Virginia

Standards of Learning Assessments

Technical Report

2014–2015 Administration Cycle

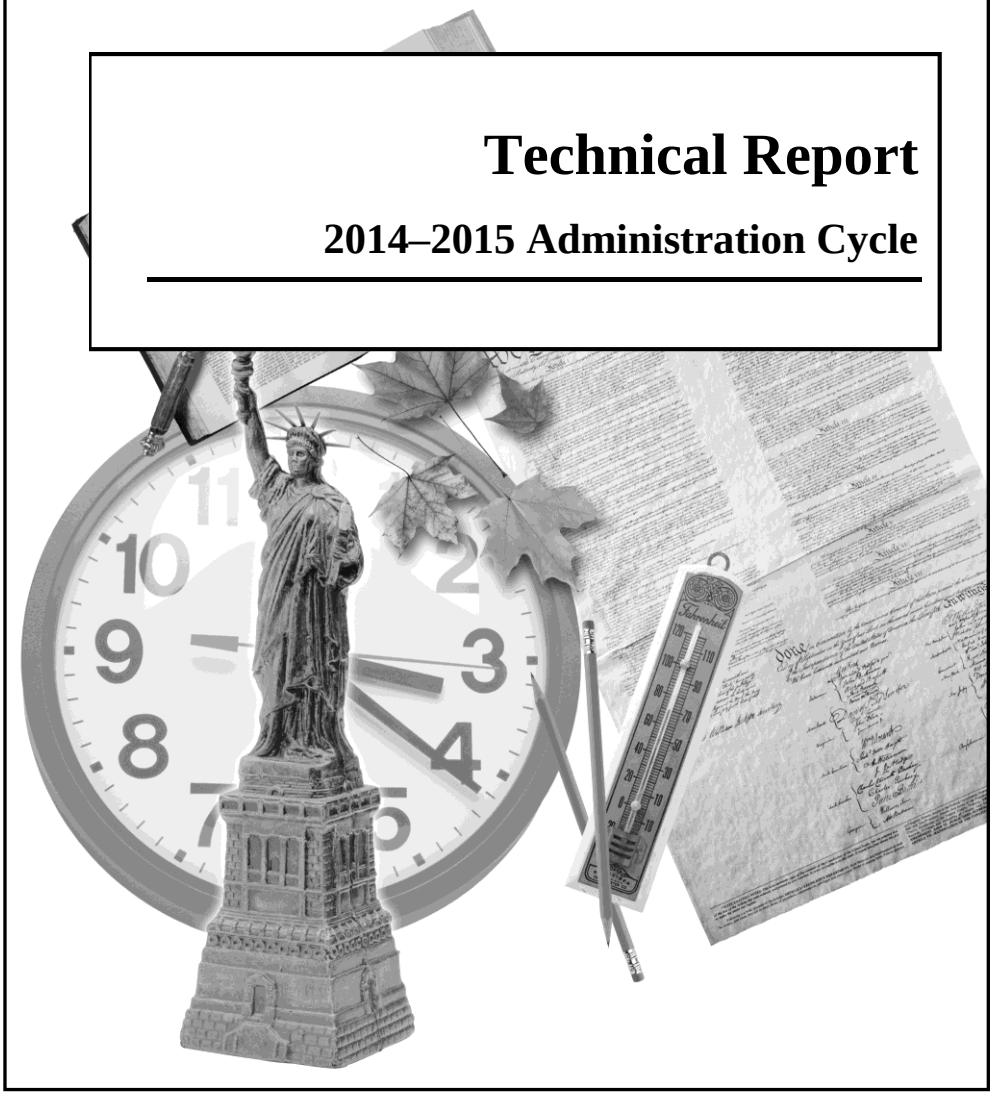


Table of Contents

Table of Contents ii
List of Tables and Figures v
PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROGRAMS 1
1. INTRODUCTION 1
2. STUDENT ASSESSMENTS IN VIRGINIA 1
2.1 Historical Overview of SOL Assessments 1
2.1.1 High Academic Standards 1
2.1.2 Tests to Measure Student Progress on the SOL 1
2.1.3 Accountability for Student Achievement 2
2.2 Overview of Current Virginia SOL Assessments 2
2.2.1 Online Testing in Virginia 3
2.2.2 Computerized Adaptive Testing 3
2.2.3 Current SOL Assessments 3
3. DEVELOPMENT OF SOL ASSESSMENTS 4
3.1 Content Standards, Curriculum Frameworks, and Test Blueprints 4
3.1.1 Standards of Learning (SOL) 4
3.1.2 Curriculum Frameworks 4
3.1.3 Test Blueprints 5
3.2 Item Development 6
3.2.1 New Item Development 6
3.2.2 New Item Content Review Committees 7
3.3 New Writing Prompt Development 7
3.3.1 Specifications and Development 7
3.3.2 New Writing Prompt Review Committees 8
3.4 Field Testing 8
3.4.1 Embedded Field Testing of MC and TEIs 8
3.4.2 Stand-Alone Field Testing of Writing Prompts 8
3.4.3 Sampling 9
3.4.4 Data Review Committees 9
3.4.5 Statistics Reviewed During Data Review Meetings 10
3.5 Test Construction 11
3.5.1 Procedures 11
3.5.2 Test Form Review Committees 12
3.5.3 Procedures for Computerized Adaptive Administrations 12
4. TEST ADMINISTRATION 13
4.1 Training and Materials 13
4.2 Testing Windows 13
4.3 Test Security Procedures 14
4.4 Testing Accommodations 14
4.4.1 Testing Accommodations for Students with Disabilities 15
4.4.2 Testing Accommodations for LEP Students 16
5. WRITING SCORING 17
5.1 Human Scorer Recruitment and Qualifications 17
5.2 Rangefinding 18
5.3 Scorer Training and Qualifying Procedures 18

5.3.1 Anchor Sets 18
5.3.2 Practice Sets 19
5.3.3 Qualifying Sets 19
5.4 Scoring Procedures 19
5.4.1 Two Ratings with Resolution for Nonadjacent Scores 19
5.4.2 Backreading 20
5.4.3 Validity Checks 20
5.4.4 General Calibration Sets 20
5.4.5 Targeted Calibration Sets 20
5.5 Rescore Process 21
6. SCORES 21
6.1 Raw Scores 21
6.2 Total Scale Scores 22
6.3 Reporting Category Scale Scores 22
6.4 Performance Levels 22
7. PERFORMANCE STANDARDS 23
7.1 Standard Setting 23
7.1.1 Performance Level Descriptors (PLDs) 23
7.1.2 Selecting Standard Setting Committees 23
7.1.3 Standard Setting Process 24
7.2 Standard Setting Schedule 25
8. CALIBRATION, SCALING, AND EQUATING 25
8.1 Rasch Partial Credit Model (RPCM) 25
8.2 Scaling 28
8.2.1 Scale Scores for the Total Test 28
8.2.2 Scale Scores for Reporting Categories 29
8.3 Equating 30
8.3.1 Field-Test Equating 30
8.3.2 Pre-Equating 31
8.3.3 Post-Equating 31
9. RELIABILITY 32
9.1 Internal Consistency 33
9.2 Inter-Rater Reliability 34
9.3 Measurement Error 34
9.3.1 Standard Error of Measurement 34
9.3.2 Conditional Standard Error of Measurement 34
9.4 Decision Accuracy and Consistency 35
10. VALIDITY 36
10.1 Validity Evidence Based on Test Content 37
10.1.1 Relation to Content Standards 37
10.1.2 Educator Input on Item Development 38
10.1.3 SOL Assessment Alignment Studies 38
10.2 Validity Evidence Based on Response Processes 38
10.2.1 Item Development 39
10.2.2 Technology Enhanced Items (TEIs) 39
10.3 Validity Evidence Based on Internal Structure 40
10.3.1 Internal Consistency 40

10.3.2 Differential Item Functioning (DIF) 40
10.3.3 Unidimensionality Evaluation 40
10.4 Validity Evidence Based on Relationships to Other Variables 41
11. ALTERNATE AND ALTERNATIVE ASSESSMENTS 41
12. RESOURCES 42
PART II: STATISTICAL SUMMARIES FOR SPRING 2015 43
1. OVERVIEW OF STATISTICAL SUMMARIES 43
1.1 Administration Results 43
1.2 Reliability Estimates for Non-Writing Assessments 43
1.3 Reliability Estimates for Writing Assessments 43
1.4 Decision Consistency and Accuracy Indices 43
1.5 CAT Pool Information 44
1.6 Raw Score-to-Scale Score Conversion Tables and Conditional SEMs 44
2. SPRING 2015 STATISTICAL SUMMARY 44
2.1 Administration Results 44
2.1.1 Participation by Mode of Administration 44
2.1.2 Percent of Students in each Performance Level 46
2.1.3 Scale Score Summary Statistics 48
2.2 Reliability Estimates for Multiple-Choice/Technology Enhanced Item Assessments 51
2.2.1 Overall Reliability Estimates 51
2.2.2 Reliability Estimates by Gender 52
2.2.3 Reliability Estimates by Ethnic Group 54
2.3 Reliability Estimates for Writing Assessments 56
2.3.1 Stratified Alpha 56
2.3.2 Inter-Rater Reliability 60
2.4 Decision Consistency and Accuracy Indices 61
2.5 CAT Pool Information 64
2.5.1 Number of Items in each CAT Pool Overall and by Reporting Category 64
2.5.2 Rasch Item Difficulty Distribution Overall and by Reporting Category 64
2.5.3 Item Exposure Rates 68
2.5.4 Pool Usage 68
2.5.5 On Target Rates for Meeting Content and Statistical Criteria 69
2.5.6 CSEM 69
2.6 Raw Score to Scale Score (RSSS) Conversion Tables and Conditional SEM 72
REFERENCES 95

List of Tables and Figures

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROGRAMS 1
Table 2.2.3.1 Virginia Standards of Learning Assessments at Each Grade Level 3
Table 7.2.1 Virginia SOL Standard Setting Schedule 25
Figure 8.1.1 Sample Item Characteristic Curve for a Dichotomous Item 27
Figure 8.1.2 Sample Category Response Curves for a Polytomous Item 28
Figure 8.3.3.1 Linking Process for the SOL Assessments 32
PART II: STATISTICAL SUMMARIES FOR SPRING 2015 43
Table 2.1.1.1 Percentage of Tests Taken by Mode: Grades 3–8 44
Table 2.1.1.2 Percentage of Tests Taken by Mode: Content-Specific History 45
Table 2.1.1.3 Percentage of Tests Taken by Mode: End-of-Course 45
Table 2.1.1.4 Percentage of Tests Taken by Mode: VMAST Mathematics and Reading 46
Table 2.1.2.1 Grades 3–8 Passing Rates 46
Table 2.1.2.2 Content-Specific History Passing Rates 47
Table 2.1.2.3 End-of-Course Passing Rates 47
Table 2.1.2.4 VMAST Grades 3–8 & End-of-Course Passing Rates 47
Table 2.1.3.1 Summary Statistics for Grades 3–8 Reading Online 48
Table 2.1.3.2 Summary Statistics for Grade 6 Mathematics Online 48
Table 2.1.3.3 Summary Statistics for Grades 5 and 8 Science Online 48
Table 2.1.3.4 Summary Statistics for High School End-of-Course Online 49
Table 2.1.3.5 Summary Statistics for Grades 8, and End-of-Course Writing Tests 49
Table 2.2.1.1 Reliability for Grades 3–8 Reading 51
Table 2.2.1.2 Reliability for Grade 6 Mathematics 51
Table 2.2.1.3 Reliability for Grades 5 and 8 Science 51
Table 2.2.1.4 Reliability for High School End-of-Course Tests 52
Table 2.2.2.1 Reliability for Grades 3–8 Reading by Gender 53
Table 2.2.2.2 Reliability for Grade 6 Mathematics by Gender 53
Table 2.2.2.3 Reliability for Grades 5 and 8 Science by Gender 53
Table 2.2.2.4 Reliability for High School End-of-Course Tests by Gender 54
Table 2.2.3.1 Reliability for Grades 3–8 Reading by Ethnic Group 54
Table 2.2.3.2 Reliability for Grade 6 Mathematics by Ethnic Group 55
Table 2.2.3.3 Reliability for Grades 5 and 8 Science by Ethnic Group 55
Table 2.2.3.4 Reliability for High School End-of-Course Tests by Ethnic Group 55
Table 2.3.1.1 Reliability for Grade 8 and End-of-Course Writing 56
Table 2.3.1.2 Reliability for Grade 8 and End-of-Course Writing by Gender 57
Table 2.3.1.3 Reliability for Grade 8 and End-of-Course Writing by Ethnic Group 59
Table 2.3.2.1 Inter-Rater Reliability for the Grade 8 Writing Assessment: Prompts 2809, 2813, 2816, 2817, 2853, and 2854 60
Table 2.3.2.2 Inter-Rater Reliability for the End-of-Course Writing Assessment: Prompts 2101, 2103, 2106, 2108, 2116, and 2124 60
Table 2.4.1 Decision Consistency and Accuracy Indices for Grades 3–8 Reading Online Tests 61
Table 2.4.2 Decision Consistency and Accuracy Indices for Grade 6 Mathematics Online Test 62

Table 2.4.3 Decision Consistency and Accuracy Indices for Grades 5 and 8 Science Online Tests 62
Table 2.4.4 Decision Consistency and Accuracy Indices for High School End-of-Course Online Tests 62
Table 2.4.5 Decision Consistency and Accuracy Indices for Grade 8 and End-of-Course Writing Tests 63
Table 2.5.1.1 Number of Items in the Mathematics CAT Pools Overall and by Reporting Category 64
Table 2.5.2.1 Rasch Item Difficulty for Items in the Mathematics CAT SOL Pool Overall and by Reporting Category 64
Table 2.5.2.2 Rasch Item Difficulty for Items in the Mathematics CAT Plain English Pool Overall and by Reporting Category 65
Table 2.5.2.3 Rasch Item Difficulty for Items in the Mathematics CAT SOL Audio Pool Overall and by Reporting Category 65
Table 2.5.2.4 Rasch Item Difficulty for Items in the Mathematics CAT Plain English Audio Pool Overall and by Reporting Category 65
Figure 2.5.2.1 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT SOL Pool 66
Figure 2.5.2.2 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Plain English Pool 66
Figure 2.5.2.3 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Audio Pool 67
Figure 2.5.2.4 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Plain English Audio Pool 67
Table 2.5.3.1 Maximum Item Exposure Rate by Mathematics CAT Pool 68
Figure 2.5.3.1 Item Exposure Rates for Grade 6 Mathematics CAT Pools 68
Table 2.5.4.1 Percentage of Items Used within each Mathematics CAT Pool 69
Table 2.5.6.1 Average CSEM for each Mathematics CAT Pool 69
Table 2.5.6.2 Average CSEM at the Basic, Proficient, and Advanced Cut Scores for each Mathematics CAT Pool 69
Figure 2.5.6.1 Average CSEM across the Score Scale for the Grade 6 Mathematics SOL CAT Pool 70
Figure 2.5.6.2 Average CSEM across the Score Scale for the Grade 6 Mathematics Plain English CAT Pool 70
Figure 2.5.6.3 Average CSEM across the Score Scale for the Grade 6 Mathematics SOL Audio CAT Pool 71
Figure 2.5.6.4 Average CSEM across the Score Scale for the Grade 6 Mathematics Plain English Audio CAT Pool 71
Table 2.6.1 RSSS Conversions for Grade 3 Reading 72
Table 2.6.2 RSSS Conversions for Grade 4 Reading 73
Table 2.6.3 RSSS Conversions for Grade 5 Reading 74
Table 2.6.4 RSSS Conversions for Grade 5 Science 75
Table 2.6.5 RSSS Conversions for Grade 6 Reading 76
Table 2.6.6 RSSS Conversions for Grade 7 Reading 77
Table 2.6.7 RSSS Conversions for Grade 8 Reading 78
Table 2.6.8 RSSS Conversions for Grade 8 Science 79
Table 2.6.9 RSSS Conversions for EOC Reading 80

Table 2.6.10 RSSS Conversions for EOC Earth Science 81
Table 2.6.11 RSSS Conversions for EOC Biology..... 82
Table 2.6.12 RSSS Conversions for EOC Chemistry..... 83
Table 2.6.13 RSSS Conversions for EOC Algebra I 84
Table 2.6.14 RSSS Conversions for EOC Geometry 85
Table 2.6.15 RSSS Conversions for EOC Algebra II..... 86
Table 2.6.16 RSSS Conversions for EOC World Geography 87
Table 2.6.17 RSSS Conversions for Grade 8 Writing Core 1 89
Table 2.6.18 RSSS Conversions for Grade 8 Writing Core 2 90
Table 2.6.19 RSSS Conversions for Grade 8 Writing Core 3 91
Table 2.6.20 RSSS Conversions for EOC Writing Core 1 92
Table 2.6.21 RSSS Conversions for EOC Writing Core 2 93
Table 2.6.22 RSSS Conversions for EOC Writing Core 3 94

PART I: HISTORICAL OVERVIEW AND SUMMARY OF PROGRAMS

1. INTRODUCTION

The Virginia Standards of Learning (SOL) Assessment Program Technical Report provides information for users and other interested parties about the development and technical characteristics of the assessments within the Virginia Assessment Program. The SOL technical report is divided into two parts. Part I presents a summary of the components of the Virginia SOL assessment program from the 2014–2015 administration cycle. Part II provides statistical information based on results from spring 2015.

2. STUDENT ASSESSMENTS IN VIRGINIA

2.1 Historical Overview of SOL Assessments

In 1994, Virginia initiated significant reform of its K–12 educational system. This reform, which has evolved over the last 20 years, consists of several major elements discussed in the following sections: high academic standards, tests to measure progress, and accountability.

