3.29**
	Female	18,492	0.93	3.20
	Male	18,819	0.93	3.34
	Asian	692	0.93	3.02
	Black or African American	11,192	0.92	3.40
	White	16,887	0.92	3.09
	Hispanic or Latino	4,805	0.94	3.46
	Two or More Races	1,759	0.93	3.19
	ML	2,426	0.91	3.63
	SWD	5,402	0.93	3.46

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

Table 4.4. Classical Reliability Indices and SEM by Subgroup for USHC Fall/Winter and Spring Core Online Forms

Administration	Subgroup	*N* Count	Cronbach's Alpha	*SEM*
Fall/Winter	**All Students**	**22,358**	**0.91**	**3.33**
	Female	10,872	0.90	3.34
	Male	11,282	0.92	3.31
	Asian	322	0.93	3.17
	Black or African American	6,831	0.88	3.40
	White	10,358	0.91	3.27
	Hispanic or Latino	3,233	0.90	3.36
	Two or More Races	1,057	0.90	3.33
	ML	1,496	0.83	3.38
	SWD	3,244	0.90	3.35
Spring	**All Students**	**36,131**	**0.93**	**3.21**
	Female	18,345	0.93	3.23
	Male	17,453	0.93	3.19
	Asian	750	0.93	2.90
	Black or African American	10,833	0.90	3.34
	White	16,866	0.93	3.10
	Hispanic or Latino	4,340	0.92	3.29
	Two or More Races	1,523	0.92	3.21
	ML	1,949	0.86	3.39
	SWD	4,853	0.92	3.30

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 Table 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) as

$$\text{CSEM}(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}$$

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

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

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 EOCEP CSEMs and is to be expected when IRT methods are used. The CSEMs at the four cut scores that define the performance levels are presented in Table 4.5. The CSEM at the A/B cut score ranged between 4.28 and 6.96 scale score points, the CSEM at the B/C cut score ranged between 3.69 and 6.18 scale score points, the CSEM at the C/D cut score ranged between 3.37 and 5.75 scale score points, and the CSEM at the D/F cut score ranged between 3.32 and 5.65 scale score points across EOCEP assessments, administrations, and forms.

Table 4.5. CSEM at EOCEP Scale Score Cuts

Administration	Form	Subject	A/B	B/C	C/D	D/F
Fall/Winter	Online	Algebra 1	5.69	4.69	4.16	4.06
		Biology 1	6.71	5.89	5.48	5.42
		English 2	4.28	3.69	3.37	3.38
		US History	6.72	6.04	5.67	5.61
	Paper-and-pencil	Algebra 1	6.05	4.80	4.18	4.05
		Biology 1	6.75	5.91	5.52	5.48
		English 2	4.28	3.69	3.37	3.38
		US History	6.76	6.08	5.71	5.63
Spring	Online	Algebra 1	5.41	4.48	4.10	4.16
		Biology 1	6.73	5.91	5.50	5.46
		English 2	4.36	3.75	3.43	3.41
		US History	6.96	6.16	5.75	5.63
	Paper-and-pencil	Algebra 1	5.41	4.50	4.10	4.16
		Biology 1	6.75	5.91	5.52	5.48
		English 2	4.62	3.78	3.47	3.41
		US History	6.96	6.18	5.75	5.65
Summer	Online	Algebra 1	5.65	4.63	4.06	4.01
		Biology 1	6.76	5.94	5.59	5.52
		English 2	4.51	3.78	3.40	3.32
		US History	6.76	6.08	5.71	5.65
	Paper-and-pencil	Algebra 1	5.41	4.50	4.10	4.16
		Biology 1	6.75	5.91	5.52	5.48
		English 2	4.28	3.69	3.37	3.38
		US History	6.76	6.08	5.71	5.63

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

The EOCEP assessments adhere 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}	N/A	N/A	N/A	$p_{1.}$
Category 2	N/A	p_{22}	N/A	N/A	$p_{2.}$
Category 3	N/A	N/A	p_{33}	N/A	$p_{3.}$
Category 4	N/A	N/A	N/A	p_{44}	$p_{4.}$
Total	$p_{.1}$	$p_{.2}$	$p_{.3}$	$p_{.4}$	$p_{..}$

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

$$\kappa = \frac{1 - p}{p - p_c}$$

where $p = p$,

$$p_c = \sum_i p_i^2$$

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

and

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

4.2.2 Classification Accuracy Index

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

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

which transforms the original raw score random variable from $X = 0, \ldots, K$ into a new random variable $X' = 0, \ldots, n$, 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 Tables 4.7 and 4.8. The two achievement cuts represent a letter grade of D and above, and the five achievement cuts correspond to the letter grades.

Table 4.7. Decision Consistency Indices for the EOCEP Fall/Winter Administration

EOCEP	Huynh				Livingston Lewis	
	Two Achievement Levels		Five Achievement Levels		Two Achievement Levels	Five Achievement Levels
	Proportion of Agreement	Kappa	Proportion of Agreement	Kappa	Proportion of Agreement	Proportion of Agreement
Algebra 1	0.871	0.717	0.639	0.515	0.875	0.651
Biology 1	0.880	0.746	0.643	0.530	0.883	0.650
English 2	0.931	0.742	0.649	0.559	0.931	0.652
USHC	0.868	0.732	0.647	0.512	0.871	0.653

Table 4.8. Decision Consistency Indices for the EOCEP Spring Administration

EOCEP	Huynh				Livingston Lewis	
	Two Achievement Levels		Five Achievement Levels		Two Achievement Levels	Five Achievement Levels
	Proportion of Agreement	Kappa	Proportion of Agreement	Kappa	Proportion of Agreement	Proportion of Agreement
Algebra 1	0.904	0.739	0.628	0.531	0.907	0.641
Biology 1	0.888	0.754	0.641	0.534	0.890	0.648
English 2	0.940	0.739	0.653	0.558	0.940	0.656
USHC	0.889	0.766	0.660	0.550	0.892	0.666

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 differential performance on a test. We then look at test performance among varying subgroups assessed by the EOCEP content areas. 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 EOCEP 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 explained in Section 4.3.1. In addition, SCDE and DRC conduct content and fairness 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 EOCEP 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 unfairness. During the test development, SCDE and DRC content area experts avoid including items that may potentially favor one demographic group over another. Section 4.3.2 explains the steps taken to evaluate EOCEP 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 EOCEP 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 references that are not related to specific content being measured that may not be understood by members of demographic subgroups. Section 2 discusses content considerations during item development and item reviews

for items included in the EOCEP assessments. These reviews are also critical in fulfilling the guidelines established in Standards 3.1 and 3.2.

4.3.1 Minimizing Unfairness 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 unfairness are summarized below.

First, careful attention was paid to content-related validity during the item development and item selection processes. Unfairness 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 unfairness is reduced.

Second, DRC and SCDE item writers followed DRC's internal fairness 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, sex, regional background, native language, race, religion, sexual orientation, and socioeconomic status. Test developers reviewed all items included in the EOCEP 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 unfairness, including the use of content and fairness committees comprising South Carolina participants, are described in more detail in Section 2 of this report.

4.3.2 Evaluating Unfairness 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 EOCEP assessments have been thoroughly reviewed for content and fairness 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 unfair; 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 unfairness 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 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 unfair (Camilli & Shepard, 1994; Green, 1975).

In order to lessen such unfairness, 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 fairness 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):

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

where F_k is the sum of scores for the focal group at the $k\,\text{th}$ 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:

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

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:

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

For selected-response items, the $\text{MH}(\chi^2_{MH})$ 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. Items flagged for large or moderate DIF after field-testing are reviewed by content experts at DRC and the SCDE to determine whether the item the item unfairly disadvantages students in a particular demographic group. Items determined to be unfair are removed from the operational item pool. Items with large DIF but for which no apparent cause of unfairness exists are avoided during future test construction. DIF statistics were computed for the following subgroups:

- **Sex**: The focal group is females; the reference group is males.
- **Race/Ethnicity**: The focal groups are students whose race/ethnicity is reported as Black, Hispanic, Asian/Pacific Islander, American Indian, or Other; the reference group is students whose race/ethnicity is reported as White.

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 4.9 and 4.10 summarize the number of moderate and large DIF flags by EOCEP assessment for each focal group that included at least 200 students for each EOCEP assessment with sufficient sample sizes. Overall, the number of items flagged for DIF in the EOCEP assessments was small and consistent across each form of an EOCEP course. Again, any items included on the EOCEP assessments (including those items flagged for DIF) have been thoroughly reviewed for content and fairness issues by South Carolina teachers, SCDE staff, and DRC test development experts.

Table 4.9. Operational DIF Summary for EOCEP Fall/Winter Administration

Course	Reference Group	Focal Group	Total No. of Items	DIF Classification				
				A	B+	B-	C+	C-
Algebra 1	Male	Female	50	49	0	1	0	0
	White	Asian	50	50	0	0	0	0
	White	Black or African American	50	50	0	0	0	0
	White	Hispanic	50	50	0	0	0	0
	White	Two or More Races	50	50	0	0	0	0
	Non-ML	ML	50	49	0	1	0	0
	Non-SWD	SWD	50	50	0	0	0	0
Biology 1	Male	Female	50	50	0	0	0	0
	White	Asian	50	47	3	0	0	0
	White	Black or African American	50	50	0	0	0	0
	White	Hispanic	50	50	0	0	0	0
	White	Two or More Races	50	50	0	0	0	0
	Non-ML	ML	50	49	1	0	0	0
	Non-SWD	SWD	50	50	0	0	0	0
English 2	Male	Female	55	52	1	2	0	0
	White	Asian	55	50	1	3	0	1
	White	Black or African American	55	54	0	1	0	0
	White	Hispanic	55	54	0	1	0	0
	White	Two or More Races	55	55	0	0	0	0
	Non-ML	ML	55	55	0	0	0	0
	Non-SWD	SWD	55	55	0	0	0	0
USHC	Male	Female	55	52	0	3	0	0
	White	Asian	55	51	3	1	0	0
	White	Black or African American	55	55	0	0	0	0
	White	Hispanic	55	55	0	0	0	0
	White	Two or More Races	55	55	0	0	0	0
	Non-ML	ML	55	55	0	0	0	0
	Non-SWD	SWD	55	55	0	0	0	0

Table 4.10. Operational DIF Summary for EOCEP Spring Administration

Course	Reference Group	Focal Group	Total No. of Items	DIF Classification				
				A	B+	B-	C+	C-
Algebra 1	Male	Female	50	48	0	2	0	0
	White	Asian	50	50	0	0	0	0
	White	Black or African American	50	45	2	2	0	1
	White	Hispanic	50	47	1	2	0	0
	White	Two or More Races	50	50	0	0	0	0
	Non-ML	ML	50	45	2	3	0	0
	Non-SWD	SWD	50	50	0	0	0	0
Biology 1	Male	Female	50	50	0	0	0	0
	White	Asian	50	44	4	2	0	0
	White	Black or African American	50	48	0	2	0	0
	White	Hispanic	50	50	0	0	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
English 2	Male	Female	55	49	2	2	0	2
	White	Asian	55	49	2	4	0	0
	White	Black or African American	55	54	0	0	0	1
	White	Hispanic	55	54	0	1	0	0
	White	Two or More Races	55	55	0	0	0	0
	Non-ML	ML	55	55	0	0	0	0
	Non-SWD	SWD	55	55	0	0	0	0
USHC	Male	Female	55	53	0	2	0	0
	White	Asian	55	54	1	0	0	0
	White	Black or African American	55	55	0	0	0	0
	White	Hispanic	55	55	0	0	0	0
	White	Two or More Races	55	55	0	0	0	0
	Non-ML	ML	55	55	0	0	0	0
	Non-SWD	SWD	55	55	0	0	0	0

4.3.3 Evaluating Unfairness 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 4.11 through 4.14 present the numbers of students, scale score means, standard deviations (SDs), effect sizes (ESs; Cohen's d), and test form reliability statistics (Coefficient Alpha; see Section 4.1) for the various subgroups of interest.