2.1.1 High Academic Standards

In 1995, the Virginia Board of Education adopted a set of statewide standards: the Virginia SOL. The Virginia SOL set forth minimum learning standards for every child from K–12 in English, mathematics, science, and history/social science. Over time, the SOL were expanded to include the areas of family life, economics and personal finance, fine arts, foreign language, computer technology, health and physical education, and driver education.

The board recognized the need for regular review and evaluation of the SOL; therefore, in September 2000, it approved a cyclical schedule for the review of the standards. This has resulted in each subject area undergoing a review and potential revision every seven years[1].

2.1.2 Tests to Measure Student Progress on the SOL

Development of tests to measure the SOL began in 1996 with heavy involvement of classroom teachers, curriculum specialists, and other local educators throughout Virginia. A statewide census field test of the new SOL test items took place in the spring of 1997. The first administration of SOL tests took place in the spring of 1998, and the program has expanded significantly since that time.

The SOL assessment program is the cornerstone of Virginia's system of accountability for public schools and is authorized in Virginia law and administrative rules (see Article I, Section 15 and

[1] The review cycle can be accessed at the following website: http://www.doe.virginia.gov/testing/assessment_committees/review_schedule.pdf

Article VIII of the Constitution of Virginia and Title 22.1 Chapter 13.2 § 22.1-253.13:3C, Code of Virginia). The purposes of the assessment program are to

- establish and communicate high levels of achievement on the SOL for Virginia public school students;
- provide communication that indicates the progress of students and schools toward meeting achievement levels on the SOL;
- provide information that can be used to improve instructional programs; and
- provide assurance of the quality of public education.

The federally enacted No Child Left Behind Act of 2001 (NCLB) reinforced many strategies already present in Virginia's public education system. For a number of years, public educators throughout the commonwealth have focused on instructional standards, student assessment, reporting of results, and continuous improvement. To respond to NCLB, Virginia has maintained its rigorous academic content standards, measuring students against defined academic performance standards, added grade-level assessments in various subjects, and reported on the progress of student subgroups at the school, the division, and the state levels. The Virginia Assessment Program has been used to meet state and federal educational requirements including:

- monitoring the progress of students and schools toward meeting established achievement levels;
- identifying the educational needs of students;
- determining which achievement levels students have attained;
- determining whether students receive a high school diploma; and
- providing accountability information for school, school division, and state levels.

2.1.3 Accountability for Student Achievement

The Standards of Accreditation (SOA) for Virginia's public schools outlines the state requirements for student testing and graduation, as well as the requirements for the accreditation of schools in the commonwealth. The SOA may be found on the website of the Virginia Department of Education: http://www.doe.virginia.gov/boe/accreditation/.

2.2 Overview of Current Virginia SOL Assessments

The Virginia SOL assessments are standards-based tests designed to measure student performance on Virginia's content standards in the areas of reading, writing, mathematics, science, and history/social science. The SOL tests contain primarily multiple-choice items, however the mathematics, English, and science assessments also include technology-enhanced items (TEI).[2] TEIs are developed in a variety of formats that allow students to indicate their responses in ways other than the multiple-choice format. The writing tests administered at grade 8 and high school include writing prompts in addition to multiple-choice and TEI items.

[2]Technology-enhanced item (TEI) types were used operationally on all assessments with the exception of paper-pencil tests and History/social science forms.

2.2.1 Online Testing in Virginia

In the 2000 session of the general assembly, legislation was passed that required and funded a statewide web-based technology initiative. The goal of this initiative was for Virginia school divisions to implement online, web-based SOL instruction, remediation, and testing in Virginia's high schools. The initiative provided funding for school divisions to purchase hardware, software, and to upgrade network and Internet capabilities.

Because the initial focus of the project was Virginia's high schools, the online testing initiative began with the end-of-course (EOC) SOL tests. The first online EOC tests were administered in fall 2001. Since that time, additional SOL tests have been move to the web-based delivery system in a phased approached so that all tests are now available in the online system.

2.2.2 Computerized Adaptive Testing

As part of the continuing effort to provide students with the best possible testing experience, Virginia added computerized adaptive testing (CAT) to the Virginia Standards of Learning Assessment Program in grade 6 mathematics during the 2014–2015 school year. The Virginia Department of Education plans to phase-in CAT for math and reading assessments in grades 3–8 during the next few years. For additional information about Virginia's implementation of CAT, refer to the *Virginia Standards of Learning Computerized Adaptive Testing Technical Manual*.

2.2.3 Current SOL Assessments

In 2014–2015, students in grades 3–8 and high school were tested using SOL assessments in the content areas listed in Table 2.2.1. The Virginia General Assembly eliminated five SOL assessments beginning with the 2014–2015 academic year: Grade 3 Science, Grade 3 History, Grade 5 Writing, US History to 1865, and US History 1865 to Present.

High school tests were designed to address specific course content, regardless of the student's current enrolled grade. The content-specific history assessments are not grade-level dependent and are typically taken in the upper elementary or middle school years.

Table 2.2.3.1 Virginia Standards of Learning Assessments at Each Grade Level

	Grade Level							
SOL Content Area	**3**	**4**	**5**	**6**	**7**	**8**	**Content-Specific History**	**High School**
English: Reading	•	•	•	•	•	•		•
English: Writing						•		•
Mathematics	•	•	•	•	•	•		
Science			•			•		
Algebra I								•
Geometry								•

SOL Content Area	Grade Level							
	3	4	5	6	7	8	Content-Specific History	High School
Algebra II								•
Virginia and U.S. History								•
World History I								•
World History II								•
World Geography								•
Earth Science								•
Biology								•
Chemistry								•
Virginia Studies							•	
Civics and Economics							•	

3. DEVELOPMENT OF SOL ASSESSMENTS

The Virginia Department of Education works jointly with Virginia educators and its testing contractor to develop a series of tests to measure student achievement on the SOL content standards. The development of the SOL assessments involves the use of test blueprints, item development specifications, multiple review committees, and field testing.

3.1 Content Standards, Curriculum Frameworks, and Test Blueprints

3.1.1 Standards of Learning (SOL)

The SOL represent a broad consensus of what parents, classroom teachers, and school administrators—as well as academic, business, and community leaders—believe schools should teach and students should learn. In each of the four core areas of English, mathematics, science, and history/social science, a curriculum framework is provided that details the specific knowledge and skills students must possess to meet the content standards for these subjects. The SOL are reviewed and updated on a seven-year cycle.

3.1.2 Curriculum Frameworks

The SOL Curriculum Frameworks[3] amplify the SOL and provide additional guidance to school divisions and their teachers as they develop an instructional program appropriate for their students. The curriculum frameworks assist teachers as they plan their lessons by identifying the essential knowledge and skills students need to learn.

School divisions use the curriculum frameworks as a resource for developing sound curricular and instructional programs, but the curriculum frameworks are not intended to limit the scope of instructional programs. Additional knowledge and skills that can enrich instruction and enhance

[3]The curriculum frameworks and test blueprints can be accessed at the following website: http://www.doe.virginia.gov/testing/

students' understanding of the content identified in the SOL should be included as part of quality learning experiences.

3.1.3 Test Blueprints

The SOL test blueprint[3] serves as a guide for test construction. Each test covers a number of SOL. In the test blueprint, SOL are grouped into categories that address related content or skills. These categories are called *reporting categories*. When the results of the SOL tests are reported, the scores will be presented in terms of scores for each reporting category and a total test score. Each SOL is assigned to only one reporting category.

The number of test items that will be assessed in each reporting category, as well as on the test as a whole can be found in the test blueprint. Because of the large number of SOL in each grade-level content area, every SOL will not be assessed on every version (form) of an SOL test. By necessity, to keep the length of a test reasonable, each test will sample from the SOL within a reporting category. However, every SOL is eligible for inclusion on each form of an SOL test. In some content areas, there are SOL that do not lend themselves to assessment within the current format of the SOL tests. The SOL not tested are listed as "Excluded from Testing" at the end of the blueprint for each test.

There is a specific blueprint for each test. Each blueprint contains three components relevant to each SOL test: general test information, a blueprint summary table, and the expanded blueprint.

The general test information section provides information about the following topics:

- test blueprint;
- reporting categories;
- assignment of SOL to reporting categories;
- Standards of Learning excluded from testing;
- coverage of SOL; and
- use of curriculum frameworks.

A summary table of the blueprint displays the following information:

- reporting categories for each test;
- number of test items in each reporting category;
- SOL included in each reporting category;
- SOL excluded from the SOL test;
- number of operational items on the test;
- number of field-test items on the test; and
- total number of items (operational and field-test items) on the test.

The expanded blueprint provides full text for each SOL. In addition, SOL that are excluded from the test are categorized by the reason they were not included.

For grade 6 mathematics, the blueprint also contains information for two types of tests, the online CAT and the traditional test. Because tests administered with CAT tailor the items to each

student's ability level, fewer items are needed to accurately measure student knowledge and skills. Therefore, the number of SOL items per reporting category and at the total test level, reflected in the blueprint, are fewer for tests administered using CAT compared to tests administered non-adaptively.

3.2 Item Development

3.2.1 New Item Development

.

Each year Virginia's assessment contractor conducts an evaluation of the items available for traditional test forms construction and CAT. Item pool distributions resulting from this evaluation map the counts of items by SOL, by item type, by Rasch item difficulty estimates, and by cognitive complexity level. Based on the evaluation of existing items, an item development plan is developed for each test. Following approval of the item development plans by the Virginia Department of Education, new items are developed to address the needs identified by the evaluation. All items assess content specified by the SOL and further defined in the associated curriculum frameworks.

The item development process is multi-phased and involves a variety of expert groups. Item writers external to the testing vendors' staff are trained on requirements for the SOL assessment program. Item writers author items in accordance with item development plan-driven assignments and the SOL item specifications. Item writers are experienced in item authoring for K-12 statewide assessments and have teaching experience in their assigned subject matter and grade span. Items are developed for the primary presentation mode of online delivery.

Testing contractors' content/assessment specialists review and edit newly submitted items for content accuracy and grade-level appropriateness and for adherence to principles for quality item construction, accessibility (i.e., universal design[4]), and fairness (e.g., bias, sensitivity, and limited English proficiency). Content/assessment specialists are usually former teachers in their designated subject matter and grade span. Each item is coded for the SOL it is intended to measure. Items are reviewed and edited to ensure the annual batch of new items meets expected distributions of item difficulty and cognitive complexity levels as required by the SOL being assessed.

In addition to the initial review by the contractor's assessment specialists, there are a series of internal item reviews involving different staff expertise. These reviews include content reviews, a professional editorial review, and a fairness review. Additional guidance and feedback is provided regarding the appropriateness of the content match to the SOL and adherence to item specifications through Virginia content review committee meetings, as well as reviews completed by Virginia Department of Education following these meetings.

[4]The application of the principles of universal design to assessments entails a blend of good test design, consideration of as many users as possible, assistive technology where appropriate, and builds in appropriate visual design (Dolan & Hall, 2001).

Item specifications are determined by the Virginia Department of Education for appropriate assessment of the SOL

3.2.2 New Item Content Review Committees

On an annual basis, Virginia educators from across the state participate in the development of the SOL assessments. Every summer, content review committees convene to review content materials for the SOL program. Content committees are composed primarily of educators teaching the subject of the test, including special education teachers. A small number of committee members may be division curriculum staff or other school division employees. They represent all grade levels—grade 3 through high school—all content areas, and the racial/ethnic diversity of Virginia students. Committee members also represent a geographical cross section of Virginia. Approximately one-third of the members of every committee are new each year in order to provide a balance of experienced educators and new members and to bring new perspectives into committee meetings. The committee members review the newly developed test items to confirm that they appropriately and fairly measure student knowledge and skills in accordance with the SOL and curriculum frameworks.

The committee members receive an orientation to the SOL assessment program, an overview of the test development process, and information about their roles. Training focuses on educators making judgments about the match of content to SOL, the appropriateness of the content for the grade level, and fairness and accessibility issues. Committees meet separately by grade level and subject. In addition to reviewing the match of the items to the SOL, content review committee members also identify and note their concerns regarding potential item bias in the areas of gender, racial/ethnic, religious, socioeconomic, and regional characteristics. Additionally, item bias concerns regarding students with disabilities and limited English proficiency (LEP) may be noted. Following discussion, the committee as a whole recommends that an item be accepted, edited, or rejected. Each committee member is also provided an individual comment (input) form. While committee recommendations are made by consensus, committee members also record their individual recommendations which may differ from the committee consensus, on the comment forms. All recommendations are tallied, and all comments are compiled into a master document that becomes an official record of the committee review. Only after committee recommendations are counted and comments recorded, is the final decision about an item made. As a result of the new item review process, some items are eliminated from the prospective field-test set, while others are edited in the manner directed for field testing.

3.3 New Writing Prompt Development

3.3.1 Specifications and Development

Writing prompts are used to assess students' writing skills on the SOL writing tests in grade 8 and EOC. New writing prompts are developed and field tested every four or five years as needed to support test construction. English language arts content specialists and item writers draft large numbers of potential writing prompts. Each writing prompt adheres to SOL specifications and is written in the form of a question, an issue, or a hypothetical situation.

3.3.2 New Writing Prompt Review Committees

As needed, the summer writing content review committees are asked to provide input on new writing prompts including evaluating the prompt's clarity, appropriateness for the SOL, similarity to prior prompt topics,[5] and perceived ability of the prompt to elicit an extended written student response. The review process is similar to that used for the review of new MC and TEIs. The committee as a whole provides a consensus recommendation, with individual members' comments captured on prompt comment forms. Based on committee feedback, edits may be made to the prompts prior to field testing.

3.4 Field Testing

Once items and prompts have been developed, reviewed, and approved by the content review committees and the Virginia Department of Education, they are eligible for inclusion on a field test.

3.4.1 Embedded Field Testing of MC and TEIs

Field-test items are embedded within the spring test forms in such a way that they appear throughout the operational test form and are not identifiable to students. This allows for the collection of data on the new items that is not impacted by motivation, as might occur if the students knew that the new items did not contribute to their score.

The position of the field-test items is pre-determined for each core form, as is the number of field-test variations per core. Each form has the same number of field-test items to keep the test length consistent. The number of field-test forms is determined based on how many new items need to be field tested in a given year.

For tests administered using CAT (i.e., grade 6 mathematics), students are randomly assigned a set of field-test items. These field-test items are not adaptive, but are constructed similarly to the field-test sets used for the traditional core forms. The field-test items are interspersed throughout the CAT test so that students do not know when a particular item is a field-test item or an operational item.

In the fall and summer administrations where the population taking the test is not representative of the state's student population, place-holder items are included in the field-test positions to maintain consistent test lengths. These items do not contribute to a student's score, nor are the data used to update item statistics.

3.4.2 Stand-Alone Field Testing of Writing Prompts

For writing tests, new prompts are field tested as needed using a separate, stand-alone field-test administration. Typically, new writing prompts are developed and field tested every four to five years. The last stand-alone field test for writing occurred during the 2011–2012 administration.

[5] New prompts are compared to old prompt pools to make sure that the same topic is not used again.

3.4.3 Sampling

During each spring test administration, test forms are distributed throughout the commonwealth in a way that will facilitate timely equating and the collection of representative field-test data. The manner in which test forms are distributed across the school divisions is called the sampling plan. The sampling procedures are based on data files containing participation counts that schools submit prior to the spring administration. These files indicate the number of students in each school who will take each test online or in paper-and-pencil format. In conjunction with the participation counts, the school division's graduation date and the date that school closes for the year are considered when assigning test forms in the sampling plan.

An attempt is made to assign test forms to divisions in such a way that approximately equal numbers of students respond to each field-test variation across the cores.[6] Also, test forms are assigned at the school division level so that all schools are administered the same core for a given test. The core that is assigned to a division by the above process is labeled the *Main* form for that division. Each division is also assigned an alternate form. The alternate form is utilized in retesting students in the case of a testing irregularity. For instance, an administrator may need to assign a different test form if the student becomes ill during a test, or if there is a disruption that prevents the student from completing the test.

The MC/TEI section of the writing tests is assigned to divisions in the same way as the non-writing tests. In addition, there are six to seven writing prompts that are administered each spring. Of the six to seven prompts, four or five are new writing prompts that must be equated; the other prompts have been equated during a previous administration. In order to obtain enough data to calibrate each new prompt for equating purposes, the new prompts are randomly assigned across the divisions.

The sampling process described above is not needed for CAT tests because a different equating design is used. This design is called pre-equating (see Section 8 for more information). Pre-equating does not require additional student data, so students can respond to any set of items in the CAT pool and still receive a score report quickly. In addition, students are administered different sets of operational items, and the set of items administered to a student during a first attempt is blocked for that student during a second attempt. This process provides a similar solution to the main and alternate forms available for non-adaptive administrations. Also, sets of field-test items are randomly assigned to students during the administration process. Therefore, by design, approximately equal numbers of students respond to each field-test item.

3.4.4 Data Review Committees

In addition to reviewing new items during the summer item content review meetings, Virginia educators review field-tested items. During the data review meeting, committee members' are asked to review the appropriateness of items' content, using the field-test item statistics to inform their judgments, as appropriate. During the data review meeting, committees recommend

[6]Each core generally contains multiple forms. The core and linking items are identical across each form, but the embedded field-test items vary.

accepting or rejecting items. As with new item review, comment (input) forms are the official record of committee activity.