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

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

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 4.11 through 4.15.

For Algebra 1, shown in Table 4.11, there is a small ES in mean for Hispanic or Latino students and students identified with two or more races compared to white students, with white students outperforming both subgroups of students. 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. There is a medium ES in mean for Asian students compared with white students, ML students compared to non-ML students, and students with disabilities compared to students without disabilities Asian, non-ML, and students without disabilities had higher mean scores than their referent or focal group in the analysis.

For Biology 1, shown in Table 4.12, there is a small to medium ES in mean for Asian students compared to white students, with Asian students outperforming white students, and a small ES for students with two or more races compared to white students with white students outperforming students with two or more races. There is a medium ES in mean for Hispanic or Latino students compared to White students, with white students outperforming Hispanic or Latino students. 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, and for ML students compared to non-ML students, with non-ML students outperforming ML students. There is a medium ES for students with disabilities compared to students without disabilities, with students without disabilities outperforming students with disabilities.

For English 2, shown in Table 4.13, there is a small ES in mean for Asian students compared to white students with Asian students outperforming white students, for students with two or more races compared to white students, with white students outperforming students with two or more races, and for female students compared to male students, with female students outperforming male students. There is a medium to

large ES in mean for Hispanic or Latino students compared to white students, with white students outperforming Hispanic or Latino students. There is a large ES in Black or African American students compared to white students, with white students outperforming Black or African American students, there is a large ES for ML students compared to non-ML students, with non-ML students outperforming ML students. There is also a large ES for students with disabilities compared to students without disabilities, with students without disabilities outperforming students with disabilities.

For USHC, shown in Table 4.14, there is a small ES in mean for Asian students compared to white students, with Asian students outperforming white students, and for students with two or more races compared to white students with white students outperforming students with two or more races. There is a medium ES in mean or Hispanic or Latino students compared to white students, with white students outperforming Hispanic or Latino students, and 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 Black or African American students compared to white students, with white students outperforming Black or African American students, and for ML students compared to non-ML students, with non-ML students outperforming ML students.

Table 4.11. Impact Analysis for Algebra 1; EOCEP Combined Fall/Winter, Spring, and Summer Administrations

Category	Subgroup	*N*	Mean	Std. Dev.	Effect Size
Ethnicity	White	30,403	75.32	15.42	N/A
	Asian	1,110	83.95	14.49	0.56
	Black or African American	20,170	64.74	13.19	0.73
	Hispanic or Latino	8,950	68.76	14.79	0.43
	Two or More Races	3,403	71.30	14.91	0.26
Sex	Male	34,531	69.86	16.05	N/A
	Female	32,305	71.24	14.99	0.09
ML	Non-ML	63,223	70.87	15.62	N/A
	ML	4,496	63.21	13.15	0.50
SWD	Without Disabilities	57,171	71.75	15.46	N/A
	With Disabilities	10,548	62.84	14.06	0.58

Table 4.12. Impact Analysis for Biology 1; EOCEP Combined Fall/Winter, Spring, and Summer Administrations

Category	Subgroup	*N*	Mean	Std. Dev.	Effect Size
Ethnicity	White	28,439	75.88	17.96	N/A
	Asian	1,117	83.51	17.42	0.43
	Black or African American	19,105	60.55	16.02	0.89
	Hispanic or Latino	8,189	65.12	18.04	0.60
	Two or More Races	3,101	70.10	17.86	0.32
Sex	Male	31,516	68.21	19.28	N/A
	Female	30,713	69.83	18.37	0.09
ML	Non-ML	58,822	69.79	18.79	N/A
	ML	3,962	55.76	14.58	0.76
SWD	Without Disabilities	53,442	70.56	18.63	N/A
	With Disabilities	9,342	59.44	17.38	0.60

Table 4.13. Impact Analysis for English 2; EOCEP Combined Fall/Winter, Spring, and Summer Administrations

Category	Subgroup	*N*	Mean	Std. Dev.	Effect Size
Ethnicity	White	29,143	83.07	13.33	N/A
	Asian	1,151	87.25	13.58	0.31
	Black or African American	19,775	72.35	13.37	0.80
	Hispanic or Latino	8,625	73.57	15.76	0.68
	Two or More Races	3,104	79.54	13.75	0.26
Sex	Male	32,488	75.73	14.98	N/A
	Female	31,896	80.17	14.27	0.30
ML	Non-ML	60,758	78.86	14.39	N/A
	ML	4,306	63.40	13.26	1.08
SWD	Without Disabilities	55,533	79.38	14.32	N/A
	With Disabilities	9,531	68.88	14.57	0.73

Table 4.14. Impact Analysis for USHC; EOCEP Combined Fall/Winter, Spring, and Summer Administrations

Category	Subgroup	*N*	Mean	Std. Dev.	Effect Size
Ethnicity	White	27,272	74.27	19.33	N/A
	Asian	1,076	79.76	19.55	0.28
	Black or African American	17,687	58.46	17.62	0.85
	Hispanic or Latino	7,580	63.39	19.35	0.56
	Two or More Races	2,587	69.18	19.47	0.26
Sex	Male	28,811	68.15	20.68	N/A
	Female	29,308	66.82	19.74	0.07
ML	Non-ML	55,251	68.32	20.13	N/A
	ML	3,448	52.53	15.29	0.79
SWD	Without Disabilities	50,533	68.78	19.98	N/A
	With Disabilities	8,166	58.80	19.64	0.50

4.4 Test Dimensionality

As another measure of the tests' internal structure, DRC examined the unidimensionality of each EOCEP assessment. One of the underlying assumptions of the IRT models used to scale EOCEP is that the tests being calibrated are unidimensional. That is, items composing each EOCEP 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 4.6.2 (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 EOCEP. 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 Table 4.15. For the EOCEP assessments, 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 16.6% of the test variance. To further investigate the unidimensionality of the EOCEP 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 Table 4.15 show that the eigenvalue of the first factor is at least five times as large as the eigenvalue of the second factor for each EOCEP. Additionally, for each EOCEP assessment, the percentage of variance explained for the second factor is substantially less than the first factor. This 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 EOCEP assessments 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 ability related to that dimension. Construct-irrelevant variance, such as factual knowledge that is irrelevant to a subject, does not appear to create significant nuisance factors.

Table 4.15. Principal Component Analysis, EOCEP Fall Administration

EOCEP	Administration	Components/Factors	Eigenvalue	% Variance Explained	Cumulative % Variance Explained
Algebra 1	Fall/Winter	First Component	9.44	18.9	18.9
		Second Component	1.87	3.7	22.6
		Ratio (First/Second)	5.06	N/A	N/A
	Spring	First Component	11.47	22.9	22.9
		Second Component	2.14	4.3	27.2
		Ratio (First/Second)	5.37	N/A	N/A
Biology 1	Fall/Winter	First Component	10.41	20.8	20.8
		Second Component	1.26	2.5	23.4
		Ratio (First/Second)	8.26	N/A	N/A
	Spring	First Component	10.83	21.7	21.7
		Second Component	1.31	2.6	24.3
		Ratio (First/Second)	8.29	N/A	N/A
English 2	Fall/Winter	First Component	12.87	23.4	23.4
		Second Component	1.33	2.4	25.8
		Ratio (First/Second)	9.65	N/A	N/A
	Spring	First Component	13.37	24.3	24.3
		Second Component	1.70	3.1	27.4
		Ratio (First/Second)	7.86	N/A	N/A
USHC	Fall/Winter	First Component	9.81	17.8	17.8
		Second Component	1.33	2.4	20.3
		Ratio (First/Second)	7.35	N/A	N/A
	Spring	First Component	11.99	21.8	21.8
		Second Component	1.53	2.8	24.6
		Ratio (First/Second)	7.83	N/A	N/A

4.5 Item Scoring

This section will describe the scoring process used for the EOCEP assessments. In particular, this section focuses on the PAS (Performance Assessment Services) process of handscoring TDA items for English 2 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 EOCEP TDA items.

This section describes how the EOCEP 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 EOCEP assessment handscoring process. Section 4.5.1.2 describes how the scorers were monitored throughout the EOCEP 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 EOCEP scoring team was staffed with many 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 in order to staff the scoring of the EOCEP 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 high school English. 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 EOCEP materials. For the typical summer administration, since the participating number of students is low, student TDA responses may be scored by the Project Manager and scoring director.

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 English 2 TDA items took place Columbia, South Carolina, using a committee comprised of South Carolina teachers and with SCDE members present. The rangefinding session was facilitated by a PAS Project Manager and a scoring director. 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 were

represented in the anchor papers. The anchor papers were used in training and qualifying 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 EOCEP English 2 assessments, 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 rescored.

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 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, adjusting 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 the EOCEP English 2 assessments are detailed in Table 4.17.

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 Report 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. 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 4.17. To aid in the interpretation of the kappa statistic, Table 4.16 provides the suggested cutoffs (Altman, 1991; Landis & Koch, 1977).

Table 4.16. 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 4.17, raters demonstrated a high % exact and adjacent agreement for the writing prompt component scoring in English 2. The exact agreement ranged from 78.7% to 83.0% for components scored using a 1–4-point rubric. The quadratic weighted kappa values were at least 0.78 for components scored using a 1–4-point rubric, indicating good to very good inter-rater agreement for these components.

Table 4.17. EOCEP TDA Reader Agreement, English 2 Fall/Winter and Spring

Administration	Reader Exact %	Reader Adjacent %	Reader Nonadjacent %	Quadratic Weighted Kappa
Fall/Winter	82.97	16.89	0.13	0.83
Spring	78.70	20.66	0.64	0.78

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 steps:

- 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 included 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 EOCEP 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 EOCEP 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 EOCEP tests are presented.

This section demonstrates adherence in the EOCEP assessments 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 explained 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 EOCEP. This section also presents item-level statistics for each EOCEP assessment with a sufficient sample size.