The same committee that reviews the writing MC items and TEIs also reviews writing prompt results (when available). As part of the training for the prompt reviews, committee members are provided with information on the scoring rubrics. The Virginia SOL Writing Field Test Prompt Evaluation Form, which Virginia's testing contractor completes as an evaluation of the prompts, is also provided to the committee. This form is a hybrid of qualitative and quantitative information. During the scoring process for field-tested prompts, scorers and team leaders record their observations about the student responses to each prompt. Team leaders then compile a qualitative report that addresses the following questions:

- Did the students understand what the prompt asked them to do?
- Did the students seem engaged by the prompt?
- Were the students able to effectively focus on a central idea and provide specific information and details?
- Did the scorers, based upon reading hundreds of student responses to the prompt, recommend that this prompt be used for live testing?

The report also includes the following pieces of information for each prompt:

- final frequency distribution of prompt scores
- suggestions and comments from the scorers
- several examples of students' written responses

Committee members review the prompt and responses to ascertain whether the prompt actually elicited responses that are complete and well elaborated. Members also review the prompt itself for appropriate content and to ensure fairness for all students. A prompt that elicits responses that are similar to lists, or a prompt that seems to confuse students is considered to be poorly performing and is usually recommended for rejection. In some circumstances, a prompt will be recommended for field testing again at a different grade level. Feedback and comments from committee members is added to the final report.

3.4.5 Statistics Reviewed During Data Review Meetings

For the purpose of reviewing the quality of new test items, reviewers are provided with various data to assist them in decision-making. These data include classical statistics and item response theory (IRT) statistics (Rasch measurement model).

The classical statistics calculated for the MC items/TEIs include

- numbers of students tested overall and by gender and ethnic group (African American, Caucasian, Hispanic, and Asian);
- traditional difficulties (*p*-values);
- item-option response distributions for all respondents by gender and ethnic group; and
- point-biserial correlations.

Classical statistics computed for field-tested writing prompts include

- numbers of students tested overall and by gender and ethnic group (African American, Caucasian, Hispanic, and Asian); and
- frequency distributions, means, and standard deviations for the writing domain raw and total scores.

To supplement the classical statistics, item difficulty parameter estimates based on the Rasch IRT model are computed. More information about the Rasch model is included in Section 8 of this report.

Mantel-Haenszel Alpha and the associated chi-square significance test are computed to evaluate differential item functioning (DIF). The Mantel-Haenszel Alpha is a log/odds probability indicating when it is more likely for one of the demographic groups to answer a particular item correctly. When this probability is significantly different across the various groups, the item is flagged for further examination.

Response distributions for each demographic group indicate whether members of a group were drawn to one or more of the answer choices for the item. If a large percentage of a particular group selected an answer choice not chosen by other groups, the item is inspected carefully. Statistical analyses merely serve to identify test items that have unusual characteristics. They do not specifically identify items that are "biased;" such decisions are made by item reviewers who are knowledgeable about the state's content standards, instructional methodology, and student testing behavior.

3.5 Test Construction

3.5.1 Procedures[7]

New core operational test forms are generally used for the first time in the spring test administration. For non-writing tests, generally three new core forms are developed annually for all EOC assessments, except EOC World Geography. Typically, three new core forms are also developed annually for all writing tests (at grade 8 and EOC). For all other SOL tests two new core forms are typically developed annually. In some cases, fewer core forms are developed and core forms from previous years are reused.

Test specifications and test construction guidelines are developed and approved by the Virginia Department of Education. Test construction guidelines provide the operational process and the established expectations (both psychometric and content characteristics) to guide SOL forms assembly. The goal is to create test forms that are equivalent in content representation and psychometric characteristics both within a year and across years.

[7] These procedures apply to non-adaptive tests only. See Section 3.5.3 for information about tests administered using CAT.

A common item linking design is used year to year. Items from a core form from the prior spring test administration are placed in the new administration's two or three core forms and serve as the anchor items. Anchor items are placed in the same, or nearly the same sequence positions in the new core form that they appeared in on the old form. For tests with items that are associated with passages (reading and writing), the passages (and associated items) are placed in as close as possible to the same position within the new test form as they were placed in the prior form. Anchor items represent approximately 20–30% of the operational forms. Content specialists select anchor items, and psychometrics and the Virginia Department of Education approve them.

Following the approval of anchor items, content specialists select the remaining operational items for each test. During the test construction process, psychometricians review each form to see whether it meets the test specification blueprint and the statistical targets established to produce forms of similar difficulty, statistical quality, and content representation within and across years.

These draft forms are reviewed by the Virginia Department of Education. Any item replacements are reviewed by psychometricians. The review process continues iteratively until the Virginia Department of Education has provided final approval.

3.5.2 Test Form Review Committees

The newly drafted operational test forms for each SOL assessment are reviewed by content review committees at the summer meetings. The new core forms are reviewed for each SOL test. Committee members receive training for this task including information on the match of the items on a form to the SOL test blueprint, the arrangement of items within the form, and the balance of topic coverage and item types. Members are asked to confirm the appropriateness of the item content and the accuracy of the keys.

Individual committee members have comment forms to record their overall evaluation of the test form, as well as comments on individual items. Individual members' comments and recommendations are compiled into a master document following the meeting. Committee review may result in the need to replace one or more items on a core form. These changes are subject to review and approval by psychometrics and the Virginia Department of Education.

Once operational test cores are finalized, content specialists select field-test items and create multiple sets that are embedded into the core forms to create multiple core form variations. These field-test sets are reviewed and approved by the Virginia Department of Education.

For tests administered with CAT, the entire item pool is used to build on-the-fly tests customized to each student's ability level. As such, a test form review for CAT tests is not used. However, the item and data review meetings are convened for new items that are under consideration for inclusion in the CAT item pool.

3.5.3 Procedures for Computerized Adaptive Administrations

Tests administered using CAT do not follow the same test construction process as traditionally administered tests. Instead of having an annual test building cycle, a process is established for

embedding test construction into the CAT algorithm which selects items on-the-fly from a large pool of SOL items. Item selection is based on statistical and content constraints that have been established and approved by the Virginia Department of Education. These constraints are built to mimic many of the same types of statistical and content constraints used to build traditional test forms (e.g., meeting the test blueprint, having items with adequate point-biserials). In order to implement these constraints, additional metadata must be created and added to the item bank for items that are part of CAT pools. These additional metadata allow for more targeted selection of items during the adaptive administration process. Therefore, the test construction procedures and timing for CAT are very different from the test construction process and timing used to build traditional test forms.

Each year, the CAT item pools will be evaluated after the administration to ensure that they are functioning as intended and allow for the adaptive algorithm to meet the content and statistical constraints. Results of this evaluation are provided in Part II of this report. For more information about CAT test construction, refer to the *Virginia Standards of Learning Computerized Adaptive Testing Technical Manual*.

4. TEST ADMINISTRATION

4.1 Training and Materials

To ensure the successful administration of the SOL assessments, Virginia Department of Education staff provides training to the division directors of testing (DDOTs) before each fall and spring test administration. DDOTs in turn provide appropriate training to the divisions' school test coordinators (STCs). STCs provide training to the schools' examiners and proctors including information about security requirements, manuals, local directions received from the DDOTs, and other pertinent information. They address training preparation of the test sites and the provision of accommodations for eligible students. In addition, the Virginia Department of Education provides a standardized training for examiners and proctors that is available to school divisions if they choose to use it.

Test implementation manuals contain detailed instructions about administration procedures. These manuals are provided on the Virginia Department of Education website: http://www.doe.virginia.gov/testing/test_administration/index.shtml.

4.2 Testing Windows

There are three test administrations: spring, summer, and fall. The spring administration is the main administration during which most students test. During the spring administration SOL assessments for all grades and subjects are provided. The summer administration is available for only grade 8 math and reading for those students who are pursuing a modified standard diploma and for all EOC tests. This administration provides an opportunity for students to retest who are enrolled in summer school for EOC courses, for students who need to retake an SOL tests to meet graduation requirements, and for transfer students who are seeking to meet the testing requirements for graduation. .The fall administration is available for only grades 6–8 and EOC. This administration is available for students who are retesting to meet graduation requirements

and for students taking courses with block schedules who complete a course during the fall. Some Virginia schools provide a block schedule for grades 6–8.

A fairly long testing window is provided for online assessments so that schools have enough time to coordinate student access to computers and to accommodate different school calendars across the state. Divisions can choose the weeks within the testing window during which they will administer the assessments, leaving sufficient time for make-up testing. The testing window for the writing assessments is earlier than the testing window for the non-writing assessments to provide extra time for the human-scoring of the short paper (essay) component of the assessment. In addition, the MC item/TEI component and the short paper component can be administered separately. Unlike the online tests, paper writing tests are administered on a specific day.

The testing calendar is posted on the Virginia Department of Education website: http://www.doe.virginia.gov/testing/test_administration/index.shtml.

4.3 Test Security Procedures

Everyone in the school division who has access to, or assists with the administration of the paper-and-pencil or online SOL assessments must read the *Test Security Guidelines* and sign the *Test Security Agreement*. The security agreement requires that those involved in the test administration exercise the necessary precautions to ensure the security of the test content and all test materials. This includes security procedures pertinent to the receipt, inventory, distribution, and storage of test materials. These forms are included in each examiner's manual and testing implementation manual.[8]

4.4 Testing Accommodations

All students in tested grade levels and courses are expected to participate in Virginia's assessment program, unless specifically exempted by state or federal law or by Board of Education regulations. Virginia's assessment system includes students with disabilities and LEP students. Students with disabilities and LEP students may take SOL tests with or without accommodations or they may be assessed through alternate or alternative assessments. The tests that comprise the Virginia Assessment Program are offered in English only; administration of the tests in other languages is not permitted.

The individualized education program (IEP) team or 504 committee has the responsibility for decisions regarding the need for and the selection of accommodations for students with disabilities. Similarly, the LEP committee determines how LEP students will participate in the SOL assessments and what, if any, accommodations should be provided to individual LEP students. Accommodations allow students with disabilities or LEP designation more appropriate access to test content so they can demonstrate their content knowledge.

[8]These manuals may be downloaded from the following website: http://www.doe.virginia.gov/testing/test_administration/

Accommodations considered for testing should be those that the student uses routinely during classroom instruction and assessments, as identified in the student's IEP, 504 plan, or LEP participation plan. The student should be familiar with an accommodation because the use of an unfamiliar accommodation during testing may have a negative impact on the student's performance. However, it is important to note that certain accommodations used for instruction or classroom assessment may not be allowed on the statewide assessment. Finally, providing an accommodation based solely on its potential to enhance performance beyond allowing for more appropriate access is inappropriate.

4.4.1 Testing Accommodations for Students with Disabilities

There are many assessment options for students with disabilities. These include the SOL assessments without accommodations, the SOL assessments with accommodations, and alternative (on-grade level) or alternate assessments including the Virginia Substitute Evaluation Program (VSEP), the Virginia Grade Level Alternative (VGLA)[9], the Virginia Modified Achievement Standards Test (VMAST)[10], and the Virginia Alternate Assessment Program (VAAP). Information on state assessment options available to students with disabilities is provided in the *Students with Disabilities: Guidelines for Assessment Participation* document available on the Virginia Department of Education's website.

http://doe.virginia.gov/testing/participation/index.shtml.

The SOL assessments must be considered by the IEP Team or 504 Committee before alternate/alternative assessments are considered. Although many students with disabilities will be able to access the SOL assessments without accommodations, others will require test accommodations to address their disabilities and individual needs. Test accommodations for students with disabilities are grouped in the following categories: time/scheduling, setting, presentation, and response. The accommodations available within each of these categories are provided in the table below and are described in more detail on the Virginia Department of Education website.

http://www.doe.virginia.gov/testing/participation/guidelines_for_special_test_accommodations.pdf.

[9] VGLA is available for qualifying students with disabilities in 3-8 writing, science, and history.
[10] VMAST assessments for End-of-Course (EOC) Algebra I and EOC Reading will be available for eligible students with disabilities pursuing a Standard Diploma with credit accommodations. VMAST assessments for grade 8 mathematics and grade 8 reading are available for students with disabilities pursing a Modified Standard Diploma. The Modified Standard Diploma is available only to students who entered the 9th grade for the first time prior to the 2013-2014 school year.

Timing/Scheduling Accommodations	**Setting Accommodations**
Adjust the scheduling of a test: • time of day • planned breaks during test * • flexible schedule (multiple test sessions) * • order of tests administration	Adjust the place in which the testing normally occurs: • individual testing (one-on-one) • special lighting • adaptive or special furniture • test administered in locations with minimal distractions • hospital/home/non-school setting
Presentation Accommodations	**Response Accommodations**
Adjust the presentation of test material and/or test directions: • visual aids * (e.g., interactive/electronic whiteboard, colored overlay, tinted screen, magnifying glass, large monitor, screen magnifier, graphic organizers, templates, masks or markers to maintain place) • amplification equipment (e.g., auditory trainer, whisper phone) • large-print test * • braille test * • Plain English version of a *Mathematics* test • reading directions to students • written directions accompanying oral directions * • reading of test items aloud * • audio version of test items * • interpreting/transliterating directions (e.g., sign language, cued speech) * • interpreting/transliterating test items (e.g., sign language, cued speech) * • specific verbal prompts *	Adjust the manner in which students respond to or answer test items: • enlarged copy of the answer document * • Examiner records responses * • augmentative communication device * • communication board or choice cards * • brailler * • word processor or word processor with speech-to-text * • spelling aids * • English dictionary * • dictation using a recording device • dictation to a scribe * • read back student response * • word prediction software * • calculator or arithmetic tables * • math aids * • calculator with additional functions * • dry erase board * • additional markers, highlighters, colored pens, and/or pencils *

4.4.2 Testing Accommodations for LEP Students

Testing accommodation determinations for LEP students, made by the LEP Committee, should be based on the evidence collected from the student's educational record, such as:

- demographic information, including grade, age, number of years in U.S., prior schooling;
- standardized testing scores, the ACCESS for ELLs[11] test scores, and other academic testing achievement;
- current academic achievement, including general education achievement and comments from general education teachers; and
- English Language Proficiency Level as reported on the ACCESS for ELLs score report.

[11] Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs®) is Virginia's English language proficiency assessment.

There are two types of accommodations available for LEP students on the Virginia SOL assessments—direct and indirect linguistic accommodations.

Direct linguistic testing accommodations involve adjustments to the language of the test. The following direct linguistic testing accommodations are available to LEP students on the SOL assessments:

- Read-Aloud Test (English only)
- Audio Test (English only)
- Bilingual Dictionary
- Dictation to a Scribe (Writing, short-paper component only)
- English Dictionary
- Plain English Mathematics Test (grades 3 through 8 and Algebra I)

The plain English mathematics test versions include items that incorporate simpler language but still measure the full breadth and depth of the SOL mathematics content standards. Item writers are trained to use the following guidelines.

- Avoid words that have double meaning
- Avoid potentially unfamiliar words
- Use short sentences
- If a large amount of information is needed, break it up into bullets

Indirect linguistic testing accommodations involve adjustments to the conditions under which LEP students take SOL tests. The following indirect linguistic testing accommodations are available to LEP students on the SOL assessments:

- Flexible schedule
- Visual Aids
- Student indicates a response

Additional information about the accommodations available for LEP students on the Virginia SOL assessments is provided on the Virginia Department of Education website.

http://www.doe.virginia.gov/testing/participation/lep_guidelines.pdf.

5. WRITING SCORING

5.1 Human Scorer Recruitment and Qualifications

The constructed response portion of the SOL writing assessment is scored by human raters. Highly qualified, experienced raters score all writing samples. These raters are drawn from a database of college graduates who completed the selection process for scorers. The need for ethnic and racial diversity is emphasized throughout the selection process. Scorers for the Virginia SOL writing test have a minimum of a bachelor's degree in an appropriate academic

discipline (e.g., English, education), demonstrated ability in performance assessment scoring, and, preferably, teaching experience at the elementary or secondary level. The selection process requires that each candidate successfully complete a personal interview, online scorer training, and attain a high enough score on a qualification activity.

In addition to the scorers, scoring supervisors, scoring directors, and content specialists are involved in the scoring process. Scoring supervisors are selected based on their proven ability to score responses accurately and communicate scoring standards to scorers. Scoring directors are chosen based on their expertise in evaluating writing and their experience training and supervising scorers. A writing content specialist monitors quality and provides support and direction for scoring directors. The content specialist is assigned based on educational background and scoring experience.

5.2 Rangefinding

The writing samples used for training scorers are selected from the samples scored during the rangefinding process. Rangefinding is the process of identifying model writing samples for the two writing domains (composing/written expression and usage and mechanics) that characterize each score point of the writing rubric (1–4). Scoring directors and the content specialists work with Virginia educators who served on the rangefinding committees to create training sets. These writing samples, and others identified by Virginia Department of Education and testing contractor staff, are used as scoring guides during scorer training, qualifying, and calibration. The primary goal of the training is to convey the decisions made during rangefinding to the scorers and to help them internalize the scoring protocol.

5.3 Scorer Training and Qualifying Procedures

Scorers are required to take several training modules where they learn about using the online scoring system and implementing the scoring rubric. Each student essay response receives a score on a scale of 1–4 points for each of the domains. The four rubric score points represent the following:

- 4 = consistent control
- 3 = reasonable control
- 2 = inconsistent control
- 1 = little or no control

Training also includes a review of the anchor sets identified during rangefinding. These sets represent each of the four rubric scores and both writing domains. The sets have been scored and annotated so that scorers learn how and why the scores were determined. Next they complete practice sets which include additional writing samples. Finally, they receive qualification writing sets which they must score accurately in order to continue participating in the scoring of student essays.

5.3.1 Anchor Sets

Scorers review an anchor set for each of the two domains. Anchor sets include clear examples of each score point (1–4) and include annotations that provide a rationale for the scores assigned during rangefinding. These papers help the scorers internalize the scoring rubric.