4.6.1.1 Test-Level Statistics

Table 4.18 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 EOCEP assessment. Due to sample size limitations, item-level information for Summer 2024 forms is not reported.

The mean *p*-value is the average of all item p-values for a specific EOCEP. The mean item-total correlation (R_{it}) is the average of all item biserial correlations for a specific assessment. The p-value and item-total correlation are explained in the next section.

Table 4.18. EOCEP Fall/Winter and Spring Means and Standard Deviations for Raw Scores, *p*-values, and Item-Total Correlation

EOCEP	Administration	No. of Items	Mean RS, *SD*	Mean *p*-value, *SD*	*p*-value Range	R_{it}, *SD*
Algebra 1	**Fall/Winter**	50	26.68, 10.16	0.54, 0.13	0.27 - 0.79	0.38, 0.10
	Spring	50	28.86, 11.03	0.58, 0.14	0.31 - 0.82	0.43, 0.11
Biology 1	**Fall/Winter**	50	27.46, 10.89	0.55, 0.10	0.37 - 0.79	0.41, 0.09
	Spring	50	28.18, 11.14	0.56, 0.10	0.36 - 0.81	0.43, 0.08
English 2	**Fall/Winter**	55	42.16, 14.20	0.64, 0.11	0.44 - 0.88	0.44, 0.10
	Spring	55	44.00, 14.10	0.68, 0.13	0.41 - 0.87	0.45, 0.08
USHC	**Fall/Winter**	55	28.50, 11.11	0.52, 0.11	0.30 - 0.76	0.38, 0.08
	Spring	55	31.56, 12.14	0.57, 0.11	0.34 - 0.77	0.42, 0.09

Note. Both English 2 Fall/Winter and Spring Administrations included 55 items totaling 70 points. English 2

4.6.1.2 Item-Level Statistics

Individual operational item level statistics on the test forms are provided in this section. Tables C.1 through C.8 in Appendix C present the item statistics for the EOCEP assessment. Due to sample size limitations, item-level information for Summer 2023 forms is not reported. A description of the item-level statistics provided and a summary of the results by EOCEP assessment is described below.

p-value: The p-value is a measure of item difficulty. For a multiple-choice item, the p-value is calculated by taking the number of students who correctly responded to an item and dividing by the total number of students who attempted the item. The value is reported as a proportion. 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 grade level. Table 4.18 also provides the range of p-values for each grade and content area.

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 EOCEP 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 the speediness 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 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, shows that all EOCEP items had omit rates well below 5%, suggesting that the majority of students were not rushed.

The average p-value for the Algebra 1 assessments was 0.54 for Fall/Winter and 0.58 for Spring, and p-values ranged from 0.27 to 0.82. The average item-total correlation was 0.38 for Fall/Winter and 0.43 for Spring.

The average p-value for the Biology 1 assessments was 0.55 for Fall/Winter and 0.56 for Spring. P-values ranged from 0.36 to 0.81 across both administrations. The average item-total was 0.41 for Fall/Winter and 0.43 for Spring.

The average p-value for the English 2 assessment was 0.64 for Fall/Winter and 0.68 for Spring, and p-values across both administrations ranged from 0.41 to 0.88. The average item-total correlation was 0.44 for Fall/Winter and 0.45 for Spring.

The average p-value for the USHC assessments was 0.52 for Fall/Winter and 0.57 for Spring, and p-values across both administrations ranged from 0.30 to 0.77. The average item-total correlation was 0.38 for Fall/Winter and 0.42 for Spring.

4.6.2 Item Calibration Using Rasch Measurement Models

Scale scores for the EOCEP 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 EOCEP 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:

$$P_{ni}(X=1) = \frac{\exp(\theta_n - D_{ij})}{1 + \exp(\theta_n - D_{ij})}$$

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.

Parameter estimation for items on the EOCEP assessments using the Rasch model was implemented using WINSTEPS (Linacre, 2020). 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. The multipoint TDA items used on the English 2 assessments also had thresholds calibrated by WINSTEPS and were put on the same scale as the dichotomous items. Operational item calibration is based on the year the item was first field-tested on an operational form or appeared on the current standard setting form.

4.6.2.1 Analysis Sample

This section describes the analysis sample 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)*

The stability (invariance) of the parameters is assessed using multiple methods (addressed in the following sections) to examine Rasch item fit, local independence, and the evaluation of the pre-equated tables. WINSTEPS 4.2.0 (Linacre, 2018) is used to estimate the change in item difficulty (displacement) between the field test and operational administrations. The assessment of these samples was based on census data or very close to census data for the EOCEP Fall/Winter or Spring assessments. They are thus representative of the South Carolina student population in regard to sex, ethnicity, and accommodation status distribution. Summer data is not included due to the low sample sizes.

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 EOCEP assessments. In this situation, it is recommended that the Zstd values be ignored if the MnSq values are acceptable (Linacre, 2019).

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 think 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 EOCEP 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 EOCEP 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. EOCEP Infit and Outfit Mean Square Statistics Fall/Winter Administration

Statistic	Algebra 1		Biology 1		English 2		USHC	
	In	Out	In	In	In	Out	In	Out
N	50	50	50	50	55	55	55	55
Mean	1.01	1.02	1.01	1.00	1.03	1.01	1.01	1.01
SD	0.12	0.17	0.12	0.17	0.13	0.20	0.10	0.15
Minimum	0.79	0.73	0.80	0.70	0.74	0.61	0.82	0.68
P_{10}	0.85	0.80	0.86	0.78	0.84	0.74	0.87	0.82
P_{25}	0.91	0.87	0.92	0.89	0.93	0.85	0.93	0.91
P_{50}	1.02	1.04	0.99	0.97	1.07	1.02	1.01	1.01
P_{75}	1.08	1.15	1.07	1.09	1.11	1.13	1.06	1.10
P_{90}	1.16	1.23	1.19	1.24	1.17	1.24	1.13	1.18
Maximum	1.29	1.36	1.27	1.38	1.31	1.47	1.23	1.33

Table 4.20. EOCEP Infit and Outfit Mean Square Statistics Spring Administration

Statistic	Algebra 1		Biology 1		English 2		USHC	
	In	Out	In	Out	In	Out	In	Out
N	50	50	50	50	55	55	55	55
Mean	1.00	1.00	1.00	1.00	1.01	1.00	1.00	1.01
SD	0.15	0.24	0.11	0.15	0.14	0.22	0.12	0.20
Minimum	0.70	0.58	0.79	0.66	0.74	0.52	0.77	0.69
P_{10}	0.82	0.74	0.87	0.81	0.80	0.62	0.84	0.75
P_{25}	0.90	0.85	0.92	0.88	0.92	0.86	0.92	0.86
P_{50}	0.98	0.95	1.00	1.01	1.04	1.03	1.01	1.00
P_{75}	1.09	1.13	1.09	1.13	1.12	1.18	1.08	1.12
P_{90}	1.20	1.32	1.13	1.21	1.20	1.28	1.13	1.26
Maximum	1.47	1.80	1.17	1.24	1.30	1.39	1.26	1.73

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 $X_1, X_2, \ldots X_n$ has LI with respect to the latent variable ϑ if, for all $x = (x_1, x_2, \ldots x_n)$ and ϑ,

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

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:

$$P\left(X_i = x_i, X_j = x_j \mid \theta\right) = P\left(X_i = x_i \mid \theta\right) P\left(X_j = x_j \mid \theta\right)$$

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 EOCEP 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 -0.02 for the EOCEP 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 ($P_{10}, P_{25}, P_{50}, P_{75}, \text{ and } P_{90}$)—for all the residual correlations for each EOCEP course and administration. The total number of item pairs and the number of pairs with residual correlations greater than 0.20 are also reported in this table. The mean residual correlations are very close to 0.00 for all EOCEP assessments and both administrations. Most of the correlations are very small suggesting local item independence generally holds for the EOCEP assessments.

Table 4.21. Summary of Residual Correlations for EOCEP Fall/Winter Administration

Statistic	Algebra 1	Biology 1	English 2	USHC
N pairs of items	1225	1225	1485	1485
Mean	-0.02	-0.02	-0.02	-0.02
SD	0.04	0.02	0.03	0.02
Minimum	-0.15	-0.09	-0.11	-0.10
P_{10}	-0.07	-0.05	-0.05	-0.04
P_{25}	-0.05	-0.04	-0.03	-0.03
P_{50}	-0.02	-0.02	-0.02	-0.02
P_{75}	0.00	-0.01	0.00	-0.01
P_{90}	0.03	0.01	0.02	0.01
Maximum	0.26	0.08	0.16	0.07
$> \lvert 0.20 \rvert$	2	0	0	0

Table 4.22. Summary of Residual Correlations for EOCEP Spring Administration

Statistic	Algebra 1	Biology 1	English 2	USHC
N pairs of items	1225	1225	1485	1485
Mean	-0.02	-0.02	-0.02	-0.02
SD	0.05	0.02	0.03	0.02
Minimum	-0.14	-0.08	-0.12	-0.10
P_{10}	-0.07	-0.04	-0.05	-0.04
P_{25}	-0.05	-0.03	-0.03	-0.03
P_{50}	-0.02	-0.02	-0.02	-0.02
P_{75}	0.00	-0.01	0.00	-0.01
P_{90}	0.03	0.00	0.02	0.01
Maximum	0.35	0.12	0.16	0.11
$> \lvert 0.20 \rvert$	5	0	0	0

4.6.2.4 Post-equating Checks

EOCEP assessments are pre-equated assessments. All items on the original standard setting item calibration are fixed at that calibration value for the creation of the pre-equated raw-score-to-scale-score conversion tables used with each new form developed. New items developed for the EOCEP assessments are co-calibrated with the pre-equated forms, and the item difficulties are placed on the standard setting difficulty origin by anchoring the operational items to their pre-equated values and estimating the difficulty of the field test items on the original origin. This process is repeated each time new items are field-tested, and the field test items maintain that difficulty in the pre-equating of future operational tests. Any item difficulty drift is monitored by performing an anchored calibration of the operational items using their original bank values. WINSTEPS calculated the estimated displacement from the bank value. These results are monitored and provided to SCDE to monitor the appropriateness of the pre-equated raw-score-to-scale-score conversion tables. When a pre-equated raw-score-to-scale-score table is deemed to be inappropriate, a post equating solution is recommended.

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 EOCEP 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 .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 EOCEP 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:

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_{diff}).
8. Calculate the interquartile range of the differences (r_{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_{e}) and current (b_{c}) item difficulties minus the median of the differences, that quantity divided by the quantity the interquartile range multiplied by 0.74:

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

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 4.23 provides equating results for the robust Z method. This table summarizes the following information for each content area: 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. The established guidelines indicate that the anchored and unanchored difficulties had high correlation (>0.95) and ratio of the standard deviations between the anchored and unanchored difficulty parameters was within the 0.90 to 1.10 range.