5.3.2 Practice Sets

After reviewing each anchor set, scorers practice on sample papers, applying a score for the domain they are reviewing. The sets include examples of each score point. After applying scores to practice papers, scorers review annotations. The annotations provide feedback on true scores (the scores approved by the rangefinding committee) and explain the correct scores for the practice papers.

5.3.3 Qualifying Sets

In order to qualify to score the Virginia SOL writing assessment, scorers take four sets of 10 papers and must achieve 70% perfect agreement and 100% adjacent agreement with rangefinding-committee-approved scores for each domain on two of the four sets. Scorers who do not meet these standards are released from the project.

5.4 Scoring Procedures

5.4.1 Two Ratings with Resolution for Nonadjacent Scores

During each administration, all writing responses are scored twice. Prior to 2014–2015, the two scores were provided by two professional human scorers. During summer 2013, a feasibility study was conducted to evaluate the feasibility of using automated scoring to provide one of the two initial scores. The study showed that automated scoring could be used successfully to score SOL writing short-paper responses in this manner. The automated scoring engine was trained on all SOL writing prompts using student responses collected during the field-testing. Beginning in 2014–2015, the automated scoring engine was used to provide the second score for each student response to facilitate more rapid reporting of student scores. The same scoring rules and quality control checks are used to ensure accurate scoring of student writing prompt responses.

After a student response is scored by a human rater and the automated scoring engine, the scores are compared. If the scores match or are adjacent, they are summed together to create the final student score. If they are not adjacent, a scoring supervisor provides a third score. If the third score matches one of the first two, then it is summed with the matching score to provide the final score. If the third score is adjacent to one (or both) of the first two scores, then the third score and the higher of the adjacent scores are summed together to provide the final score. If the three scores do not match, and are not adjacent, then the response is scored a fourth time by a scoring director. If the fourth score matches any of the first three scores, then it is summed with the matching score which provides the final score. If the fourth score is adjacent to any of the first three scores, then it is summed with the highest adjacent score to provide the final score.

For responses that receive a backread score, this third score can overrule the outcome from the first two scores. For example, if the first two scores were 2, but the backread provided a score of

3, then instead of adding the first two scores together to provide the final score, the backread score and the highest adjacent score would be added together. So in this example, the student would receive a final score of 5 instead of the original score of 4. If the backread score was 4 in this example, it would not match or be adjacent to the first two scores, so the response would require a fourth score.

5.4.2 Backreading

Backreading is a system that allows a scoring supervisor and/or a scoring director to monitor an individual rater's score. Scoring supervisors read responses already scored by scorers they are monitoring. While backreading, the scoring supervisor can evaluate the scorer's performance, provide feedback, and—if necessary—adjust an assigned score. The scoring supervisor may also halt the scoring activity of an individual or group of scorers whose performance has declined. The inter-rater reliability requirements are 65% perfect agreement and 96% perfect plus adjacent agreement for all scorers. Approximately 5% of papers are included in the backreading process.

5.4.3 Validity Checks

Throughout scoring, scorers receive and score validity papers. These papers are pre-scored according to rangefinding standards. All scores on validity papers are approved by the Virginia Department of Education. Validity papers are used to monitor consistency in scoring over time; they are interspersed with and indistinguishable from other student responses. Approved true scores for these papers are loaded into the scoring system, and a report is run that indicates what percentage of accuracy a scorer achieves on validity papers in scoring against the true score. Validity papers are used as a check to ensure that scorers, as well as scoring supervisors, do not drift from the rubric but continue to score accurately. For scorers that do not meet the 65% perfect and 96% perfect plus adjacent agreement, targeted calibration sets are sent to further evaluate a scorer and provide re-training.

5.4.4 General Calibration Sets

Calibration is a process whereby scorers apply scores to student papers that had been scored previously by scoring directors and possibly a scoring supervisor. Calibration sets include student responses and are used as a training tool to improve agreement among scorers. After scorers take a calibration set, annotations provide an explanation for the correct scores for the responses.

Calibration is a form of training that creates consensus and accuracy among scorers. Calibration sets may focus on particular scoring issues—including clarifying a scoring line, showing a response that is unusual or problematic to score, or showing a range of responses or performance skills for a particular score point. The scoring directors present scorers with calibration sets daily throughout scoring.

5.4.5 Targeted Calibration Sets

Targeted calibration sets are similar to the general calibration sets. They include pre-scored student responses. However, like the qualifying sets, scorers must achieve 70% exact agreement

and 100% exact plus perfect agreement in order to be allowed to continue scoring. Scorers who do not attain the target accuracy rates are dismissed from the scoring process.

5.5 Rescore Process

The primary purpose of the rescore process is to provide an additional step to ensure that the score assigned to the student's writing sample produced as part of the writing tests is an accurate representation of the student's achievement.

In fall of 2014 the automatic rescore process was eliminated for term graduates. All rescoring requests now follow the same process which is initiated by school divisions. Requests to rescore a student's writing sample may be initiated by parents or by school personnel. All requests must be reviewed and approved by the school division before being submitted. Requests should be considered only if there is substantial evidence that the writing sample should have received a higher score. School division staff familiar with the rubric used to score this assessment must review the writing sample. Rescore requests should be approved by the school division only if the reviewers agree that the paper should have received a higher score according to the rubric. A school division may request that a student's writing sample be rescored if

- the student failed the test; **AND**
- there is evidence that the writing sample produced by the student for the writing test should have received a higher score. Evidence of this requires that at least two school division staff familiar with the rubric used to score the writing short-paper portion of the writing test review the paper and agree that it should have received a higher score.

6. SCORES

6.1 Raw Scores

A raw score represents the number of points a student received for correctly answering questions on a test. For the SOL non-writing tests that consist of MC items and TEIs only, the raw score that a student earns is equal to the number of items the student answers correctly. For the SOL writing tests that have a MC/TEI component and a short-paper component, the raw score of the short-paper component is calculated as the weighted sum of the ratings given for each domain[12] and the total raw score is the sum of the raw scores on the two components (MC/TEI plus short paper).

Because different tests have different item types and different numbers of questions, the raw score is only useful in relation to that test or content area. For example, consider a student who receives a raw score of 59 on mathematics and a raw score of 43 on reading. Now imagine that there were 75 total items on the mathematics test and 50 total items on the reading test. In simple terms, this can mean the respective percentage correct would be 79% for the mathematics test and 86% for the reading test.

[12]Each essay is scored on two elements: 1) composing and written expression and 2) usage and mechanics.

Raw scores cannot be used for comparing student performance across different tests or test forms because they are affected by test length and difficulty. Raw scores are comparable only within a given test form. To make comparisons of student performance within a test across years, raw scores must be converted to scale scores.

6.2 Total Scale Scores

A scale score is a conversion of a student's raw score on a test to a common scale. The common scale allows for a numerical comparison of student scores across different years and versions of a specific test. Because Virginia uses multiple versions of a test within a grade and subject, the scale is used to control slight variations in difficulty from one version of a test to the next. For all SOL tests, the scale scores range from 0 to 600. Sometimes a scale score can be estimated as less than 0 (for very low raw scores), or greater than 600 (for very high raw scores). In these cases, to keep the range consistent, scale scores below 0 are set to 0 and scale scores above 600 are set to 600. In addition, at a raw score of 0, the scale score is always set to be 0, and at the top raw score (100% correct), the scale score is always set to 600 (see Section 8.2 for more information on Scaling).

While the scale scores can be used for comparisons across test forms within an SOL test, they cannot be compared across different SOL tests. For example, scale scores cannot be used to reliably determine whether a student or group of students is stronger in reading than in mathematics.

6.3 Reporting Category Scale Scores

SOL are grouped into categories that address related content or skills. These categories are called reporting categories.[13] For each SOL assessment, reporting category scale scores are reported in addition to the overall test scale score. There are varying numbers of reporting categories for the SOL assessments. For each assessment, the reporting category scale scores range from 0 to 50, with a 30 indicating approximate mastery of the content covered by that reporting category.(see Section 8.2 for more information on Scaling).

Reporting category scale scores allow only comparisons within a given reporting category. A reporting category scale score cannot be used to reliably determine whether a student or group of students is stronger on one reporting category compared to another reporting category.

6.4 Performance Levels

In addition to test scores, performance levels are reported on all SOL assessments. Students are classified into performance levels on the basis of their scale scores as compared with the performance level cut scores, which were adopted by the Virginia Board of Education based on recommendations of educators who participated in standard setting meetings. For the reading and

[13]A list of the reporting categories for a given SOL assessment can be found in the test blueprints: http://www.doe.virginia.gov/testing/

mathematics assessments in grades 3–8, there are four performance levels: fail/below basic, fail/basic, pass/proficient, and pass/advanced. For all other assessments, there are three performance levels: fail/does not meet, pass/proficient, and pass/advanced or advanced/college path for EOC Reading, EOC Writing, and Algebra II. For all SOL assessments, the cut score for the pass/proficient level corresponds to a scale score of 400, and the cut score for the pass/advanced level corresponds to 500.

Regardless of which form or test administration a student takes a particular SOL test, the same level of achievement is required to obtain a scale score of 400 for pass/proficient and a scale score of 500 for pass/advanced. For each SOL assessment, the scale scores that represent pass/proficient and pass/advanced remain the same over years, but they may correspond to different raw scores across test forms and administrations. The fluctuation of raw scores does not mean that the requirements for the performance levels have changed. It only reflects changes in difficulty across test forms.

7. PERFORMANCE STANDARDS

7.1 Standard Setting

Performance standards relate levels of test performance directly to what students are expected to learn, as defined in the statewide curriculum. This is done by establishing cut scores that distinguish between performance levels. Standard setting is the process of establishing the cut scores that define the performance levels for an assessment.

7.1.1 Performance Level Descriptors (PLDs)

Performance level descriptors (PLDs) are statements of what a student should know and be able to do at each performance level given the content standards being assessed. In grades 3–8 reading and mathematics, there are four performance levels that a student may achieve: fail/below basic, fail/basic, pass/proficient, and pass/advanced. For all other assessments, three performance levels exist: fail/does not meet, pass/proficient, and pass/advanced or advanced/college path[14] for EOC Reading, EOC Writing, and Algebra II.

7.1.2 Selecting Standard Setting Committees

Standard setting committee meetings are used to establish cut scores on the SOL assessments that operationalize the PLDs. Virginia educators participate in the meetings and recommend cut scores. These educators are recruited by the Virginia Department of Education based on the following criteria:

[14]A student obtaining the proficiency level of advanced/college path on the Algebra II test should have the necessary knowledge and skills for enrollment, without remediation, in an introductory credit-bearing college mathematics course with Algebra II as the highest prerequisite. A student obtaining a proficiency level of advanced/college path on the EOC Reading test should have the necessary knowledge and skills for enrollment, without remediation, in an introductory credit-bearing college course with a substantial reading load. A student obtaining a proficiency level of advanced/college path on the EOC Writing test should have the necessary knowledge and skills for enrollment, without remediation, in an introductory credit-bearing college course with a substantial writing load.

- instructional training and experience in the content area;
- in-depth knowledge of the SOL content standards;
- instructional experience with students who have disabilities and/or LEP students; and
- balanced regional representation.

Additionally, higher education faculty are recruited to participate in setting the EOC performance standards for tests that have an advanced/college path performance level (e.g., EOC Reading, EOC Writing and Algebra II).

7.1.3 Standard Setting Process

Standard setting panelists are guided through an established process to help them identify cut scores that make sense and reflect the PLDs.

Standard setting panelists participate in a general session in which the reason for setting new performance standards is described and an overview of the standard setting procedure is given. Following the general session, committees break into subject- and grade-specific groups, where they review the PLDs and define threshold descriptions which further define what "just barely basic", "just barely proficient" and "just barely advanced" students know and can do. After discussion and general agreement about what the threshold level descriptions are for the basic, proficient and advanced performance levels, the panelists review the operational test in order to experience it as a student would. For the SOL writing tests, they also review examples of short paper responses as well as scoring rubrics.

The methods in the Angoff (1971) family of standard setting procedures are among the most widely used approaches employed for selected-response (MC) exams. The particular variation of Angoff used in to set standards for the Virginia SOL assessments is called the Yes/No method. For MC items, panelists review each operational test item and evaluate whether or not the "just barely proficient" and "just barely advanced" student would be likely (at least 2/3rds of the time) to get the item correct (yes or no). MC score recommendations for each performance level are calculated by recoding each 'yes' to the value of one (1) and each 'no' to the value of zero (0) and then summing across all items for each individual.

For the short paper component on the writing assessments, another variation of Angoff known as the Expected Task Score approach is used. Using this method, panelists evaluate whether or not the "just barely proficient" and "just barely advanced" student would be likely (at least 2/3rds of the time) to get each rubric score (1, 2, 3, and 4). Although each domain is scored on a 1–4 point scale, the domains are differentially weighted when finding the total score. Essay score recommendations for each performance level are calculated by summing together the weighted rubric score across the two domains for each individual. Individual total cut score recommendations for each performance level are calculated by summing together the selected response score recommendations and short paper score recommendations for each panelist.

Panelists practice the Modified-Angoff rating methods using sample items and then engage in three rounds of ratings for the individual items. Cut scores are calculated by aggregating the

committee-level data for each performance level. The cut scores are provided to the panelists as feedback after each round of ratings. Panelists then engage in committee level discussions about how the individual items were rated as well as the resulting raw score cuts.

After the third and final round of ratings, panelists complete an evaluation of the standard setting process, their confidence in the results, and their satisfaction with the standard setting facilitator/facilities. These surveys are used to ensure that all panelists understood the process used to facilitate their recommendations and were confident in their recommendations and satisfied with the standard setting facilitator/facilities.

Next, a smaller group of committee members is convened to review the recommendations coming from Round 3. This articulation committee reviews the PLDs of all assessments included in a particular standard setting meeting (e.g., grades 3–8 reading) and discusses the results of Round 3 with regards to the estimated impact data. Impact data provides the percentage of students who would fall into each performance level if the Round 3 cut scores were adopted. The articulation committee can suggest modifications to the cut score recommendations if there is justification based on the content or the impact data. The final recommended cut scores are submitted to the Virginia Board of Education for approval.

Approved cut scores for all SOL tests can be found on the Virginia Department of Education website.

http://www.doe.virginia.gov/testing/scoring/sol-cut-scores.pdf

7.2 Standard Setting Schedule

Each subject area undergoes a review and potential revision to the content standards every seven years. After the content standards are revised, the performance standards must also be revised. Therefore, standard setting in Virginia also occurs on a seven-year cycle. There were no standard-setting events during the 2014–2015 test administration cycle. The table below provides years of the previous standard settings and future standard settings.

Table 7.2.1 Virginia SOL Standard Setting Schedule

Subjects	Previous Standard Setting	Upcoming Standard Setting
History	2010–2011	2017–2018
Mathematics	2011–2012	2018–2019
Reading and Writing	2012–2013	2019–2020
Science	2012–2013	2019–2020

8. CALIBRATION, SCALING, AND EQUATING

8.1 Rasch Partial Credit Model (RPCM)

For the Virginia SOL assessments, the unidimensional IRT Rasch Partial Credit Model (RPCM; Masters, 1982; Rasch, 1980) is used for item calibration. The RPCM model is commonly used for calibration, scaling, and equating in large assessment programs. The WINSTEPS software program (Linacre, 2006) is used to calibrate items, that is, to calculate item difficulty estimates, and is also used to calculate student proficiency estimates.

There are several benefits to using the RPCM model:

- The model can handle both dichotomous (e.g., MC/TEI) and polytomous (e.g., prompt) items.
- Unlike raw scores or percent correct values, estimates of item difficulty and student proficiency based on RPCM are not dependent upon the particular items included on a test form. Therefore, item difficulty and student proficiency values can be compared across forms and administrations.
- The underlying proficiency scale that is created using RPCM can be used to maintain performance levels across forms so that the interpretation of passing the SOL test means the same thing regardless of the particular form a student happens to receive.
- There is a one-to-one relationship between raw scores and Rasch item difficulty estimates so that it is possible to develop a raw score to theta table, where theta (θ) is the student proficiency value estimated using RPCM.
- Item difficulty and student proficiency (theta) are estimated simultaneously, and the values are on the same scale. This makes it possible to predict which items a student is likely to get right or wrong based on their proficiency.

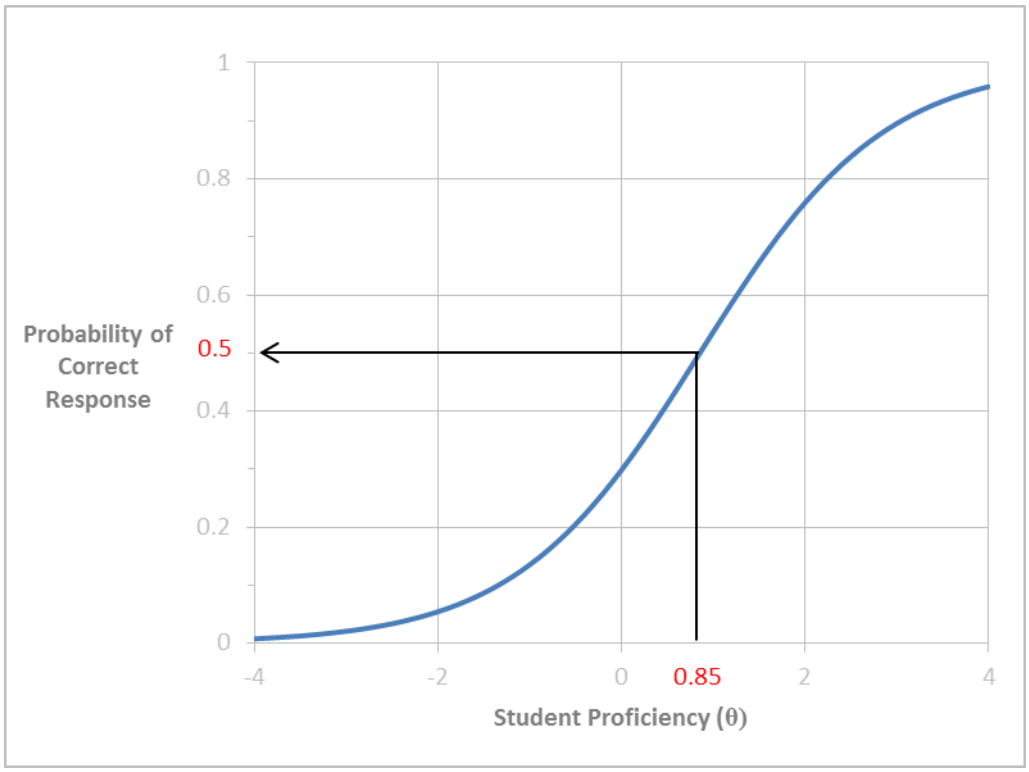
Masters (1982) provided a formula for calculating the probability of a student with proficiency θ obtaining a raw score of x on item/task i involving m_i+1 score categories with difficulties D_{ij}.

$$P_{xi} = \frac{e^{\sum_{j=0}^{x}(\theta - D_{ij})}}{\sum_{k=0}^{m_i}\left[e^{\sum_{j=0}^{k}(\theta - D_{ij})}\right]}, \quad x = 0, 1, ..., m_i \qquad (8.1)$$

For dichotomous items, the result of the equation 8.1 can be displayed as an item characteristic curve (ICC) where the vertical axis is the probability of a correct response, ranging from 0 to 1, and the horizontal axis is the student's proficiency estimate. The theta/proficiency scale is continuous and unbounded in theory. In practice, it typically ranges from -4 to +4 logits.

An example ICC is provided in Figure 8.1.1. The difficulty of an item corresponds to the location on the ICC where a student has a 0.5 probability of answering the item correctly. Because item difficulty and student proficiency are on the same scale, in the example ICC, the item difficulty is 0.85 and the student theta corresponding to a 0.5 probability of a correct response is also 0.85.

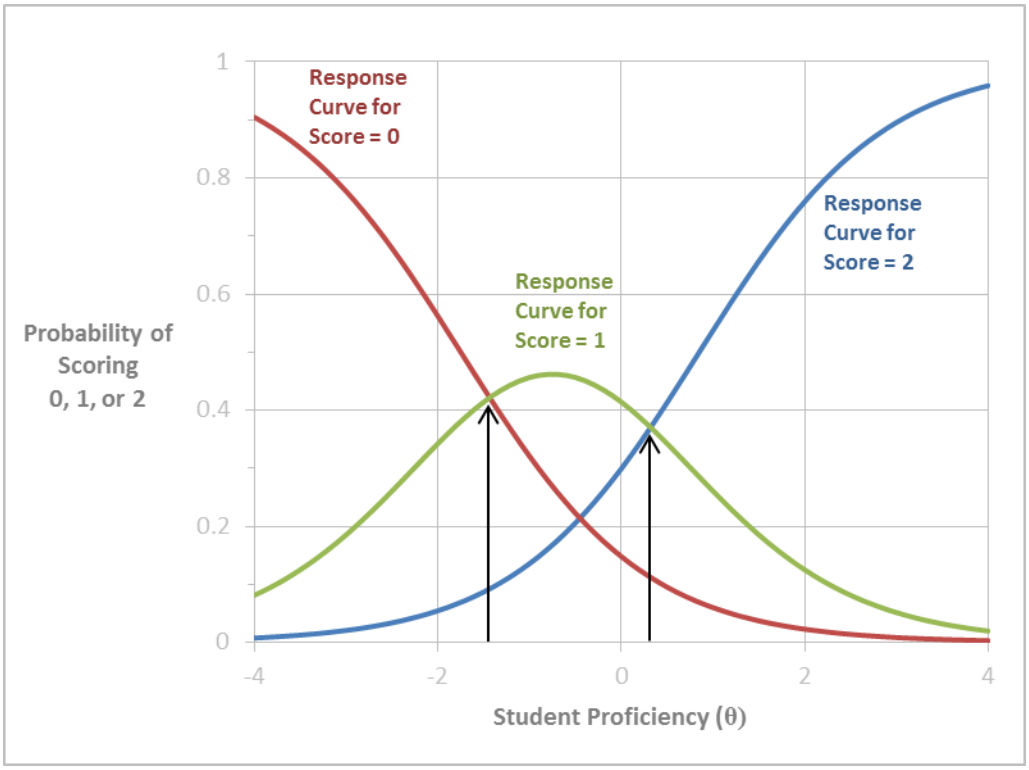
Figure 8.1.1 Sample Item Characteristic Curve for a Dichotomous Item



For polytomous items, a more complex graph is needed to display the probability of obtaining each of the possible score points based on student proficiency. For example, with an item with three score points (0, 1, 2), three response probability curves are needed. An example of the response curves for a polytomous item is provided in Figure 8.1.2. The left-most curve in Figure 8.1.2 represents the probability of students getting a score of 0 (completely incorrect) on the item given their proficiency. Students with very low proficiency (i.e., below –3) are likely to be in the score of 0 category and are more likely to be in this category than in the other two categories. Those receiving a score of 1 tend be moderately proficient. The final, right-most curve represents the probability for those receiving a score of 2 (completely correct). High-proficiency students are more likely to be in this category than in the other two categories, but there are still some of middle and low proficiency students that may obtain full credit on this item.

The arrows in Figure 8.1.2 indicate "thresholds" where the response curves cross. Students with proficiencies lower than the proficiency indicated by the left arrow have a higher probability of receiving a score of 0 compared to a score of 1. But students with proficiencies higher than the proficiency indicated by the left arrow have a higher probability of receiving a score of 1 compared to a score of 0. Likewise, the right arrow indicates the threshold between students who are more likely to obtain a score of 1 and those that are more likely to receive a score of 2.

Figure 8.1.2 Sample Category Response Curves for a Polytomous Item



8.2 Scaling

The raw score scale is commonly used in classrooms. It is intuitive to the public. However, it is not as useful as it appears for standardized assessments. This is because the goal is to be able to compare student scores regardless of the particular set of items they took and to be able to make inferences about their proficiency that is not form specific or only accurate for a particular administration. In order to be able to make the desired score comparisons, it is important to be able to place all of the items from all test forms onto a common scale. This can be done with the RPCM.

8.2.1 Scale Scores for the Total Test

The RPCM is used to create a single scale for item difficulty and student proficiency values. However, the resulting scale has negative numbers and numbers with decimals. These values are not intuitive to the public. To create a more user-friendly scale, a linear transformation is applied to the theta (proficiency) scale to obtain scale scores. On the Virginia SOL tests the scale scores range from 0 to 600, with higher scores indicating higher proficiency. To make score interpretation even easier, the cut scores on the theta scale, which are different for every grade and subject, as determined during the standard setting process, have been set to scale scores of 400 for proficient and 500 for advanced for all SOL tests.

To determine the appropriate linear transformation (see Wright & Stone, 1979) to get from the

theta scale to the 0–600 scale, theta is multiplied by a slope constant (γ) and then an intercept constant (α) is added to the result.

$$\textit{Scale Score} = \alpha + \gamma \cdot \textit{theta}, \tag{8.2}$$

where the intercept of the linear transformation is

$$\alpha = \left(D_1 d_2 - D_2 d_1\right) / \left(d_2 - d_1\right), \tag{8.3}$$

and the slope is

$$\gamma = \left(D_2 - D_1\right) / \left(d_2 - d_1\right). \tag{8.4}$$

Here $D_1 = 400$ (the proficient cut scale score), $D_2 = 500$ (the advanced cut scale score), d_1 is the theta value associated with the proficient cut score, and d_2 is the theta value associated with the advanced cut score. Again, D_1 and D_2 are constant across SOL tests, but d_1 and d_2 differ based on the specific standards for each test.[15]

Although scale scores are a linear transformation of the theta scale, they are a nonlinear transformation of the raw scores from which they were obtained. That is, the distance between scale scores does not remain the same for each change in the raw scores. Typically, for the middle of the scale (around the 350–400 range), the increments are smaller than they are near the top or bottom of the scale. Sometimes a scale score can be estimated as less than 0 (for very low raw scores), or greater than 600 (for very high raw scores). In these cases, to keep the range consistent, scale scores below 0 are set to 0 and scale scores above 600 are set to 600. In addition, at a raw score of 0, the scale score is always set to be 0, and at the top raw score (100% correct), the scale score is always set to 600.

8.2.2 Scale Scores for Reporting Categories

Each assessment covers a number of SOL, which are grouped into categories that address related content or skills. These categories are called *reporting categories*. Reporting category scale scores are calculated to provide an interpretation of student performance in each reporting category in relation to the performance standard on the test as a whole. For example, Rasch item difficulty estimates for items in reporting category 1 (obtained from the calibration and equating process conducted at the total test level) are used to create a raw-score-to-theta table specific to reporting category 1. Once the raw-score-to-theta table is produced, the following formula is used to create a scale for reporting category 1:

$$\text{Reporting Category 1 Scale Score} = 30 + 7 \times \left(\frac{\theta_X - \theta_{PROF}}{\sigma_R}\right), \tag{8.5}$$

where θ_X is the theta value associated with the reporting category raw score, θ_{PROF} is the theta value associated with the passing cut on the overall test (equivalent to d_1 in equations 8.3 and

[15] Although the *proficient* and *advanced* scale score cuts are constant across SOL tests, for the tests that also have a *basic* cut score, the basic scale score cuts vary. This is because with a linear transformation it is not possible to fix three points.

8.4), and σ_R is the standard deviation of the theta values associated with the reporting category. The standard deviations used in these calculations are derived from the first administration of a test after new performance standards are set. The reporting category scale scores range from 0 to 50. Sometimes a reporting category scale score can be estimated as less than 0 (for very low reporting category raw scores), or greater than 50 (for very high reporting category raw scores). In these cases, to keep the range consistent, reporting category scale scores below 0 are set to 0 and reporting category scale scores above 50 are set to 50. In addition, at a reporting category raw score of 0, the reporting category scale score is always set to be 0, and at the top reporting category raw score (100% correct), the reporting category scale score is always set to 50.

The same process and formula is used to create scale scores for the rest of the reporting categories on a test.

8.3 Equating

Equating is a statistical procedure that adjusts for slight differences in difficulty between test forms. Once the adjustment is applied, scores can be compared across forms and students taking one form of a test are neither advantaged nor disadvantaged when compared to students taking a different form of a test.

The equating design used for the SOL assessments is the common item non-equivalent group design (Kolen & Brennan, 2014). In this design, a set of common items (also called linking items or anchor items) that was administered on a previous form is also included on a new form. This set of items provides a mechanism for estimating differences in form difficulty while taking differences in student proficiency into consideration.

8.3.1 Field-Test Equating

As described in Section 8.1, the items on the SOL assessments are calibrated using the RPCM. Once the items are calibrated, the item difficulty statistics are obtained. Each time a set of items is calibrated using the Winsteps software (with item centered scaling), the average difficulty for the set is always estimated arbitrarily as zero. However, some item sets may be easier than others, so the item difficulty statistics must be transformed to be on the same scale. An initial set of items can be used to set the scale. Then, each time new sets of items are calibrated, they can be transformed so their item difficulties are on this initial or *base scale*. The process for testing out new items is called *field testing*, and the estimation of item difficulties and transformation to the base scale is called *field-test equating*.

In field-test equating, rather than using equating to estimate the difficulty of a test form, the goal is to estimate the difficulty of new items. These new items are included on forms with operational items that have already been field tested during previous administrations. The operational items serve as the common items for equating. These items were placed onto the base scale previously. Rasch values for the common (operational) items are fixed while Rasch values for the field-test items are freely calibrated. By fixing the common item Rasch values, the Winsteps software can estimate the field-test values on the same underlying base Rasch scale.

8.3.2 Pre-Equating

One of the benefits of IRT is that item difficulty and student proficiency (θ) are estimated simultaneously, with both estimates on the same scale. Using this relationship, a raw score to theta table can be generated using the following equation:

$$Raw\ Score = \sum_{i=1}^{I} \sum_{j=0}^{m_i} j \times P_{ij}(\theta), \quad (8.6)$$

where $P_{ij}(\theta)$ is the probability of a correct response for each of the $i = 1, ..., I$ items given that the item categories (j) are scored $0, ..., m_i$. The scaling constants described in Section 8.2 can be applied to the theta values and a raw score to scale score (RSSS) table can be produced.

Pre-equating makes use of this relationship by using existing item difficulty estimates, obtained during field testing or previous post-equating analyses, to generate a RSSS table for a test form prior to the test being administered. No transformation of item difficulties is needed in this case because previous equating analyses have already placed all of the item statistics on the same scale.

Pre-equating is completed for all new SOL forms as a way to evaluate the difficulty of the test and the position of the cut scores that differentiate the performance levels on the form prior to administration. Test difficulty values can be compared to established targets and used to determine whether modifications should be made to a potential test form during the test construction process.

Pre-equating is also used to provide scores for students who take a test administered using CAT. Each student takes a set of items tailored to his or her ability level. Therefore, many students receive a different set of items, and a single RSSS table will not suffice. Instead, each student's theta is calculated based on the specific set of items he or she responds to, and the scaling constants (see Section 8.2) are applied to generate the student's scale score.

Although pre-equating is sufficient for generating the RSSS table needed for reporting student scores, often post-equating is completed after the administration of traditional (non-adaptive) test forms in order to update the RSSS table, taking possible changes in item difficulty into account.

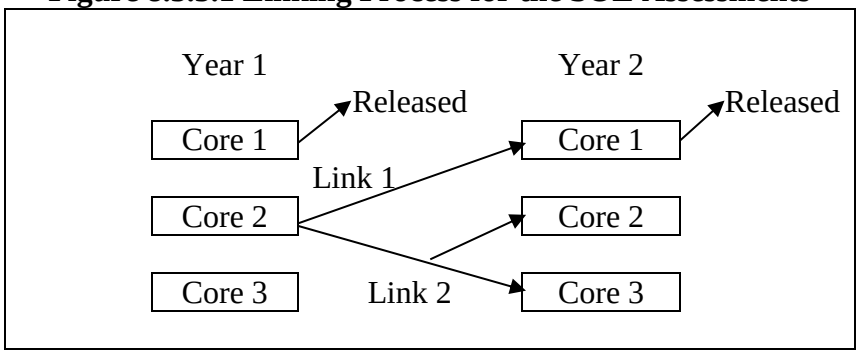
8.3.3 Post-Equating

Post-equating allows for new item difficulty estimates to be computed and used to generate scoring tables for reporting purposes. As with pre-equating, the goal of post-equating is to estimate form difficulty rather than item difficulty, and to generate a RSSS table. Typically, new SOL forms are post-equated.

In order to post-equate, a linking design is implemented where a set of common items is included on a new form that has already been administered on a previous form. The common item set is made up of 20–30% of the operational items and is representative of the total test in terms of content and statistical properties. Figure 8.3.3.1 provides a diagram of an example linking design. Because Core 1 is typically released to the public annually, items on Core 1 cannot be carried forward to future forms. Therefore, linking items from the first year that link forward to

the second year can only come from Cores 2 and 3. In this example, one set of linking sets comes from Core 2 (Year 1) and is included in Core 1 (Year 2); a second set of linking items comes from Core 2 (Year 1) and is included in both Core 2 (Year 2) and in Core 3 (Year 2). Therefore, Cores 2 and 3 (Year 2) share linking items. The linking design is similar for tests that have only two core forms, except Core 2 in Year 1 provides the linking items for both Core 1 and Core 2 in Year 2.

Figure 8.3.3.1 Linking Process for the SOL Assessments



The Rasch item difficulty values for the linking items are already on the base scale. To post-equate the non-linking items, the Rasch values for the linking items are anchored (or fixed) while the Rasch values for the non-linking items are freely estimated. The Rasch values estimated by Winsteps for the non-linking items are on the base scale because of the anchoring of the linking item Rasch values during the estimation process. The anchored Rasch values for the linking items, and the newly estimated Rasch values for the non-linking items are used to generate a RSSS table for the new core form.

Post-equating is conducted on a sample of the student population so that scores can be reported back to schools quickly. Once scores are available for at least 3,000 students on a core form, the post-equating process begins. Data are evaluated to make sure the sample will adequately support the post-equating process, and item-level statistics are created to double check scoring accuracy. Additionally, linking items are evaluated for stability. Linking items should have stable item difficulty estimates over time or they will not provide a stable link to the base scale. Therefore, if linking items have a displacement value greater than 0.5 (indicating poor stability), they are removed from the link set. Content representation of the linking set is evaluated if items are dropped. Post-equating is replicated by at least two psychometricians from the testing contractor and an additional replication is provided by an external verification group. Based on this process, a RSSS table is generated and verified for reporting.

9. RELIABILITY

Because tests contain a limited number of items given to students during a single administration, the score each student obtains is only an estimate of their true proficiency. Reliability is a way of quantifying the level of stability in test scores. High reliability indicates that there is a high degree of stability in the observed score representing the student's true proficiency level, and that if the student was retested on the same content, they would likely obtain a very similar score.

Conversely, low reliability indicates that there is a low degree of stability in the observed score and that if the student was tested again on the same content, they might obtain a very different score. Highly reliable scores are essential to making inferences about what students know and can do. Reliability estimates for new SOL core forms are provided in Part II of this report.