For both administrations, the anchor sets remained intact, because they generally met the criteria established for the equating guidelines. For all Robust Z evaluations, 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 are similar, indicating that no scoring table adjustments were needed.

Table 4.23. EOCEP Post-Equating Summary

Administration	Course	No. OP Items	No. Iterations	No. OP Items Fixed	% Items Recommended to be Freely Calibrated	*r*	SD Ratio	UW Link Constant
Fall/Winter	Algebra 1	50	1	49	2.00%	0.95	0.98	0.04
	Biology 1	50	6	44	12.00%	0.90	1.00	0.18
	English 2	55	3	52	5.45%	0.93	1.03	0.37
	USHC	55	1	54	1.82%	0.95	0.97	0.04
Spring	Algebra 1	50	0	50	0.00%	0.96	0.96	0.26
	Biology 1	50	4	46	8.00%	0.95	0.97	0.14
	English 2	55	0	55	0.00%	0.96	0.99	0.36
	USHC	55	0	55	0.00%	0.96	0.98	-0.04

Note. r = correlation. SD = Standard Deviation. UW = Unweighted.

4.6.3 Analyses by Reporting Categories

Three sets of analyses were conducted at the reporting category level for the EOCEP assessments in an additional attempt to assess their internal structure. 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.

To assess the internal structure of the EOCEP 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 reporting 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 4.24 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.

The uncorrected PPM in Tables 4.24 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 each EOCEP assessment.

For Algebra 1, at the reporting category level, the correlations ranged from 0.62 to 0.82. For Biology 1, the correlations are not reported because reporting category scores were not reported for the 2023–2024 assessments. (The table remains in this report as a placeholder.) For English 2, the correlations ranged from 0.72 and 0.84. For USHC, the correlations ranged from 0.63 and 0.75. 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. It is expected 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:

$$CAPPM = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

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 4.24 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 4.24 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.

For Algebra 1, at the reporting category level, the correlations ranged from 0.87 to 0.99. For Biology 1, the correlations are not reported because reporting category scores were not reported for the 2023–2024 assessments. (The table remains in this report as a placeholder.) For English 2, the correlations ranged between 0.90 and 1.00. For USHC, the correlations ranged between 0.94 and 1.00. 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.

Table 4.24. Correlation Coefficients among Reporting Categories, Algebra 1

Administration	No.	Reporting Category	No. of Items	Uncorrected Correlation Coefficient		Corrected Correlation Coefficient	
				1	2	2	3
Fall/Winter	1	Algebra	21	N/A	N/A	0.97	0.93
	2	Functions	21	0.79	N/A	N/A	0.99
	3	Number and Quantity; Interpreting Data	8	0.64	0.66	N/A	N/A
Spring	1	Algebra	23	N/A	N/A	0.96	0.87
	2	Functions	18	0.82	N/A	N/A	0.98
	3	Number and Quantity; Interpreting Data	9	0.62	0.67	N/A	N/A

Table 4.25. Correlation Coefficients among Reporting Categories, Biology 1

Adm.	No.	Reporting Category	Items	Uncorrected Correlation Coefficient					Corrected Correlation Coefficient				
				1	2	3	4	5	2	3	4	5	6
Fall/ Winter	1	Structure and Processes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	2	Ecosystems	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	3	DNA and Heredity	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	4	Biological Evolution	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Spring	1	Structure and Processes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	2	Ecosystems	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	3	DNA and Heredity	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	4	Biological Evolution	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note. These correlations are omitted because the 2023–2024 Biology 1 assessments did not report at the reporting category level.

Table 4.26. Correlation Coefficients among Reporting Categories, English 2

Administration	No.	Reporting Category	No. of Items	Uncorrected Correlation Coefficient			Corrected Correlation Coefficient		
				1	2	3	1	2	3
Fall/Winter	1	Informational Text	20	N/A	N/A	N/A	N/A	0.98	0.99
	2	Literary Text	19	0.84	N/A	N/A	N/A	N/A	0.97
	3	Writing and Communication	11*	0.73	0.72	N/A	N/A	N/A	N/A
Spring	1	Informational Text	19	N/A	N/A	N/A	N/A	1.00	1.00
	2	Literary Text	19	0.81	N/A	N/A	N/A	N/A	0.90
	3	Writing and Communication	13*	0.75	0.76	N/A	N/A	N/A	N/A

Note. Fall/Winter Number 3: Writing and Communication Reporting Category includes 11 items with a total of 26 points. Spring Number 3: Writing and Communication Reporting Category includes 13 items with a total of 28 points.

Table 4.27. Correlation Coefficients among Reporting Categories, US History

Adm.	No.	Reporting Category	Items	Uncorrected Correlation Coefficient				Corrected Correlation Coefficient			
				1	2	3	4	2	3	4	5
Fall/ Winter	1	Foundations of American Republicanism	11	N/A	N/A	N/A	N/A	1.00	0.99	0.98	0.94
	2	Expansion and Union	11	0.68	N/A	N/A	N/A	N/A	1.00	0.99	0.97
	3	Capitalism and Reform	11	0.66	0.71	N/A	N/A	N/A	N/A	1.00	0.96
	4	Modernism and Interventionism	11	0.65	0.69	0.69	N/A	N/A	N/A	N/A	1.00
	5	Legacy of the Cold War	11	0.64	0.64	0.63	0.65	N/A	N/A	N/A	N/A
Spring	1	Foundations of American Republicanism	11	N/A	N/A	N/A	N/A	0.98	0.97	0.98	0.94
	2	Expansion and Union	11	0.70	N/A	N/A	N/A	N/A	1.00	0.99	0.97
	3	Capitalism and Reform	11	0.69	0.71	N/A	N/A	N/A	N/A	1.00	0.98
	4	Modernism and Interventionism	11	0.73	0.74	0.74	N/A	N/A	N/A	N/A	0.99
	5	Legacy of the Cold War	11	0.69	0.70	0.71	0.75	N/A	N/A	N/A	N/A

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 4.31 for the EOCEP 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. As with the correlation coefficients, note that the Biology 1 reporting category summary statistics are omitted here because only overall scores were reported. Table 4.29 remains in the report as a placeholder.

Table 4.28. Reliability and Standard Error of Measurement of Reporting Categories EOCEP, Algebra 1 Reporting Category Level

Administration	No.	Reporting Category	No. of Items & Points	*N* Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	*SEM*
Fall/Winter	1	Algebra	21	17,043	11.29	4.88	0.83	2.01
	2	Functions	21	17,043	11.30	4.36	0.79	2.01
	3	Number and Quantity; Interpreting Data	8	17,043	4.09	1.93	0.57	1.26
Spring	1	Algebra	23	50,422	14.47	5.64	0.88	1.96
	2	Functions	18	50,297	9.85	4.33	0.83	1.81
	3	Number and Quantity; Interpreting Data	9	50,422	4.56	2.07	0.58	1.35

Table 4.29. Reliability and Standard Error of Measurement of Reporting Categories EOCEP, Biology 1 Reporting Category Level

Administration	No.	Reporting Category	No. of Items & Points	*N* Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	*SEM*
Fall/Winter	1	Structure and Processes	N/A	N/A	N/A	N/A	N/A	N/A
	2	Ecosystems	N/A	N/A	N/A	N/A	N/A	N/A
	3	DNA and Heredity	N/A	N/A	N/A	N/A	N/A	N/A
	4	Biological Evolution	N/A	N/A	N/A	N/A	N/A	N/A
Spring	1	Structure and Processes	N/A	N/A	N/A	N/A	N/A	N/A
	2	Ecosystems	N/A	N/A	N/A	N/A	N/A	N/A
	3	DNA and Heredity	N/A	N/A	N/A	N/A	N/A	N/A
	4	Biological Evolution	N/A	N/A	N/A	N/A	N/A	N/A

Note. Reporting category summary statistics are omitted because the 2023–2024 Biology 1 assessments did not report at the reporting category level.

Table 4.30. Reliability and Standard Error of Measurement of Reporting Categories EOCEP, English 2 Reporting Category Level

Administration	No.	Reporting Category	No. of Items & Points	*N* Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	*SEM*
Fall/Winter	1	Informational Text	20	26,710	12.45	4.86	0.85	1.87
	2	Literary Text	19	26,710	12.43	4.78	0.87	1.75
	3	Writing and Communication	11 (26 points)	26,710	10.63	3.22	0.63	1.95
Spring	1	Informational Text	19	37,736	12.36	4.39	0.83	1.79
	2	Literary Text	19	37,736	13.45	4.39	0.86	1.64
	3	Writing and Communication	13 (28 points)	37,736	11.72	4.07	0.75	2.05

Table 4.31. Reliability and Standard Error of Measurement of Reporting Categories EOCEP, USHC Reporting Category Level

Administration	No.	Reporting Category	No. of Items & Points	*N* Count	RS Mean	RS Std. Dev.	Cronbach's Alpha	*SEM*
Fall/Winter	1	Foundations of American Republicanism	11	22,358	6.13	2.48	0.65	1.48
	2	Expansion and Union	11	22,358	5.68	2.71	0.71	1.46
	3	Capitalism and Reform	11	22,358	6.08	2.64	0.68	1.48
	4	Modernism and Interventionism	11	22,358	5.43	2.68	0.69	1.49
	5	Legacy of the Cold War	11	22,358	5.16	2.49	0.63	1.52
Spring	1	Foundations of American Republicanism	11	36,131	6.97	2.64	0.72	1.40
	2	Expansion and Union	11	36,131	6.19	2.68	0.70	1.46
	3	Capitalism and Reform	11	36,131	6.55	2.65	0.70	1.45
	4	Modernism and Interventionism	11	36,131	6.02	3.00	0.78	1.41
	5	Legacy of the Cold War	11	36,131	5.82	2.82	0.73	1.46

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 EOCEP assessment 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 EOCEP assessments are described and evaluated.

4.7.1 Description of EOCEP Raw Scores

The basic summary statistic on all EOCEP assessments is the raw score. A raw score is calculated for each examinee. The raw score is the number of score points earned on the assessment. The English 2 assessments contain one 4-point TDA item per form that is scored by two readers. The score is then doubled before calculating the total raw score. By itself, the raw score has limited utility; it can only be interpreted in reference to the total number of items on a content-area assessment, and raw scores should not be compared across tests or administrations.

4.7.2 Creating the EOCEP Scale

The structure of each EOCEP assessment scale score metric was determined by SCDE staff. Scale scores are intended to make scores more meaningful by defining a scale of measurement that is not tied to a particular test form. The range of scale scores was set to have a minimum score of 0 and maximum score of 100. Additionally, the scale is constructed so that each standard letter grade of A, B, C, D, and F corresponds to scale score values of 90, 80, 70, 60, and 50.

4.7.3 Description of EOCEP Scale Scores

When new test forms are developed, the new set of items may require slightly different levels of content-area skill to answer correctly. This depends on the difficulty of the specific questions used on each form. To be fair to students and to permit valid comparison of test scores across administrations, the skills represented by each scale score point must remain consistent from year to year.

As noted previously, scale scores adjust for slight shifts in underlying difficulty levels at each score point and provide valid points of comparison across all test administrations within a particular content area. With scale scores, schools can reasonably compare the demonstrated knowledge and performance of groups of students across years.

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. To assist in maintaining equivalent passing standards across different administrations, SCDE, in collaboration with DRC,

constructs all tests to be of similar difficulty. This similarity is maintained from administration to administration at the total test level and, as much as possible, at the reporting standard level (see Section 2.1.8).

New raw-score-to-scale-score conversion tables, along with the standard error of measurement for each raw score, were developed for the pre-equated SC EOCEP forms. Scale scores were calculated for every raw score for each EOCEP assessment using the formulas provided in Table 4.32. Note that ϑ_{RS} is the value of theta corresponding to that raw score.

Table 4.32. Table of Scale Score Conversion Tables for EOCEP Assessments

EOCEP	Conversion Equation
Algebra 1	$\mathrm{SS} = \ \mathrm{floor}\ [63.12636 + \vartheta_{\mathrm{RS}} * 13.58696]$
Biology 1	$\mathrm{SS} = \ \mathrm{floor}\ [61.16009 + \vartheta_{\mathrm{RS}} * 17.92970]$
English 2	$\mathrm{SS} = \ \mathrm{floor}\ [61.48610 + \vartheta_{\mathrm{RS}} * 12.66900]$
USHC	$\mathrm{SS} = \ \mathrm{floor}\ [63.18734 + \vartheta_{\mathrm{RS}} * 20.12207]$

Table 4.33 contains the cut scores (both in the scale score and theta metrics) and the LOSS and HOSS for each EOCEP assessment. Cut score values were obtained from the EOCEP standard settings (see Chapter 6 for more details on standard setting activities).

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 and HOSS were first converted to a theta estimate.

Table 4.33. EOCEP Scale Score Cuts, Rasch Ability Cuts, and LOSS and HOSS

Achievement Level	A/B		B/C		C/D		D/F		LOSS	HOSS
EOCEP	Scale Score Cut	Rasch Ability Cut	Scale Score Cut	Rasch Ability Cut	Scale Score Cut	Rasch Ability Cut	Scale Score Cut	Rasch Ability Cut		
Algebra 1	90	1.9779	80	1.2419	70	0.5059	60	-0.2301	0	100
Biology 1	90	1.6805	80	1.0508	70	0.4930	60	-0.0647	0	100
English 2	90	2.2507	80	1.4614	70	0.6975	60	-0.1173	0	100
USHC	90	1.3325	80	0.8355	70	0.3386	60	-0.1584	0	100

4.7.4 Cautions for Score Use

As with any assessment, student scores at the minimum or maximum ends of the score range will have large standard errors of measurement and should be viewed cautiously. For instance, the maximum raw score for the EOCEP Algebra 1 assessment is equal to the number of points on the form. If a student achieves this score, it cannot be determined whether the student would have achieved a higher scale score if that score was possible. All that is known is that the student's scale score, as revealed by this test, is at least the maximum scale score for the perfect raw score. Because of this ceiling effect, extreme scale scores may vary from one administration to the next, depending on the distribution of the item difficulties on the form. Even if the number of items tested remains the same, making comparisons of students that score at the extreme ends of the score distribution is difficult. To minimize confusion and the potential for misinterpretation, the maximum scale score possible on the EOCEP has been fixed at 100 to reduce the likelihood that the maximum score will change across forms.

4.7.5 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 conversion tables were produced for each operational form of the EOCEP assessments administered during the 2023–2024 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. Scoring tables can be found in Appendix B.

4.7.6 EOCEP Scale Evaluation

The EOCEP assessments each have an established scale. The scale score distribution results are discussed in more detail in Section 6.3. However, to summarize the scale score distribution and behavior, the scale scores increase as the percentile rank increases (as expected), showing increasing student ability along the scale for all grades in all content areas.

The Test Characteristic Curves (TCC) and Test Information Function (TIF) curves for the EOCEP assessments are shown in figures 4.1 through 4.4. The TCCs are S shaped, indicating the increasing probability of a higher test score as a student's ability increases. Additionally, the Fall/Winter, Spring, and Summer administration TCCs are similar for each EOCEP.

The CSEM curves for the EOCEP assessment forms are a 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 over the score scale was found to be reasonable for the EOCEP assessments and administrations.