9.1 Internal Consistency

There are many ways to estimate reliability, but one way that is commonly used on large-scale assessments where students only receive one form is to compute internal consistency reliability. Internal consistency quantifies the stability of scores by estimating how consistently individuals respond to items. A basic estimate of internal consistency reliability is Cronbach's coefficient alpha statistic (Cronbach, 1951). Coefficient alpha is equivalent to the average correlation between scores based on all possible divisions of a test into two halves. Coefficient alpha can be used on any combinations of dichotomous (two score categories) and polytomous (three or more score categories) test items and is computed using the following formula:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{j=1}^{n} S_j^2}{S_X^2}\right), \quad (9.1)$$

where *n* is the number of items,

S_j^2 is the variance of students' scores on item *j*, and

S_X^2 is the variance of the total-test scores.

Cronbach's alpha ranges in value from 0.0 to 1.0, where higher values indicate greater reliability. Two factors affect estimates of internal consistency: test length and homogeneity of items. The longer the test, the higher the internal consistency value tends to be. Likewise, the more similar the items, the more likely examinees will respond consistently across items within the test, which also leads to higher reliability.

Stratified alpha (Cronbach, Schonemann, & McKie, 1965) provides a more appropriate reliability estimate for tests that are made up of homogeneous items clustered within heterogeneous components. For example, the SOL writing tests contain a combination of MC items paired with an open-ended writing prompt. Coefficient alpha for each component (MC and prompt) is included in calculating reliability for total test scores using the following stratified alpha formula:

$$\boldsymbol{Stratified\ \alpha = 1 - \frac{\sum \sigma_i^2 (1-\rho_{ii'})}{\sigma_T^2}}, \quad (9.2)$$

where $\boldsymbol{\sigma_i^2}$ = variance of scores on item type *i*,

σ_T^2 = variance of total scores, and

$\rho_{ii'}$ = reliability coefficient of scores on item type *i*.

For tests administered using CAT, coefficient alpha and stratified alpha cannot be applied because each student receives a unique test "form". However, it is still possible to calculate the

reliability of scores obtained from an adaptive test. Reliability for Grade 6 Mathematics was estimated using a method by Thissen (1990):

$$\bar{\rho} = \frac{\sigma_\theta^2 - average(CSEM_{true}^2)}{\sigma_\theta^2} \quad (9.3)$$

Where $\bar{\rho}$ is the average reliability, σ_θ^2 is the variance of the distribution of proficiency, and $CSEM_{true}^2$ is the squared true conditional standard error of measurement which is averaged across the proficiency distribution. To obtain these values, the number of students and the distribution of proficiency from the operational administration was used as inputs to the CAT simulation program. For more information about the CAT simulator, refer to the *Virginia Standards of Learning Computerized Adaptive Testing Technical Manual.*

9.2 Inter-Rater Reliability

For assessments that are scored by human raters, for example, the SOL writing test prompt component, the consistency with which raters provide scores is an important measure of reliability. Inter-rater reliability is calculated as the percentage of agreement between raters on the same student essay. Both perfect agreement and adjacent agreement are calculated. Perfect agreement occurs when the two independent scorers assign the same score point to the same piece of student work. Adjacent agreement occurs when the two independent scorers assign adjacent score points to the same piece of student work. The inter-rater reliability standards for the SOL writing assessments are 65% exact agreement and 96% exact plus adjacent agreement (see Section 5).

9.3 Measurement Error

9.3.1 Standard Error of Measurement

Whereas reliability coefficients provide an indication of the stability in test scores, measurement error quantifies the level of instability or uncertainty in test scores. The standard error of measurement (SEM) is inversely related to the reliability of a test; therefore, the greater the reliability, the lower the SEM. With a lower SEM, there is more confidence in the accuracy, or precision, of the observed test scores. The SEM is calculated using the following equation:

$$SEM = \sigma_X\sqrt{1 - \rho_{XX'}}, \quad (9.4)$$

where σ_X is the population standard deviation of observed scores, and

$\rho_{XX'}$ is the population reliability coefficient.

Using coefficient alpha, SEM can be estimated as follows:

$$Est(SEM) = S_X\sqrt{1 - \alpha}, \quad (9.5)$$

where S_X is the sample standard deviation of observed scores.

9.3.2 Conditional Standard Error of Measurement

The SEM provides a single measure of uncertainty in test scores. However, the uncertainty in test scores can vary depending upon the proficiency of the student (Andrich & Luo, 2004). The conditional standard error of measurement (CSEM) value based on IRT can vary across test scores. For example, if a person gets either a few or a large number of items correct (i.e., scores at the extremes of the score distribution), the CSEM will be greater in value than it will be if the person gets a moderate number of items correct. Scores near the middle of the score distribution typically have lower CSEM compared to the extremes because many tests are comprised of a large proportion of moderately difficult items which are suited to measuring students of moderate proficiency. By providing the CSEM, the error band quantifying uncertainty in a student's score is more precise than using the SEM.

Under the Rasch model, the CSEM for each person is as follows:

$$\sigma_{\hat{\theta}} = \frac{1}{\sqrt{\sum_{i=1}^{L} p_{vi}(1-p_{vi})}}, \tag{9.6}$$

where

v is for a particular person,

i is for a particular item,

L is the number of items on the test,

$\widehat{\theta}$ is the student proficiency estimate, and

p_{vi} is the probability that a person answers an item correctly and is defined as

$$p_{vi} = \frac{e^{\theta v-\delta_i}}{1+e^{\theta v-\delta_i}}, \tag{9.7}$$

where θ_v is person *v*'s proficiency and δ_i is item *i*'s difficulty.

An approximate 68% confidence interval for $\hat{\theta}$ is given by $\hat{\theta} \pm CSEM$. To interpret this confidence interval, consider a student who takes the test 100 times. Assuming measurement error is normally distributed, the student's true proficiency would fall within the confidence interval 68% of the time (or 68 times out of 100).

9.4 Decision Accuracy and Consistency

Students are classified into performance levels based on a comparison of their scale scores to the performance level cut scores. The most important cut score on the SOL assessments is the pass/proficient cut score. Students who obtain a scale score less than the pass/proficient cut score have failed the test whereas students who obtain a scale score greater than or equal to the pass/proficient cut score have passed the test. However, observed scores always have some degree of measurement error. This error may lead to misclassifications. A misclassification occurs when a proficient student fails a test (false negative) or when a non-proficient student

passes a test (false positive). Because of the decisions made based on these classifications, it is important to evaluate decision misclassifications to verify that they are kept to a minimum.

Decision accuracy is the extent to which the decision made based on students' scores on a particular test form (observed classifications) would agree with the decisions that would be made if each student were tested with all possible parallel forms of the assessments or an infinite number of independent administrations of the test (true classification). *Decision consistency* is the extent to which two observed classifications, based on taking two equally difficult forms of the test would classify students into the same performance category. For traditional SOL test forms, both decision accuracy and decision consistency are calculated using the Livingston and Lewis (1995) equations as implemented in the Manual for BBClass (Brennan, 2004). For SOL tests administered using CAT, decision accuracy is estimated using the Rudner (2005) approach:

$$P(Level_k) = \sum_{\theta=c}^{d}\left(\varphi\left(\frac{b-\theta}{SE(\theta)}\right) - \varphi\left(\frac{a-\theta}{SE(\theta)}\right)\right) f(\theta) \quad (9.8)$$

Where $P(Level_k)$ is the proportion of students with true scores in performance level *k*, bounded by cut scores *c* and *d*, who are classified into the performance level bounded by cut scores *a* and *b*. φ is the normal cumulative distribution function, and $f(\theta)$ is the density function associated with the true score. Classification consistency for CAT tests is estimated using the following equation from Lunz and Bergstrom (1991):

$$P_j(ConsistentDecision) = P(Fail|\hat{\theta}_j)^2 + P(Pass|\hat{\theta}_j)^2, \quad (9.9)$$

where $P\left(Pass\middle|\hat{\theta}_j\right) = \Phi\left(\frac{\hat{\theta}_j - \theta_c}{SE_{\hat{\theta}_j}}\right)$,

$P(Fail|\hat{\theta}_j) = 1 - P(Pass|\hat{\theta}_j)$,

θ_c is the cut score,

θ_j is the ability estimate of candidate *j*,

$SE_{\hat{\theta}_j}$ is the standard error of θ_j, and

Φ(Z) is the cumulative standard normal distribution evaluated at Z. Once these probabilities are obtained for each student, the probability that each student would obtain a consistent decision is calculated. The average of this probability over all students is the estimated decision consistency of the SOL assessment.

10. VALIDITY

As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." Sources of validity evidence are often clustered into the following categories:

- Evidence Based on Test Content

- Evidence Based on Response Processes
- Evidence Based on Internal Structure
- Evidence Based on Relationships to Other Variables

However, these sources of validity evidence are not considered independent pieces that must be equally represented and studied. Instead, these categories are used in this chapter to provide a framework for validating the interpretation(s) of scores that result from the SOL assessments.

> "These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use." (AERA, APA, & NCME, 2014)

10.1 Validity Evidence Based on Test Content

Content validity answers the question: "Does the test include items that measure all relevant aspects of the content while excluding irrelevant content?" Content validity is frequently defined in terms of the sampling adequacy of test items. That is, content validity is the extent to which the items in a test adequately represent the construct of interest (Suen, 1990). In educational testing, the state curriculum defines the content that is to be taught and assessed. Consequently, content validity provides judgmental evidence in support of the domain relevance and representativeness of the content in the test (Messick, 1989).

Section 3 of this technical report provides a good deal of information about the development of the SOL assessments including item development, prompt development, field testing of new items, and construction of new test forms. Virginia educators are included during each step of the test development cycle so that the SOL tests are very closely aligned with the SOL content standards from the beginning. The sections below provide a short summary of how this alignment provides content validity evidence for the SOL program.

10.1.1 Relation to Content Standards

Each Virginia SOL assessment is built to a specified test blueprint so that student scores reflect a consistent measure of the breadth of the commonwealth's content standards within each subject. This blueprint specifies the number of items to be used from each content strand within each reporting category. The content in the SOL test blueprint derives directly from the SOL curriculum framework.[16] The SOL curriculum framework amplifies the SOL and defines the content knowledge, skills, and understandings that are measured by the SOL tests. The curriculum framework provides additional guidance to school divisions and their teachers as they develop an instructional program appropriate for their students. It assists teachers as they plan

[16]The SOL curriculum frameworks and test blueprints may be accessed at the following website: http://www.doe.virginia.gov/testing/

their lessons by identifying essential content knowledge and skills students need to acquire. This supplemental framework delineates in greater specificity the minimum content that all teachers should teach and all students should learn. This direct relationship of the SOL curriculum frameworks with the SOL test blueprint and the SOL assessments lends support to the content validity of the SOL assessments.

10.1.2 Educator Input on Item Development

The Virginia Department of Education and its testing contractor are engaged in a continuous item and test development cycle. All SOL items go through several review rounds both internally and with Virginia educators (see Section 3). New SOL test forms are also rigorously reviewed and vetted with Virginia educators to ensure a tight alignment between the content the tests measure and the Virginia content standards that are to be assessed. These steps include item bias reviews where educators look for cases where an item may be measuring more than just the intended construct, especially where unintended constructs would disadvantage certain student subgroups.

10.1.3 SOL Assessment Alignment Studies

In addition to Virginia's rigorous item and test development process, the Department of Education contracted with the Virginia Commonwealth University to conduct external reviews of the alignment between the SOL tests and the SOL content standards. The alignment studies were conducted using procedures developed by Norman Webb (Webb, 2005) and focused on four different alignment criteria:

- Categorical Concurrence
- Depth of Knowledge Consistency
- Range of Knowledge Correspondence
- Balance of Representation

Using a large panel of trained alignment evaluators, results indicated that the assessments and the standards were well aligned based on the four criteria used to evaluate the alignment. A few discrepancies were identified including cases where test items required a lower depth of knowledge than was anticipated based on the SOL. For some assessments, the balance of representation was also not perfectly aligned. However, in most cases all four alignment criteria were met. Alignment studies were conducted for all of the SOL reading, mathematics, and science assessments at grades 3–8 and EOC. These results provide evidence that the test development process has been successful in building assessments that measure the knowledge and skills identified as important in the Virginia content standards.

10.2 Validity Evidence Based on Response Processes

Response processes are the cognitive strategies that students use to respond to items. For example, if a test is measuring reading comprehension, then it is reasonable to expect that the student would have to read a selection and respond to a question that requires understanding of the selection the student read. If the student read a passage about planets and was asked a

question that he or she knew the answer to because of prior learning in an astronomy class, then his or her response to the question may not accurately reflect reading comprehension skills. Instead, the item might be better characterized as measuring astronomy content knowledge.

10.2.1 Item Development

SOL reading assessments include both passage-based items and stand-alone items that measure the reading SOL. Passages contain a variety of lengths, topics, and genres. Some items are even tied to a pair of passages so that item responses are dependent upon integrating concepts across two passages. Passages are selected to be grade-level appropriate and engaging for students. These features of reading test development are intended to elicit student responses that accurately reflect proficiency in relation to the reading content standards.

Likewise, a variety of math items are developed. For many math tests some items are developed to measure math content knowledge in the context of calculator use, whereas other items are developed to measure math content knowledge where calculator use is not be permitted. Even among MC items, there is a wide range of math items, some of which require a simple computation, while others require problem solving and/or multi-step calculations.

History items contain excerpts from historical documents, images of historical figures, and maps of important locations. Likewise, science items contain diagrams, scenarios, and graphs, similar to what students see in the classroom.

Finally, the writing assessments include MC /TEI and prompt item types. The MC and TEI items allow for quick measures of skills like punctuation, grammar, and revision skills, while the prompt items require students to create a piece of writing. Without the prompt, it would be difficult to truly evaluate the important writing response process of students creating a piece of writing. For writing, the prompt scoring is also a critical part of the validity evidence. Without appropriate and accurate application of the prompt rubric, valid interpretations of student scores would not be possible. Detailed information about the professional scoring of students' prompt responses is provided in Section 5.

Significant planning goes into creating items that assess the breadth and depth of the content standards creatively with items that can be easily administered and scored under high-stakes conditions. Moreover, item development specifications provide guidance for developing items that measure the intended content, without measuring additional factors that are not what the test is intended to measure. For example, mathematics items are developed with English learners in mind. Unnecessarily complex language and double-meaning words are avoided.

10.2.2 Technology Enhanced Items (TEIs)

TEIs have been introduced into the SOL assessments over time with the first items field tested for math in spring 2011. TEIs were first field tested for reading, writing, and science in spring 2012. TEIs will be field tested for history in a future administration. These items allow students to demonstrate content knowledge in different ways than they have been able to with MC items alone. For example, students can plot points on a coordinate plane to provide a response to a

math item. They can select several components of a food web in science. They can identify relevant regions on a map for history. The Virginia Department of Education continues to field test these item types on SOL tests to measure content with greater fidelity and to a greater depth of knowledge. TEIs comprise approximately 15% of the SOL assessments and are an important component of Virginia's commitment to developing items that best measure the intended response processes.

10.3 Validity Evidence Based on Internal Structure

10.3.1 Internal Consistency

Section 9.1 detailed the statistics used to evaluate the internal consistency of the SOL assessments. These reliability statistics include coefficient alpha and stratified alpha. Reliability of the SOL assessments is evaluated each year for the new core forms. Part II of the technical report includes these statistics for the overall group and broken out by gender and the two largest ethnic subgroups in Virginia—African Americans and Caucasians. The reliability values for the SOL assessments are quite high, indicating highly reliable assessments. The non-writing subjects have values at or above 0.85 for the total group. Writing reliability values are 0.80 and above for the total group. Similar reliability values are found for each of the subgroups. These values provide evidence that the SOL assessments are reasonably homogenous measures of each content domain (e.g., reading, math), as expected.

10.3.2 Differential Item Functioning (DIF)

As described in Section 3.4, each new SOL item is field tested before it is used on an operational assessment. As part of the field-test process, statistics are generated to verify that the items are functioning as intended. DIF is evaluated for males and females and for African American and Caucasian students. Items that exceed customary thresholds are flagged for additional scrutiny. If the Virginia Department of Education or teacher reviewers identify a potential concern with the item, it is eliminated from the pool of items used to measure students' knowledge of the SOL. This process helps to ensure that the items on the Virginia SOL assessments are measuring the content standards and not measuring other, unintended constructs or disadvantaging particular student subgroups.

10.3.3 Unidimensionality Evaluation

During the first administration of a new SOL test, developed based on revised content standards, an exploratory factor analysis is conducted to verify that the assessment measures one primary factor. For example, the mathematics assessments should be measuring mathematics competence and not the combined effects of mathematics competence, reading and language proficiency, and the ability to work quickly. Often called the requirement for "test unidimensionality," evidence for the validity of the unifactor or unidimensionality assumption for an assessment can come from performing a factor analysis.

Data from student responses to test items were analyzed to look for relationships amongst the items and to identify the factor or factors the test items are measuring. An exploratory factor

analysis will frequently identify multiple factors—although not all factors are of the same strength relative to their prevalence in the data. It is therefore important to review the eigenvalues associated with each factor as these provide an indication of the relative strength of each factor. Reviewing the "eigenvalue plot" that is obtained from analyzing the data is one method for conducting this evaluation. However, Divgi's index (Divgi, 1980; similar to a method proposed by Lord, 1980) provides a simple evaluation when testing for unidimensionality (one factor). This index provides the ratio of the difference between the eigenvalues associated with the first and second factors to the difference between the eigenvalues associated with the second and third factors. A value that is greater than 3.0 implies that the test in question is characterized by a dominant first dimension.

In every case, the first eigenvalue was substantially larger than the second and the second and the third eigenvalues were of similar magnitude. In addition, the results showed that all values of Divgi's index exceeded 3.0. These results suggest that the Virginia SOL mathematics, reading, writing, and science assessments across all grades and core forms are characterized by a dominant primary dimension, indicating that they are measuring a dominant trait or main factor.