Figure 4.1. Test Characteristic Curves for EOCEP Algebra 1 Administrations

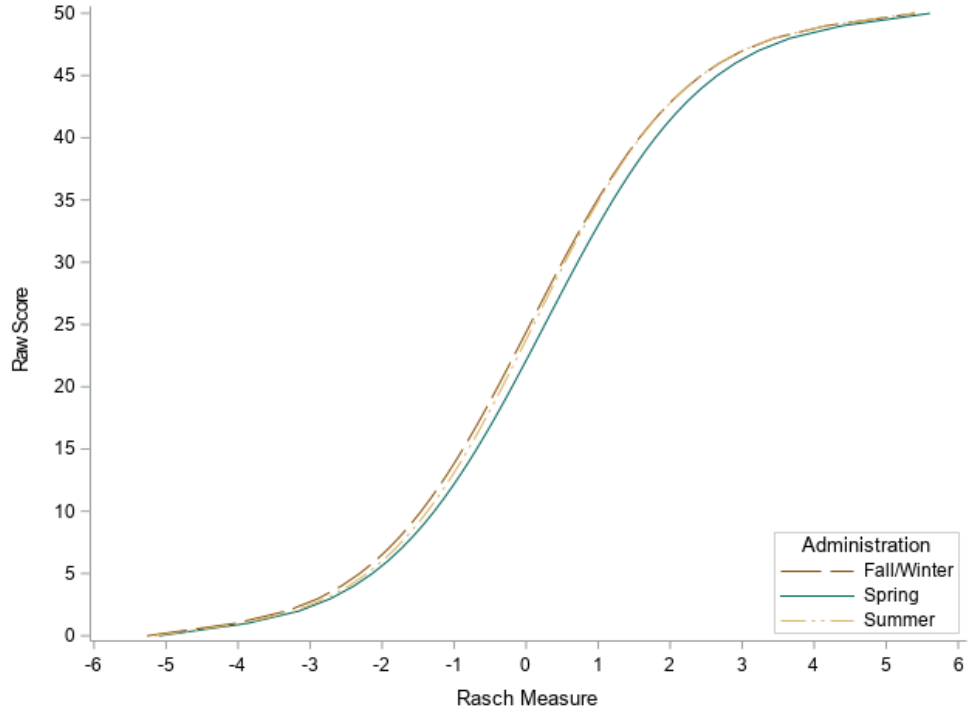


Figure 4.2. CSEM Curves for EOCEP Algebra 1 Administrations

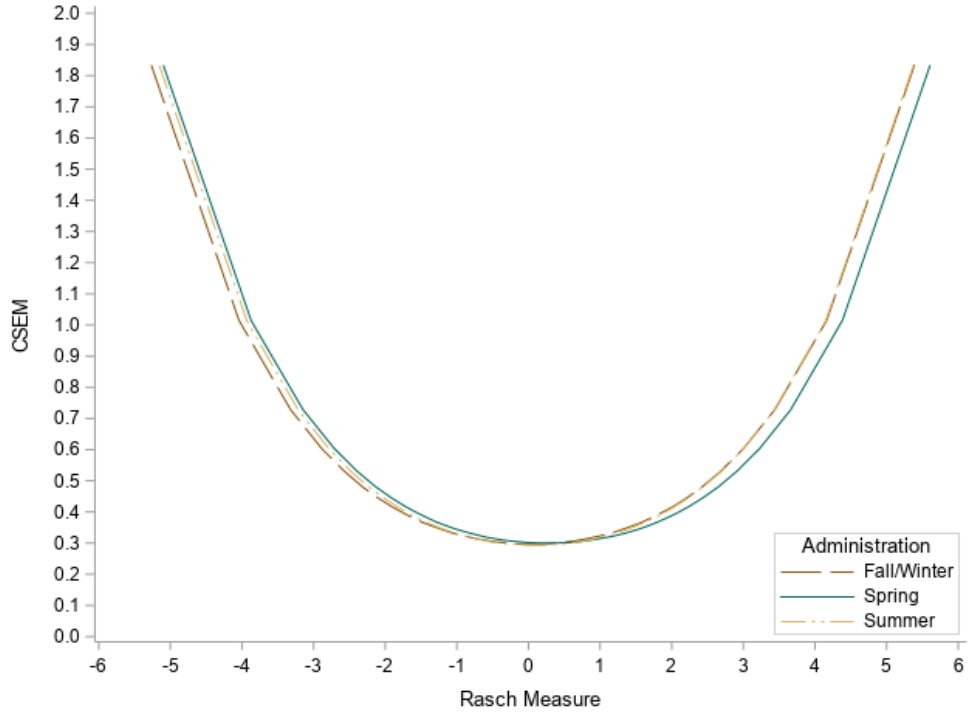


Figure 4.3. Test Characteristic Curves for EOCEP Biology 1 Administrations

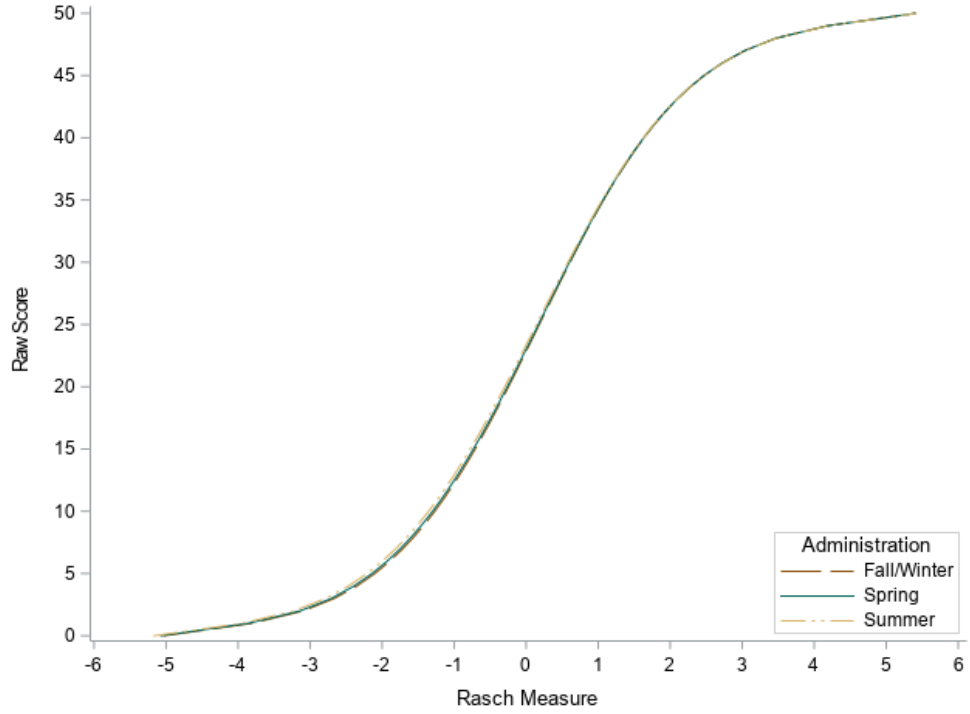


Figure 4.4. CSEM Curves for EOCEP Biology 1 Administrations

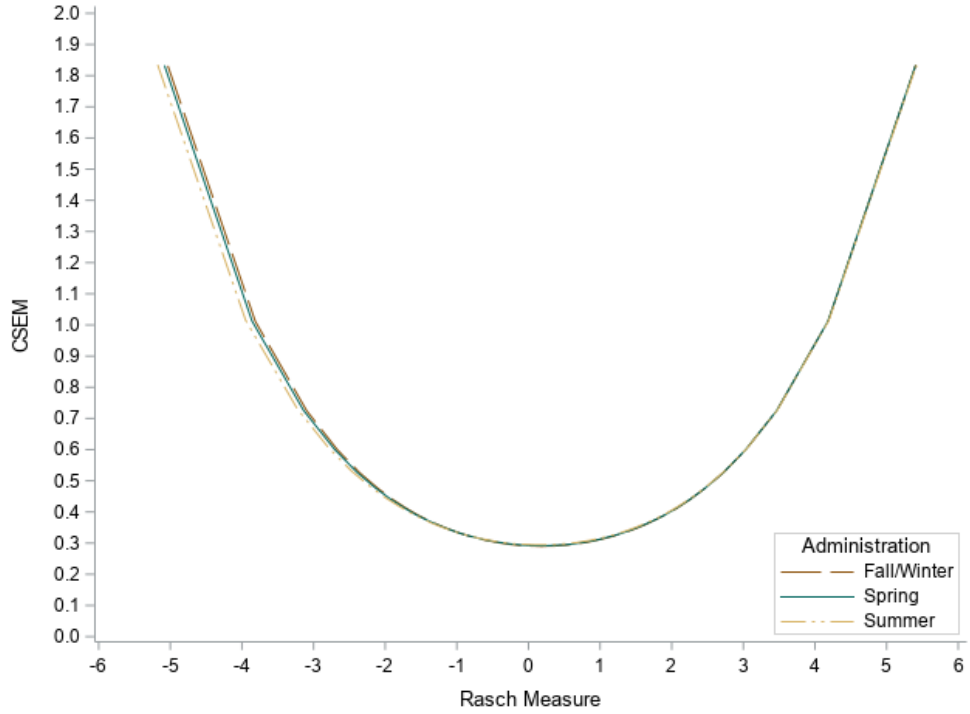


Figure 4.5. Test Characteristic Curves for EOCEP English 2 Administrations

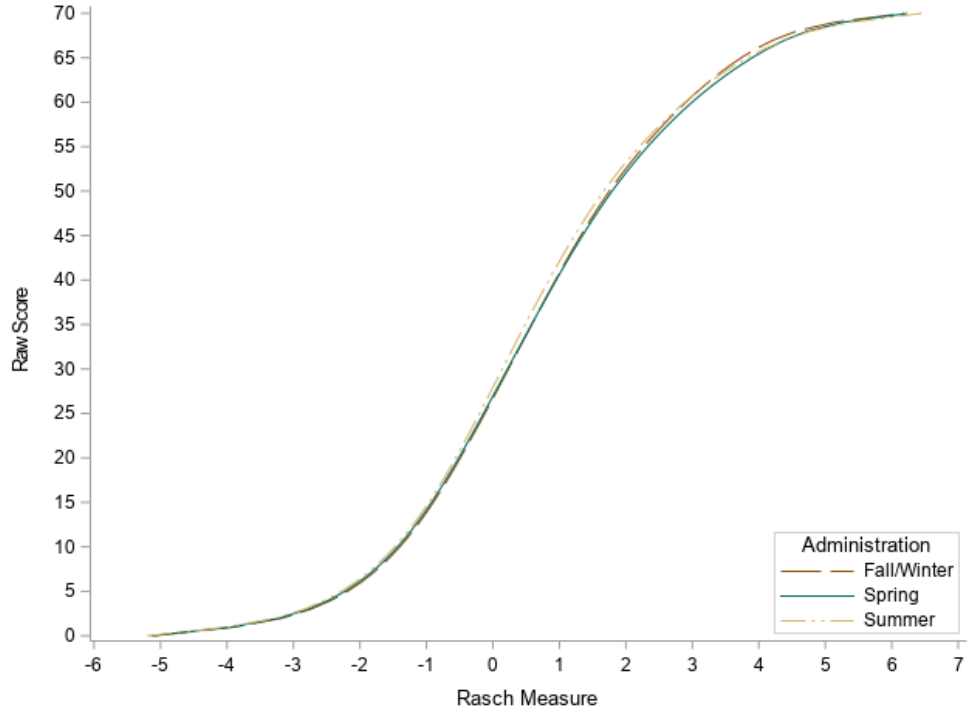


Figure 4.6. CSEM Curves for EOCEP English 2 Administrations

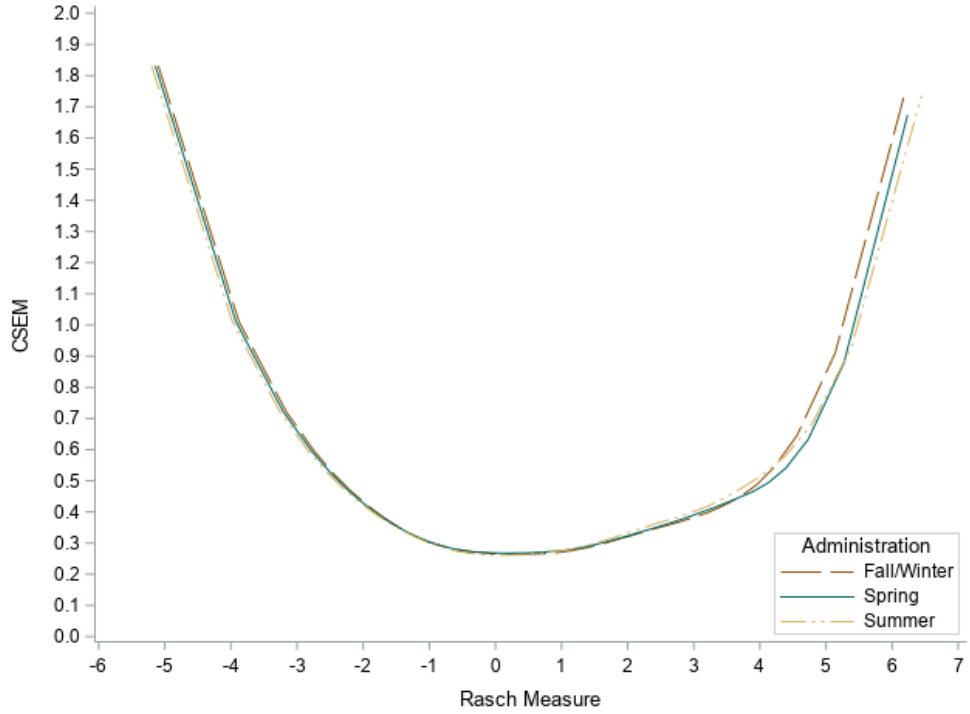


Figure 4.7. Test Characteristic Curves for EOCEP USHC Administrations

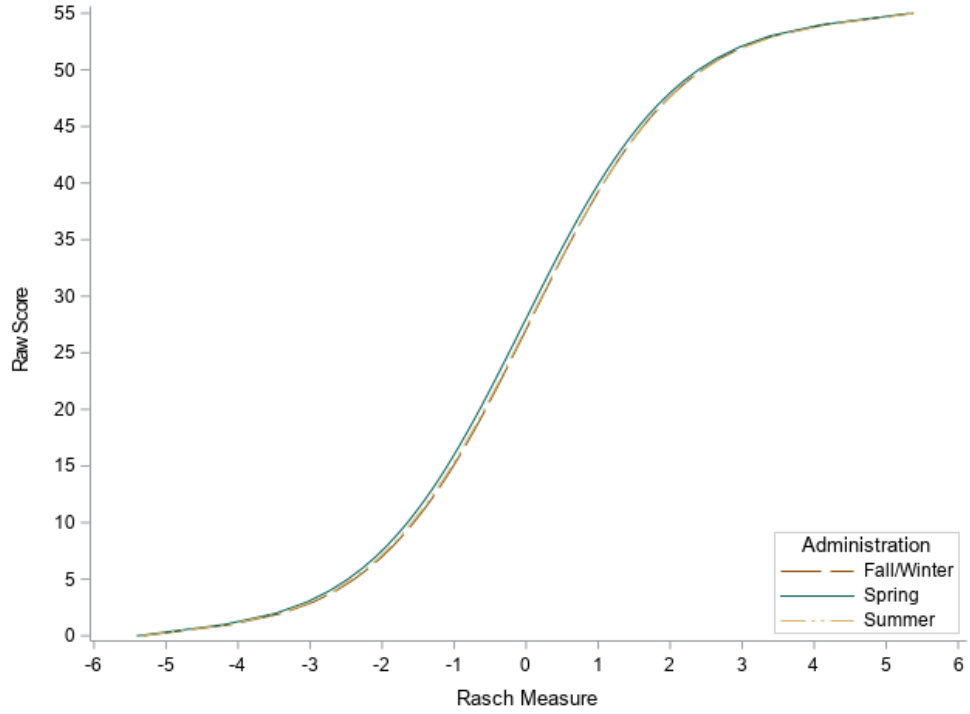
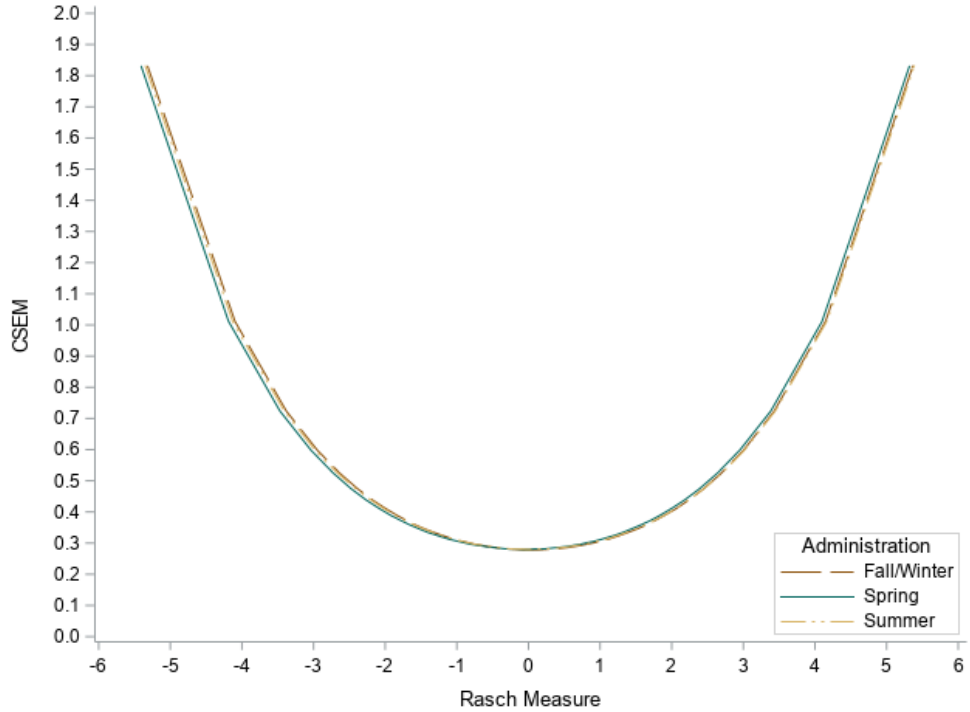


Figure 4.8. CSEM Curves for EOCEP USHC Administrations



4.8 Technical Analyses and Ongoing Maintenance

The appropriate item banks were updated with the operational and field test statistics from the 2023–2024 EOCEP administrations. 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 Section 504 Plan, must participate in the EOCEP English 2 (starting in Fall/Winter 2019–20), Algebra 1, and Biology 1 assessments by the end of their third year in high school. Students with an IEP/Section 504 Plan who are enrolled in U.S. History and the Constitution must participate in the USHC EOCEP. 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 courses can be found at SCDE's website. The website includes information for testing students with IEPs including the testing process guide, allowable accommodations, frequently asked questions, and the South Carolina Accessibility Support Document.

5.2 Procedures for Including Multilingual Learners

The EOCEP assessments are not available in alternate language formats; all ML students must take these tests in English. TAs may not translate any part of the English 2 assessments 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.

Translated versions of the Test Administration Manual (TAM) test directions are available for online and paper/pencil testing. Separate documents are available for each of the following languages: Spanish (Latin America), Russian, Vietnamese, Chinese (Simplified), Portuguese (Brazil), Arabic, Gujarati, Ukrainian, Telugu, and Tamil.

Documentation of procedures for determining student eligibility for accommodations and guidance on selection of appropriate accommodations for ML students in tested courses can be found in the SCDE's Accessibility Support Document. The document 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 IEP, Section 504 Plan, or ILAP. Accommodation code definitions can be found in the TAM.

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

Universal tools and accommodations are permitted on the EOCEP assessments. 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 highlighter or 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 EOCEP 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 Section 504 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 Section 504 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. 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, the TAM contains the list of universal tools and accommodations permissible for the EOCEP 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 DRC INSIGHT Portal.

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 accommodations are properly coded in DRC's INSIGHT portal. There are two means to monitor accommodations use: through the state IEP system 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 circumstances can be found in Appendix C of the TAM. Table 5.1 presents the percentages of students using accommodations for the combined Fall/Winter 2023, Spring 2024, and Summer 2024 test administrations.

Table 5.1. EOCEP Percentage of Students Using Accommodations, Combined Fall/Winter, Spring, and Summer Administrations

Accommodations	Algebra 1	Biology 1	English 2	USHC
Total N	67,719	62,784	65,064	58,699
Setting	5.34	4.71	4.93	4.03
Timing	0.38	0.39	0.36	0.30
Scheduling	0.03	0.02	0.02	0.02
Response Options	0.02	0.01	0.03	0.02
Presentation	0.09	0.08	0.07	0.05
Supplemental Materials	0.05	0.05	0.04	0.05

5.5 Monitoring Test Administration for Special Populations

The state has a monitoring process for reviewing IEPs or Section 504 Plans. 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 Section 504 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 Section 504 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.*

6.1 State Adoption of Academic Achievement Standards for All Students

This section briefly describes the EOCEP assessment standard settings and presents the cut scores derived from each 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 specific documentation from each subject area standard setting. Specifically, USHC was most recently completed in 2014 (Data Recognition Corporation, 2014), Algebra 1 in 2016 (Data Recognition Corporation, 2017) Biology 1 in 2017 (Data Recognition Corporation, 2018), and English 2 in 2019 (Data Recognition Corporation, 2019).

Note that the 2023–2024 Biology 1 assessments were built to updated content standards. Following discussions between test content specialists from both DRC and the SCDE, and on the advice of the Technical Advisory Committee (TAC), scale scores and performance levels for the 2023–2024 assessments were reported using the previously established reporting scale (although reporting category scores were not reported). A standard setting study was conducted for the Biology 1 assessment in summer 2024 using the spring 2024 assessments and performance data. The cut scores from that standard setting study will be incorporated beginning with the Fall/Winter 2024–2025 Biology 1 assessment.

The process of the performance level settings for the EOCEP assessments 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 EOCEP 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

Due to the timing of different content standards being implemented in different academic years, the standard settings for the EOCEP assessments were conducted at different times.

After the development of college- and career-ready content standards, South Carolina proceeded with the determination of challenging academic achievement standards. The SCDE and DRC conducted a standard setting for the Algebra 1 EOCEP tests in Columbia, South Carolina, from July 26 through 29, 2016. The SCDE and DRC also conducted a standard setting meeting for Biology 1 in Columbia, South Carolina, from June 20 through 21, 2017. Additionally, the SCDE and DRC conducted a standard setting meeting for English 2 in Columbia, South Carolina, from July 23 to 24, 2019. For these subjects, three achievement level cut scores (the transitions between Does Not Meet Expectations/Minimally Meets Expectations, Minimally Meets Expectations/Meets Expectations, and Meets Expectations/Exceeds Expectations) were recommended. Additionally, students receive a letter grade based on their EOCEP performance: F, D, C, B, or A.

In 2019, the South Carolina State Board of Education adopted The South Carolina Social Studies College- and Career-Ready Standards (2019). On June 15–16, 2022, SCDE partnered with DRC to conduct a standard setting for the EOCEP USHC assessment. Cut scores for the assessment were developed to divide students into four performance levels: Does Not Meet Expectations, Minimally Meets Expectations, Meets Expectations, and Exceeds Expectations. Additionally, students receive a letter grade based on their USHC performance: F, D, C, B, or A. The four performance levels correspond to the South Carolina Uniform Grading Policy (UGP) where B and C are combined and correspond to Meets Expectations. Refer to the South Carolina End-of-Course Examination Program (EOCEP) U.S. History & the Constitution Standard Setting 2022 Technical Report for full details.

6.1.2 Methodology

For each of the EOCEP standard settings, South Carolina educators from across the state participated in the Bookmark Standard Setting Procedure (BSSP) (Lewis, Mitzel, & Green, 1996; Lewis, Mitzel, et al., 2012). In the BSSP, these educators recommended cut scores for each of the EOCEP assessments.

6.1.3. Workshops

During the performance level setting, participants studied the South Carolina performance level descriptors (PLDs) and South Carolina Content Standards 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 the South Carolina

Content Standards specific to each relevant course. Participants then discussed the content-based expectations for students at the threshold of each performance level (e.g., a student who is barely at the level of “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 committees to examine proposed standards across the EOCEP courses and recommend possible adjustments. After the committee meetings, SCDE considered the 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 each of the EOCEP 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 in Section 6.1.3).

The following general verbal descriptions of the EOCEP performance levels were given to the standard setting committees for Algebra 1, English 2, Biology 1, and USHC:

- **Does Not Meet Expectations**—The student does not meet the expectations of the course content standards.
- **Minimally Meets**—The student minimally meets the expectations of the course content standards.

- **Meets Expectations**—The student meets the expectations of the course content standards.
- **Exceeds Expectations**—The student exceeds the expectations of the course content standards.

More detailed descriptions of the specific concepts and skills are provided for each grade level in the PLDs. PLDs are 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 revised the PLDs.

6.2 Challenging & Aligned Academic Achievement Standards

SCDE follows the practice of making policy-based adjustments (when desired) to panel recommendations within confidence intervals around the cut score recommendations.

There were four major policy criteria to be satisfied in determining final cut scores:

- The cut scores should be based on College- and Career-Ready performance, in line with the stated SCDE policy.
- The cut scores should be linear with respect to the South Carolina UGS (as revised in 2016), on which EOCEP results are reported.
- The cut scores should produce reasonable distributions of scale scores and letter grades. Scale scores range from 0 to 100. Each scale score is assigned a letter-grade equivalent (A, B, C, D, or F) in accordance with the UGS.
- At cut scores that do not translate to integer scale scores, it is possible for a raw score to correspond to a theta below the theta-level cut score yet still translate to a scale score value equal to the scale score cut. In such cases, the reported scale score is reduced by one point to fall below the scale score cut, thereby making the theta and scale score metrics consistent.

The theta value identifying the approximate percentages of students equivalent to that of South Carolina students meeting or exceeding the ACT® college readiness benchmark for the appropriate subject for the 2016 statewide ACT® assessment (Mathematics for Algebra 1, English for English 2, Science for Biology 1) was used to link to college- and career-ready standards by determining the point corresponding to a scale score of 80 (the minimum score that is equivalent to a B). Also, the approximate percentages of students reaching the minimum score (Level 4) on the subject area test necessary to earn an NCRC Silver Certificate (South Carolina's career-ready criterion) on the 2016 statewide ACT® WorkKeys® administration were used to determine the points corresponding to a scale score of 60 (the minimum score that is equivalent to a D). The WorkKeys tests used were Applied Mathematics for Algebra 1 and Biology 1

(since no science test was available) and a combination of the Locating Information and Reading for Information tests for English 2.

6.2.1 Cut Scores

Table 6.1 provides the EOCEP cut scores in terms of the scale scores.

Table 6.1. EOCEP Scale Score Ranges

EOCEP	Scale Score Range	Performance Letter Grade	Performance Level
Algebra 1, Biology 1, English 2, and USHC	0–59	F	Does Not Meet
	60–69	D	Minimally Meets
	70–89	B, C	Meets
	90–100	A	Exceeds

6.3 Reporting

This section contains information on the results of the combined 2023–2024 administrations of the South Carolina EOCEP 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, Minimally Meets 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 on the EOCEP assessments.

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 EOCEP assessments and should not be used for state accountability purposes.

6.3.1 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 South Carolina End-of-Course Examination Program Score Report User's Guide (SCDE, 2024).

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 South Carolina End-of-Course Examination Program Score Report User's Guide (SCDE, 2024) is accessible to parents, teachers, and laypeople alike.

In addition to the student's grade report, 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 2023–2024 administration year, DRC reported the SC EOCEP 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.

The preliminary score report for Algebra 1, Biology 1, and USHC is posted electronically within thirty-six hours after an online test is submitted or after a paper/pencil test is checked in by the testing contractor.