10.4 Validity Evidence Based on Relationships to Other Variables

Additional validity evidence can be provided by linking scores from the assessment of interest to scores from other assessments or measures (e.g., course grades). For example, linking a reading assessment to another measure of reading would be expected to show a stronger relationship than linking a reading assessment to a mathematics assessment. During the 2014–2015 academic year, the SOL assessment results were not linked empirically to other assessments or scores that would provide validity evidence for SOL scores. However, Virginia plans to carry out additional research in this area to support the validity of the SOL assessment program in the coming years.

11. ALTERNATE AND ALTERNATIVE ASSESSMENTS

The commonwealth of Virginia offers the following alternative and alternate assessments:

- The Virginia Grade Level Alternative (VGLA) Program
- The Virginia Substitute Education Program (VSEP)
- The Virginia Modified Achievement Standards Test (VMAST)
- The Virginia Alternate Assessment Program (VAAP)

The purpose of these assessments is to evaluate the performance of students who are unable to participate in the Virginia SOL statewide testing program, even with accommodations. Information about these assessments can be found in the *Virginia Alternative and Alternate Assessments Overview*.

12. RESOURCES

In addition to the information presented in this technical report (Part I and Part II), other resources are available that provide specific details on a variety of topics pertaining to the Virginia SOL assessments. These include testing implementation and examiner's manuals, Electronic Practice Assessment Tools (ePAT) applications, released versions of the SOL assessments, and practice items. These resources are available on the Virginia Department of Education website.

http://www.doe.virginia.gov/testing/test_administration/index.shtml

PART II: STATISTICAL SUMMARIES FOR SPRING 2015

1. OVERVIEW OF STATISTICAL SUMMARIES

This section contains an overview of the statistical summaries for the spring 2015 administration of the Virginia SOL and VMAST assessments. For more information about reliability, the CSEM, decision consistency, and decision accuracy, see Part I, Section 9, Reliability.

1.1 Administration Results

Three sets of tables are included in the Administration Results section. The first set shows the percentages of students who participated in online or paper-and-pencil administration in the spring 2015 administration. The second set shows the percentages of students in the proficient and advanced performance levels and the overall pass rate for each of the SOL and VMAST[17] assessments in the spring 2015 administration. The last set shows the scale score summary statistics for all newly constructed SOL assessments for the spring 2015 online administration.

1.2 Reliability Estimates for Non-Writing Assessments

Coefficient alpha reliability coefficients are provided overall, and by gender and ethnic subgroups (white and black) for all newly constructed core forms of the SOL assessments. Reliability was estimated using an alternative approach for scores obtained from the CAT administration of Grade 6 Mathematics (see Section 9 for more information). Reliability was not recalculated for reused forms.

1.3 Reliability Estimates for Writing Assessments

The reliability of all the writing assessments was estimated using stratified alpha. Stratified alpha is provided for each combination of MC core and writing prompt for the administrations in grade 8 and high school EOC. Analyses were done for the overall group and by subgroups based on gender and ethnicity (white and black).

Inter-rater reliability values are also provided for each prompt.

1.4 Decision Consistency and Accuracy Indices

Results are provided for the decision consistency and accuracy analyses for the 2015 new online SOL assessments. Although there is no general rule to determine the acceptable levels of decision accuracy and consistency needed for educational assessments, the Virginia SOL assessments have decision accuracy and consistency levels comparable to those that are reported in the Livingston and Lewis (1995) paper that describes the procedure. As expected, decision accuracy is generally higher than decision consistency.

[17] During 2014–2015, the VMAST forms were reused so only the online testing rates and the pass rates are included in the Administration Results section.

1.5 CAT Pool Information

There are several additional analyses that are relevant for tests administered using CAT. For each of the four CAT pools (SOL, plain English, SOL audio, and plain English audio), the following information is provided:

- The number of items overall and by reporting category
- The distribution of Rasch item difficulty values overall and by reporting category
- A frequency distribution of item exposure rates as well as the maximum observed exposure rate across all items
- Percent of the item pool used during an administration
- Rates for meeting the content and statistical targets
- Average CSEM values, CSEM values at each of the cut scores, and CSEM across the score scale

1.6 Raw Score-to-Scale Score Conversion Tables and Conditional SEMs

RSSS conversions tables are provided, which include the CSEM at each scale score level for all newly constructed cores (1, 2, and 3) of the non-writing assessments. For the writing assessments, these tables are provided for each combination of MC core and writing prompt.

2. SPRING 2015 STATISTICAL SUMMARY

2.1 Administration Results

2.1.1 Participation by Mode of Administration

The following tables show the number of tests administered in the online and paper modes of administration. Each table shows the grade and subject area of the test, the total number of valid tests administered, and the percentages of tests that were administered online and on paper.

Table 2.1.1.1 Percentage of Tests Taken by Mode: Grades 3–8

Grade	Subject	Total Number	Mode	
			Online (%)	Paper (%)
3	Reading	103,027	99	1
	Mathematics	104,958	99	1
4	Reading	101,640	99	1
	Mathematics	100,265	99	1
5	Reading	101,182	99	1
	Mathematics	94,847	99	1

Grade	Subject	Total Number	Mode	
			Online (%)	Paper (%)
	Science	102,171	99	1
6	Reading	101,021	99	1
	Mathematics	90,996	99*	1
7	Reading	100,217	99	1
	Mathematics	89,116	99	1
8	Reading	101,861	99	1
	Mathematics	71,756	99	1
	Science	90,980	99	1
	Writing	91,253	100	0

*The % of online testers for Grade 6 Mathematics came from CAT.

Table 2.1.1.2 Percentage of Tests Taken by Mode: Content-Specific History

Grade	Subject	Total Number	Mode	
			Online (%)	Paper (%)
CSH	Virginia Studies	95,595	99	1
	Civics and Economics	85,580	99	1

Table 2.1.1.3 Percentage of Tests Taken by Mode: End-of-Course

Grade	Subject	Total Number	Mode	
			Online (%)	Paper (%)
End-of-Course	Earth Science	80,032	99	1
	Biology	94,720	99	1
	Chemistry	61,366	100	0
	Algebra I	121,259	99	1
	Geometry	95,707	100	0
	Virginia & U.S. History	87,944	100	0
	World History I	79,551	99	1
	World History II	72,735	100	0
	World Geography	20,792	100	0
	English: Reading	86,595	100	0
	Algebra II	67,715	100	0
	Writing	107,285	99	1

Table 2.1.1.4 Percentage of Tests Taken by Mode: VMAST Mathematics and Reading

Grade	Subject	Total Number	Mode	
			Online (%)	Paper (%)
8	VMAST Reading	16	100	0
	VMAST Mathematics	26	100	0
End-of-Course	VMAST Reading	89	100	0
	VMAST Mathematics	133	100	0

2.1.2 Percent of Students in each Performance Level

The results in this section are based on all tests that produced a valid score. The tables below show the grade and subject area, the total number of tests taken, the percentages passing at the proficient and advanced performance levels, and the overall passing rate. Tests taken on paper and online are combined in the calculation of the passing rates.

Table 2.1.2.1 Grades 3–8 Passing Rates

Grade	Subject	N-Count	Proficiency Level		Overall Pass Rate (%)
			Proficient (%)	Advanced (%)	
3	Reading	103,027	50.0	17.2	67.2
	Mathematics	104,958	50.6	16.5	67.1
4	Reading	101,640	52.0	17.8	69.8
	Mathematics	100,265	51.3	26.2	77.5
5	Reading	101,182	51.1	21.1	72.3
	Mathematics	94,847	48.7	23.2	71.9
	Science	102,171	55.2	16.9	72.1
6	Reading	101,021	52.9	16.0	68.9
	Mathematics	90,996	61.3	13.7	74.9
7	Reading	100,217	56.2	18.0	74.2
	Mathematics	89,116	50.5	12.8	63.3
8	Reading	101,861	57.9	9.6	67.5
	Mathematics	71,756	55.7	8.0	63.7
	Science	90,980	61.3	10.3	71.6
	Writing	91,253	52.6	18.9	71.4

Table 2.1.2.2 Content-Specific History Passing Rates

Subject	N-Count	Proficiency Level		Overall Pass Rate (%)
		Proficient (%)	Advanced (%)	
Virginia Studies	95,595	41.5	41.2	82.7
Civics and Economics	85,580	52.1	29.9	81.9

Table 2.1.2.3 End-of-Course Passing Rates

Subject	N-Count	Proficiency Level		Overall Pass Rate (%)
		Proficient (%)	Advanced (%)	
Earth Science	80,032	62.4	7.3	69.7
Biology	94,720	60.3	12.1	72.4
Chemistry	61,366	65.7	14.0	79.7
Algebra I	121,259	58.5	7.6	66.1
Geometry	95,707	55.7	11.0	66.7
Virginia & U.S. History	87,944	64.7	12.6	77.3
World History I	79,551	58.2	17.4	75.6
World History II	72,735	59.3	17.3	76.6
World Geography	20,792	64.3	11.6	75.9
English: Reading	86,595	67.4	7.0	74.5
Algebra II	67,715	57.1	20.6	77.7
Writing	107,285	52.1	19.5	71.7

Table 2.1.2.4 VMAST Grade 8 & End-of-Course Passing Rates

Grade	Subject	N-Count	Proficiency Level		
			Proficient (%)	Advanced (%)	Overall Pass Rate (%)
8	VMAST Reading	16	31.3	0	31.3
	VMAST Mathematics	26	3.8	0	3.8
End-Of-Course	VMAST Reading	89	49.4	1.1	50.6
	VMAST Mathematics	133	19.5	0	19.5

2.1.3 Scale Score Summary Statistics

Tables 2.1.3.1 through 2.1.3.4 show the scale score summary statistics for each newly constructed Virginia SOL online assessment administered for the first time during the 2015 spring administration. Each table shows the grade and subject area of the test, number of examinees taking each test per core, as well as the number of test items, observed scale score mean, median, standard deviation, and minimum and maximum total scale scores.

Table 2.1.3.1 Summary Statistics for Grades 3–8 Reading Online

Subject	Grade	Core	Items	N	Mean	Med	SD	Min	Max
Reading	3	1	40	37,555	444.0	448	79.1	117	600
		2	40	41,988	433.7	430	70.5	115	600
	4	1	40	36,691	440.1	437	74.6	171	600
		2	40	42,774	437.8	438	65.8	126	600
	5	1	40	36,941	445.8	443	69.2	106	600
		2	40	42,036	445.9	444	71.2	114	600
	6	1	45	36,435	438.9	440	68.8	166	600
		2	45	41,681	436.9	439	64.1	124	600
	7	1	45	36,639	446.8	450	67.1	175	600
		2	45	41,355	443.3	440	58.7	209	600
	8	1	45	36,778	428.9	429	57.8	204	600
		2	45	39,003	432.3	432	64.4	207	600

Table 2.1.3.2 Summary Statistics for Grade 6 Mathematics Online

Subject	Grade	Items	N	Mean	Med	SD	Min	Max
Mathematics	6	45	86,644	441.0	440	54.6	233	600

Table 2.1.3.3 Summary Statistics for Grades 5 and 8 Science Online

Subject	Grade	Core	Items	N	Mean	Med	SD	Min	Max
Science	5	1	40	35,470	449.3	450	68.5	199	600
		2	40	42,258	439.6	435	67.0	139	600
	8	1	50	31,783	442.5	441	49.3	233	600
		2	50	38,456	435.2	430	57.9	197	600

Table 2.1.3.4 Summary Statistics for High School End-of-Course Online

Subject	Core	Items	N	Mean	Med	SD	Min	Max
Earth Science	1	50	33,563	436.3	437	46.7	267	600
	3	50	12,117	430.9	431	43.0	310	600
Biology	1	50	47,770	447.7	446	50.8	207	600
	3	50	13,590	445.5	444	53.9	268	600
Chemistry	1	50	34,285	450.8	448	49.4	261	600
	3	50	8,388	444.5	444	47.5	279	600
Algebra I	3	50	16,407	424.2	422	43.4	240	600
Geometry	3	50	13,042	433.1	427	52.5	248	600
World Geography	2	60	9,771	441.2	439	50.1	266	600
English: Reading	1	55	21,988	443.8	448	38.9	289	600
	2	55	21,226	446.5	446	46.5	287	600
	3	55	14,860	435.9	437	44.1	277	600
Algebra II	3	50	8,807	451.6	447	58.0	273	600

Table 2.1.3.5 shows the scale score summary statistics for grade 8 and EOC Virginia SOL writing tests taken during the spring 2015 administration. The table presents the number of examinees tested for every grade/core/prompt combination, as well as the observed scale score mean, median, standard deviation, and minimum and maximum values.

Table 2.1.3.5 Summary Statistics for Grades 8, and End-of-Course Writing Tests

Grade	Core	Prompt	N	Mean	Med	SD	Min	Max
8	1	2809	5,891	445.3	445	64.9	213	600
		2813	5,812	437.8	435	61.5	203	600
		2816	5,882	446.2	445	64.6	228	600
		2817	5,889	445.6	443	64.8	193	600
		2853	5,919	447.1	443	64.7	194	600
		2854	5,849	444.9	441	64.2	192	600
	2	2809	5,321	451.7	443	69.2	236	600
		2813	5,304	443.6	443	65.3	214	600
		2816	5,368	450.1	443	68.0	186	600

Grade	Core	Prompt	N	Mean	Med	SD	Min	Max
		2817	5,470	452.6	451	68.8	236	600
		2853	5,363	451.8	451	67.9	209	600
		2854	5,358	452.8	449	70.0	223	600
	3	2809	2,852	447.2	442	69.6	245	600
		2813	2,741	439.0	432	67.8	226	600
		2816	2,848	447.2	442	70.2	234	600
		2817	2,917	450.5	450	69.7	235	600
		2853	2,820	446.5	440	70.6	236	600
		2854	2,741	444.9	438	69.9	222	600
EOC	1	2101	4,960	457.2	457	60.0	252	600
		2103	4,965	457.9	454	62.9	257	600
		2106	4,882	456.2	451	62.8	238	600
		2108	4,813	456.3	457	61.2	237	600
		2116	4,956	458.9	452	62.5	240	600
		2124	4,962	458.8	456	62.6	266	600
	2	2101	4,659	465.1	461	63.4	262	600
		2103	4,671	466.4	466	65.1	248	600
		2106	4,626	467.4	464	64.8	224	600
		2108	4,697	468.9	462	65.4	237	600
		2116	4,668	468.2	465	64.1	207	600
		2124	4,721	468.6	461	64.4	237	600
	3	2101	2,539	456.7	453	61.4	264	600
		2103	2,570	457.1	450	64.4	250	600
		2106	2,558	458.0	455	64.4	240	600
		2108	2,570	458.7	453	62.9	250	600
		2116	2,513	457.5	449	62.4	228	600
		2124	2,585	458.6	453	65.1	259	600

2.2 Reliability Estimates for Multiple-Choice/Technology Enhanced Item Assessments

2.2.1 Overall Reliability Estimates

This section addresses the overall reliability estimates for each newly constructed SOL test administered for the first time in spring 2015. Each table shows the number of students used in the analyses and the associated Cronbach's Alpha reliability for each grade/core combination. For Grade 6 mathematics, reliability was calculated using results from the CAT simulator and a formula provided in Thissen (1990). In all cases, the reliability coefficients were above the desired lower limit of 0.80.

Table 2.2.1.1 Reliability for Grades 3–8 Reading

Subject	Grade	Core	Online	
			N	Alpha
Reading	3	1	37,555	0.90
		2	41,988	0.88
	4	1	36,691	0.89
		2	42,774	0.87
	5	1	36,941	0.86
		2	42,036	0.87
	6	1	36,435	0.90
		2	41,681	0.88
	7	1	36,639	0.90
		2	41,355	0.87
	8	1	36,778	0.88
		2	39,003	0.90

Table 2.2.1.2 Reliability for Grade 6 Mathematics

Subject	Grade	Online	
		N	Reliability
Mathematics	6	86,644	0.93

Note: An alternative reliability formula appropriate for use with CAT was used to estimate reliability for Grade 6 Mathematics.

Table 2.2.1.3 Reliability for Grades 5 and 8 Science

Subject	Grade	Core	Online	
			N	Alpha
Science	5	1	35,470	0.87

Subject	Grade	Core	Online	
			N	Alpha
		2	42,258	0.87
	8	1	31,783	0.89
		2	38,456	0.92

Table 2.2.1.4 Reliability for High School End-of-Course Tests

Subject	Core	Online	
		N	Alpha
Earth Science	1	33,563	0.90
	3	12,117	0.89
Biology	1	47,770	0.89
	3	13,590	0.91
Chemistry	1	34,285	0.89
	3	8,388	0.89
Algebra I	3	16,407	0.90
Geometry	3	13,042	0.91
World Geography	2	9,771	0.91
English: Reading/Lit. & Res.	1	21,988	0.84
	2	21,226	0.89
	3	14,860	0.88
Algebra II	3	8,807	0.88

2.2.2 Reliability Estimates by Gender

This section presents subgroup reliability results broken down by gender for newly constructed SOL tests administered for the first time in spring 2015. Each table shows the number of students used in the analyses and the associated reliability coefficients for each grade/core/gender combination. For all SOL tests, the reliability estimates are above the desired lower limit of 0.80. Students not reporting their gender were excluded from these results.