For English 2 online testing, the reading section of the English 2 tests is scored within thirty-six hours; however, handscoring the writing TDA may require up to ten days. Consequently, the English 2 scores are posted electronically ten days after the writing responses are submitted or thirty-six hours after the reading responses are submitted, whichever is later.

Since the number of students who test on paper is very small, the English 2 scores for paper/pencil tests are posted electronically five days after the answer documents are checked in by the contractor.

Score reports are generated for each district and school. Scores for students testing online and students testing with paper are aggregated together. For each fall, spring, and summer administration, schools receive two paper copies of each ISR and one student label per student per test. For each alternative window administration, ISRs are posted electronically. For each fall, spring, and summer administration, district data files, district rosters, school rosters, class summary reports, and class rosters are posted electronically.

Once per year, the following cumulative reports, which combine summer, fall, and spring administrations, are produced, and posted electronically: district summary report, district summary by grade level, district summary by school, school summary report, and school summary by grade level.

6.3.2 Description of Each Type of Report

In this section, descriptions of the following reports are provided: Individual Student Report, Student Score Label, Student Rosters, Summary Reports, Summary by Grade, and District Summary by School.

In compliance with AERA, APA, & NCME (2014) Standard 12.18, the EOCEP 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)*

6.3.2.1 Individual Student Report

The Individual Student Report (ISR) is available as a paper material and through the DRC INSIGHT Portal. The one-page ISRs (printed on both sides) 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. In a table with identifying course information, the student's scale scores and letter grade are provided for all courses. For English 2, the student's TDA score and a Lexile range is provided. A Quantile range is provided for Algebra 1. At the bottom of the page are the descriptions of the performance levels.

On page two, the ISR provides the student's performance for each reporting category for each EOCEP assessment; reporting category performance is classified as "Low,"

“Middle,” or “High.” This classification is based on the subset of items that assess the reporting category.

Sample ISRs for EOCEP assessments are provided in the South Carolina End-of-Course Examination Program Score Report User’s Guide (SCDE, 2024).

6.3.2.2 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 an administration of an EOCEP test. 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 each EOCEP test. The label lists the student’s information, the EOCEP taken, the scale score, and the letter grade. 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 South Carolina End-of-Course Examination Program Score Report User’s Guide (SCDE, 2024).

6.3.2.3 Student Rosters

Student Rosters are provided at district, school, and class levels for each EOCEP assessment and administration. The rosters include the overall scale score range, test administration, course number, student grade level, sex, race/ethnicity, and teacher name. Total test scale scores and performance level letter grades, as well as the course number, are displayed in a table with footnotes at the bottom. A sample Student Roster is provided in the South Carolina End-of-Course Examination Program Score Report User’s Guide (SCDE, 2024).

6.3.2.4 Summary Reports

Summary Reports are provided at district, school, and class levels for each EOCEP assessment and for all EOCEP courses combined across administrations. The summary reports include the overall scale score range, number of students tested, the number and percentage of students with scores in each letter grade category, and descriptive statistics for each level of group (district, school, and class). The descriptive statistics of the scale scores included the standard deviation, mean, median, as well as the lowest and highest observed scale scores for the student group. Footnotes are provided at the bottom of the page. A sample Summary Report is provided in the South Carolina End-of-Course Examination Program Score Report User’s Guide (SCDE, 2024).

6.3.2.5 Summary Reports by Grade

Available from the DRC INSIGHT Portal is a Summary Report generated at district and school levels for each EOCEP assessment and for all EOCEP courses combined across administrations. The summary reports include the overall scale score range, number of students tested, the mean scale score and standard deviation for each grade level. Footnotes are provided at the bottom of the page. A sample Summary Report by

Grade is provided in the South Carolina End-of-Course Examination Program Score Report User's Guide (SCDE, 2024).

6.3.2.6 Summary Reports by School

Available from the DRC INSIGHT Portal is a Summary Report generated at district level for each EOCEP assessment and for all EOCEP courses combined across administrations. The summary reports include the overall scale score range, number of students tested, the mean scale score and standard deviation, and the number and percentage of students with scores in each letter grade category for each school. Footnotes are provided at the bottom of the page. A sample Summary Report by School is provided in the South Carolina End-of-Course Examination Program Score Report User's Guide (SCDE, 2024).

6.3.2.7 The DRC INSIGHT Portal

Schools and districts can access summary level reports through the DRC INSIGHT Portal, DRC's online assessment management and reporting system. The DRC INSIGHT Portal allows school district personnel with appropriate permissions to access EOCEP 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://wbte.drcedirect.com/SC/portals/sc. Each school and/or district is assigned a username and password to access the site.

6.4 Interpreting Test Results

A student's correct responses to the assessment questions are used to derive that student's EOCEP assessment scale scores. The scale score describes performance on a continuum that ranges from 0 to 100 for all EOCEP assessments.

The EOCEP 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 South Carolina Academic Standards for each course.

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 South Carolina End-of-Course Examination Score Report User's Guide (SCDE, 2024), which was developed collaboratively by DRC and SCDE staff.

6.5 Current Administration Results

Results for the 2023–2024 EOCEP combined administrations can be found in Tables 6.2 and 6.3, which provides summaries of the total test scale scores based on the state population for those three administrations, as well as state-level percentages for students in each performance level across all four subjects.

Table 6.2. State-Level EOCEP Scale Score Summary Statistics, Combined 2023-2024 Administrations

EOCEP	N	Mean SS	SD SS	Percentile				
				10th	25th	50th	75th	99th
Algebra 1	67,719	70.36	15.59	50	57	70	82	100
Biology 1	62,784	68.90	18.87	44	53	67	84	100
English 2	64,660	77.84	14.82	57	67	79	90	100
USHC	58,699	67.39	20.23	42	51	66	84	100

Note. Data combines Fall/Winter, Spring, and Summer Administrations.

Table 6.3. State-Level Percentages of Students in Each Performance Level, Algebra 1, Biology 1, English 2, and USHC

EOCEP	N	Does Not Meet – Letter Grade F	Minimally Meets – Letter Grade D	Meets – Letter Grades C & B	Exceeds – Letter Grade A	Meets + Exceeds – Letter Grades C, B & A
Algebra 1	67,719	27.56	21.76	37.42	13.26	50.68
Biology 1	62,784	37.13	15.63	28.11	19.13	47.24
English 2	64,660	14.18	16.29	43.72	25.80	69.52
USHC	58,699	41.25	14.99	25.30	18.47	43.76

Note. Data combines Fall/Winter, Spring, and Summer Administrations.

6.6 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. Table 6.4 shows the cross-year mean scale score and the performance level distributions for Algebra 1, Biology 1, English 2, and USHC. Note that the English 2 administration began in 2019. Additionally, for all EOCEP content areas, the impact of the Covid-19 pandemic and the cancelation of the Spring 2020 test administration may yield results that are not directly comparable across administration years.

Table 6.4. State Level Scale Score Means and Performance Distributions 2018–24, Algebra 1, Biology 1, English 2, and USHC

EOCEP	Year	Student Count	Scale Score Mean	% A	% B	% C	% D	% F
Algebra 1	2018	60,489	68.4	9.1	14.0	20.9	24.0	32.0
	2019	61,278	68.3	10.1	13.0	20.4	25.1	31.4
	2020	11,649	63.7	3.6	8.8	17.0	30.8	39.8
	2021	52,842	65.8	8.7	10.5	15.3	26.8	38.7
	2022	64,923	68.1	10.6	11.6	20.2	23.7	34.0
	2023	67,214	69.1	13.3	11.7	19.8	23.6	31.5
	2024	67,719	70.4	13.3	15.0	22.4	21.8	27.6
Biology 1	2018	56,738	69.5	16.2	16.3	16.7	18.3	32.6
	2019	57,521	68.8	16.1	13.1	17.8	20.2	32.8
	2020	18,978	68.0	16.0	13.2	15.4	19.9	35.5
	2021	51,623	65.2	14.2	10.4	14.7	18.0	42.7
	2022	60,285	66.4	16.8	10.9	14.8	14.9	42.5
	2023	65,080	66.8	16.8	11.4	14.7	16.3	40.8
	2024	62,784	68.9	19.1	12.7	15.4	15.6	37.1
English 2	2020	17,579	75.7	17.6	25.9	21.4	19.8	15.2
	2021	49,863	76.5	22.4	23.8	21.1	16.1	16.6
	2022	60,327	76.5	21.8	23.4	21.7	17.5	15.7
	2023	63,484	77.7	28.6	19.8	19.1	16.6	15.9
	2024	64,660	77.8	25.8	23.2	20.6	16.3	14.2
USHC	2022	53,055	65.1	14.5	10.7	14.2	17.1	43.5
	2023	56,149	67.5	19.5	11.1	14.1	16.2	39.1
	2024	58,699	67.4	18.5	11.2	14.1	15.0	41.3

Note. 2020 includes the data from the Fall/Winter and Summer administrations.

6.7 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. Chapman and Hall.

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

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. McKay.

Brennan, R. L. (2004). BB-Class (Version 1.0). University of Iowa, Center for Advanced Studies in Measurement & Assessment.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Sage Publications.

Cattell, R. B. (1952). Factor analysis. Harper.

Center for Universal Design. (1997). What is universal design? 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/10.3102/10769986022003265.

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

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

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

Data Recognition Corporation. (2017). South Carolina End of Course Examination Program (EOCEP) Algebra 1 and English 1 Standard Setting 2017 Technical Report.

Data Recognition Corporation. (2018). South Carolina End of Course Examination Program (EOCEP) Biology 1 Standard Setting 2018 Technical Report.

Data Recognition Corporation. (2019). South Carolina End of Course Examination Program (EOCEP) English 2 Standard Setting 2019 Technical Report.

Data Recognition Corporation. (2022). South Carolina End of Course Examination Program (EOCEP) U.S. History & the Constitution Standard Setting 2022 Technical Report.

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

Draba, R. (1977). The Identification and Interpretation of Item Bias. Research Memorandum No. 25, Statistical Laboratory, Department of Education, University of Chicago.

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. 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, 1–8.

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.

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). Routledge.

Linacre, J. M. (2018). WINSTEPS Rasch measurement (Version 4.2.0). https://www.winsteps.com.

Linn, R., & Gronlund, N. (1995). Measurement in assessment and teaching (7th ed.). Prentice-Hall.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

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.

O'Neill, T., Peabody, M., Tan, R., & Du, Y. (2013). How Much Item Drift is Too Much? Rasch Measurement Transactions, 27(3), 1423–1424.

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

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

SCDE (2024). South Carolina End-of-Course Examination Program Score Report User’s Guide.

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

Wright, B.D., & Douglas, G. A. (1976). Rasch item analysis by hand. Research Memorandum No. 21, Statistical Laboratory, Department of Education, University of Chicago.

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

Wright, B.D., & Linacre, M. (1994). Reasonable Mean-Square Fit Statistics. Rasch Measurement.

Transactions, 8(3), 370. Retrieved from: https://www.rasch.org/rmt/rmt83b.htm.

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.