Table 2.2.2.1 Reliability for Grades 3–8 Reading by Gender

Subject	Grade	Core	Online			
			Female		Male	
			N	Alpha	N	Alpha
Reading	3	1	18,664	0.90	18,891	0.91
		2	20,737	0.87	21,251	0.88
	4	1	17,988	0.89	18,703	0.90
		2	20,815	0.87	21,959	0.87
	5	1	18,208	0.86	18,733	0.86
		2	20,554	0.87	21,482	0.88
	6	1	18,202	0.89	18,233	0.90
		2	20,377	0.88	21,304	0.89
	7	1	18,051	0.89	18,588	0.90
		2	20,159	0.86	21,196	0.87
	8	1	17,964	0.87	18,814	0.88
		2	19,305	0.89	19,698	0.90

Table 2.2.2.2 Reliability for Grade 6 Mathematics by Gender

Subject	Grade	Online			
		Female		Male	
		N	Reliability	N	Reliability
Mathematics	6	42,923	0.92	43,721	0.93

Note: An alternative reliability formula appropriate for use with computer adaptive testing was used to estimate reliability for Grade 6 Mathematics.

Table 2.2.2.3 Reliability for Grades 5 and 8 Science by Gender

Subject	Grade	Core	Online			
			Female		Male	
			N	Alpha	N	Alpha
Science	5	1	17,643	0.86	17,827	0.87
		2	20,940	0.86	21,318	0.88
	8	1	15,761	0.88	16,022	0.89
		2	19,155	0.91	19,301	0.92

Table 2.2.2.4 Reliability for High School End-of-Course Tests by Gender

Subject	Core	Online			
		Female		Male	
		N	Alpha	N	Alpha
Earth Science	1	16,472	0.89	17,091	0.90
	3	5,784	0.89	6,333	0.89
Biology	1	23,801	0.89	23,969	0.90
	3	6,605	0.91	6,985	0.91
Chemistry	1	17,858	0.89	16,427	0.89
	3	4,498	0.88	3,890	0.89
Algebra I	3	8,107	0.90	8,300	0.90
Geometry	3	6,578	0.91	6,464	0.91
World Geography	2	4,946	0.91	4,825	0.91
English: Reading/Lit. & Res.	1	10,962	0.83	11,026	0.86
	2	10,382	0.88	10,844	0.89
	3	7,411	0.87	7,449	0.88
Algebra II	3	4,626	0.88	4,181	0.89

2.2.3 Reliability Estimates by Ethnic Group

Reliability results are broken down by ethnic group for newly constructed SOL tests administered for the first time in spring 2015. In Virginia, only the white and black ethnic groups have a large enough population to calculate reliability statistics. Each table shows the number of students used in the analyses and the associated reliability coefficients for each grade/core/ethnic combination. In all instances, the reliability estimates were above the desired lower limit of 0.80. Students not reporting their ethnicity are excluded from these results.

Table 2.2.3.1 Reliability for Grades 3–8 Reading by Ethnic Group

Subject	Grade	Core	Online			
			Black		White	
			N	Alpha	N	Alpha
Reading	3	1	10,282	0.90	18,051	0.89
		2	9,161	0.87	19,018	0.87
	4	1	9,893	0.88	17,924	0.88
		2	9,409	0.85	19,271	0.85
	5	1	9,847	0.85	18,450	0.85
		2	8,856	0.85	19,341	0.86
	6	1	9,955	0.89	18,039	0.88
		2	9,114	0.86	19,263	0.87

Subject	Grade	Core	Online			
			Black		White	
			N	Alpha	N	Alpha
	7	1	10,255	0.89	18,312	0.87
		2	9,043	0.85	19,105	0.86
	8	1	10,465	0.87	18,600	0.85
		2	8,150	0.88	18,549	0.89

Table 2.2.3.2 Reliability for Grade 6 Mathematics by Ethnic Group

Subject	Grade	Online			
		Black		White	
		N	Reliability	N	Reliability
Mathematics	6	21,860	0.92	43,270	0.92

Note: An alternative reliability formula appropriate for use with CAT was used to estimate reliability for Grade 6 Mathematics.

Table 2.2.3.3 Reliability for Grades 5 and 8 Science by Ethnic Group

Subject	Grade	Core	Online			
			Black		White	
			N	Alpha	N	Alpha
Science	5	1	9,511	0.85	17,888	0.83
		2	8,772	0.85	19,212	0.85
	8	1	8,742	0.87	16,469	0.86
		2	8,152	0.88	17,312	0.91

Table 2.2.3.4 Reliability for High School End-of-Course Tests by Ethnic Group

Subject	Core	Online			
		Black		White	
		N	Alpha	N	Alpha
Earth Science	1	10,167	0.87	15,974	0.88
	3	2,803	0.87	6,099	0.88
Biology	1	11,996	0.87	22,730	0.87
	3	3,056	0.89	6,584	0.89
Chemistry	1	6,126	0.87	17,949	0.87
	3	1,465	0.88	4,767	0.88
Algebra I	3	4,772	0.87	9,103	0.90
Geometry	3	3,611	0.87	7,418	0.91

Subject	Core	Online			
		Black		White	
		N	Alpha	N	Alpha
World Geography	2	2,577	0.90	5,268	0.90
English: Reading/Lit. & Res.	1	4,573	0.83	11,588	0.82
	2	5,193	0.88	9,455	0.87
	3	4,031	0.86	8,287	0.87
Algebra II	3	1,938	0.85	5,490	0.88

2.3 Reliability Estimates for Writing Assessments

2.3.1 Stratified Alpha

The tables below present Stratified Alpha reliability estimates in writing for grade 8 and EOC. Results are provided for the overall sample, as well as by gender and ethnic group, for each combination of MC/TEI core and writing prompt.

For grade 8, Stratified Alpha ranged from 0.84 to 0.87. For EOC, Stratified Alpha ranged from 0.87 to 0.88.

Table 2.3.1.1 Reliability for Grade 8 and End-of-Course Writing

Grade	Core	Prompt	Online	
			N	Alpha
Writing 8	1	2809	5,891	0.85
		2813	5,812	0.85
		2816	5,882	0.85
		2817	5,889	0.85
		2853	5,919	0.85
		2854	5,849	0.84
	2	2809	5,321	0.86
		2813	5,304	0.86
		2816	5,368	0.86
		2817	5,470	0.86
		2853	5,363	0.85
		2854	5,358	0.86
	3	2809	2,852	0.85
		2813	2,741	0.87
		2816	2,848	0.86
		2817	2,917	0.86
		2853	2,820	0.86

Grade	Core	Prompt	Online	
			N	Alpha
		2854	2,741	0.86
Writing End-of-Course	1	2101	4,960	0.87
		2103	4,965	0.87
		2106	4,882	0.88
		2108	4,813	0.87
		2116	4,956	0.87
		2124	4,962	0.87
	2	2101	4,659	0.88
		2103	4,671	0.88
		2106	4,626	0.88
		2108	4,697	0.87
		2116	4,668	0.87
		2124	4,721	0.88
	3	2101	2,539	0.87
		2103	2,570	0.88
		2106	2,558	0.88
		2108	2,570	0.87
		2116	2,513	0.87
		2124	2,585	0.88

Table 2.3.1.2 shows the reliability results for grade 8 and EOC writing broken down by gender for each combination of core and writing prompt. For grade 8, Stratified Alpha ranged from 0.82 to 0.86 for females and from 0.85 to 0.87 for males. For EOC, Stratified Alpha ranged from 0.85 to 0.88 for females and from 0.87 to 0.89 for males.

Table 2.3.1.2 Reliability for Grade 8 and End-of-Course Writing by Gender

Grade	Core	Prompt	Online			
			Female		Male	
			N	Alpha	N	Alpha
Writing 8	1	2809	2,974	0.84	2,917	0.85
		2813	2,854	0.83	2,958	0.85
		2816	2,981	0.84	2,901	0.85
		2817	2,948	0.83	2,941	0.86
		2853	2,912	0.84	3,007	0.85
		2854	2,881	0.82	2,968	0.85
	2	2809	2,697	0.84	2,624	0.86
		2813	2,685	0.85	2,619	0.86

Grade	Core	Prompt	Online			
			Female		Male	
			N	Alpha	N	Alpha
		2816	2,706	0.85	2,662	0.86
		2817	2,711	0.84	2,759	0.87
		2853	2,660	0.84	2,703	0.86
		2854	2,718	0.85	2,640	0.86
	3	2809	1,411	0.84	1,441	0.86
		2813	1,349	0.86	1,392	0.87
		2816	1,444	0.85	1,404	0.86
		2817	1,419	0.85	1,498	0.86
		2853	1,413	0.85	1,407	0.86
		2854	1,317	0.85	1,424	0.86
Writing End-of-Course	1	2101	2,402	0.85	2,558	0.87
		2103	2,476	0.87	2,489	0.87
		2106	2,401	0.87	2,481	0.88
		2108	2,384	0.86	2,429	0.87
		2116	2,391	0.86	2,565	0.87
		2124	2,462	0.86	2,500	0.88
	2	2101	2,317	0.87	2,342	0.88
		2103	2,327	0.88	2,344	0.88
		2106	2,316	0.87	2,310	0.88
		2108	2,354	0.87	2,343	0.87
		2116	2,336	0.86	2,332	0.88
		2124	2,364	0.87	2,357	0.88
	3	2101	1,274	0.86	1,265	0.87
		2103	1,274	0.87	1,296	0.88
		2106	1,257	0.88	1,301	0.88
		2108	1,259	0.86	1,311	0.87
		2116	1,250	0.86	1,263	0.87
		2124	1,308	0.87	1,277	0.89

Table 2.3.1.3 shows the reliability results for grade 8 and EOC writing broken down by ethnic group for each combination of MC/TEI core and writing prompt. For grade 8, Stratified Alpha ranged from 0.82 to 0.85 for black students and from 0.83 to 0.85 for white students. For EOC, Stratified Alpha ranged from 0.83 to 0.87 for black students and from 0.85 to 0.88 for white students.

Table 2.3.1.3 Reliability for Grade 8 and End-of-Course Writing by Ethnic Group

Grade	Core	Prompt	Online			
			Black		White	
			N	Alpha	N	Alpha
Writing 8	1	2809	1,176	0.85	3,652	0.83
		2813	1,228	0.83	3,534	0.83
		2816	1,157	0.83	3,642	0.84
		2817	1,167	0.85	3,652	0.84
		2853	1,180	0.84	3,730	0.83
		2854	1,176	0.82	3,622	0.84
	2	2809	1,312	0.85	2,395	0.84
		2813	1,292	0.85	2,463	0.84
		2816	1,349	0.84	2,450	0.85
		2817	1,326	0.84	2,520	0.85
		2853	1,280	0.84	2,512	0.84
		2854	1,326	0.84	2,454	0.84
	3	2809	704	0.83	1,573	0.85
		2813	625	0.85	1,583	0.85
		2816	645	0.85	1,611	0.85
		2817	708	0.85	1,632	0.84
		2853	691	0.84	1,545	0.85
		2854	661	0.83	1,538	0.85
Writing End-of-Course	1	2101	922	0.85	3,160	0.86
		2103	971	0.87	3,136	0.86
		2106	941	0.87	3,155	0.87
		2108	929	0.85	3,079	0.86
		2116	905	0.86	3,227	0.86
		2124	923	0.86	3,184	0.87
	2	2101	1,124	0.85	2,186	0.86
		2103	1,134	0.85	2,238	0.87
		2106	1,095	0.85	2,173	0.86
		2108	1,163	0.85	2,121	0.85
		2116	1,166	0.84	2,133	0.86
		2124	1,151	0.86	2,139	0.86
	3	2101	685	0.83	1,403	0.86
		2103	709	0.84	1,354	0.86
		2106	686	0.86	1,383	0.88
		2108	686	0.83	1,385	0.85
		2116	692	0.84	1,312	0.85
		2124	667	0.86	1,395	0.87

2.3.2 Inter-Rater Reliability

There were a total of 6 writing prompts administered during the spring 2015 administration in grade 8 and EOC writing. The following tables provide the inter-rater reliability for each prompt/trait combination. Across all grades, the perfect agreement rate was above 62% and the perfect plus adjacent agreement rate was 99% or higher.

Table 2.3.2.1 Inter-Rater Reliability for the Grade 8 Writing Assessment: Prompts 2809, 2813, 2816, 2817, 2853, and 2854

Prompt	Trait	Online			
		N	Perfect Agree (%)	Adjacent (%)	Non-Adjacent (%)
2809	Comp/Written Expression	15,415	72	28	0
	Usage and Mechanics	15,415	65	34	1
2813	Comp/Written Expression	15,196	72	28	0
	Usage and Mechanics	15,196	67	32	0
2816	Comp/Written Expression	15,465	73	27	0
	Usage and Mechanics	15,465	66	34	1
2817	Comp/Written Expression	15,627	72	27	0
	Usage and Mechanics	15,627	67	32	1
2853	Comp/Written Expression	15,463	75	25	0
	Usage and Mechanics	15,463	68	31	0
2854	Comp/Written Expression	15,318	72	28	0
	Usage and Mechanics	15,318	69	31	0

Table 2.3.2.2 Inter-Rater Reliability for the End-of-Course Writing Assessment: Prompts 2101, 2103, 2106, 2108, 2116, and 2124

Prompt	Trait	Online			
		N	Perfect Agree (%)	Adjacent (%)	Non-Adjacent (%)
2101	Comp/Written Expression	14,999	75	25	0
	Usage and Mechanics	14,999	62	36	1
2103	Comp/Written Expression	13,708	72	28	0
	Usage and Mechanics	13,708	66	34	0
2106	Comp/Written Expression	13,561	73	27	0
	Usage and Mechanics	13,561	64	36	1
2108	Comp/Written Expression	13,507	77	23	0
	Usage and Mechanics	13,507	64	35	1
2116	Comp/Written Expression	13,616	77	23	0

Prompt	Trait	Online			
		N	Perfect Agree (%)	Adjacent (%)	Non-Adjacent (%)
	Usage and Mechanics	13,616	66	34	0
2124	Comp/Written Expression	13,725	75	25	0
	Usage and Mechanics	13,725	65	34	1

2.4 Decision Consistency and Accuracy Indices

The number of examinees taking each newly constructed Virginia SOL assessment administered for the first time during the spring 2015 administration and the proportion of accurate classifications, false positives, false negatives, and consistent classifications are provided in this section. Classification indices were computed using the cut score that separates passing scores from failing scores.

Decision accuracy was at or above 91% and decision consistency was at or above 87% for all of the non-writing SOL tests. Decision accuracy was at or above 90% and decision consistency was at or above 86% for all of the combinations of MC and prompts for writing SOL tests.

Table 2.4.1 Decision Consistency and Accuracy Indices for Grades 3–8 Reading Online Tests

Subject	Grade	Core	N	Accuracy	False Positive	False Negative	Consistency
Reading	3	1	37,555	0.92	0.04	0.04	0.89
		2	41,988	0.91	0.04	0.05	0.87
	4	1	36,691	0.92	0.04	0.04	0.88
		2	42,774	0.92	0.04	0.04	0.88
	5	1	36,941	0.92	0.04	0.05	0.88
		2	42,036	0.91	0.04	0.05	0.88
	6	1	36,435	0.92	0.04	0.04	0.89
		2	41,681	0.92	0.04	0.04	0.88
	7	1	36,639	0.93	0.03	0.04	0.90
		2	41,355	0.92	0.03	0.04	0.89
	8	1	36,778	0.92	0.04	0.04	0.88
		2	39,003	0.92	0.04	0.04	0.89

Table 2.4.2 Decision Consistency and Accuracy Indices for Grade 6 Mathematics Online Test

Subject	Grade	N	Accuracy	False Positive	False Negative	Consistency
Mathematics	6	86,644	0.94	0.03	0.03	0.92

Table 2.4.3 Decision Consistency and Accuracy Indices for Grades 5 and 8 Science Online Tests

Subject	Grade	Core	N	Accuracy	False Positive	False Negative	Consistency
Science	5	1	35,470	0.92	0.04	0.04	0.89
		2	42,258	0.91	0.04	0.05	0.87
	8	1	31,783	0.93	0.03	0.04	0.91
		2	38,456	0.93	0.04	0.04	0.90

Table 2.4.4 Decision Consistency and Accuracy Indices for High School End-of-Course Online Tests

Subject	Core	N	Accuracy	False Positive	False Negative	Consistency
Earth Science	1	33,563	0.93	0.03	0.04	0.90
	3	12,117	0.93	0.03	0.04	0.90
Biology	1	47,770	0.94	0.03	0.03	0.91
	3	13,590	0.94	0.03	0.04	0.91
Chemistry	1	34,285	0.94	0.03	0.03	0.92
	3	8,388	0.94	0.03	0.04	0.91
Algebra I	3	16,407	0.91	0.04	0.04	0.88
Geometry	3	13,042	0.92	0.04	0.04	0.89
World Geography	2	9,771	0.94	0.03	0.03	0.91
English: Reading/Lit. & Res.	1	21,988	0.95	0.02	0.03	0.93
	2	21,226	0.95	0.02	0.03	0.92
	3	14,860	0.93	0.03	0.04	0.90
Algebra II	3	8,807	0.92	0.03	0.04	0.89

Table 2.4.5 Decision Consistency and Accuracy Indices for Grade 8 and End-of-Course Writing Tests

Grade	Core	Prompt	N	Accuracy	False Positive	False Negative	Consistency
8	1	2809	5,891	0.91	0.04	0.05	0.87
		2813	5,812	0.90	0.05	0.05	0.86
		2816	5,882	0.92	0.04	0.05	0.88
		2817	5,889	0.91	0.04	0.05	0.87
		2853	5,919	0.91	0.05	0.05	0.87
		2854	5,849	0.90	0.05	0.05	0.86
	2	2809	5,321	0.91	0.04	0.05	0.87
		2813	5,304	0.90	0.05	0.05	0.87
		2816	5,368	0.91	0.04	0.05	0.87
		2817	5,470	0.91	0.04	0.05	0.87
		2853	5,363	0.91	0.04	0.05	0.87
		2854	5,358	0.91	0.04	0.05	0.88
	3	2809	2,852	0.90	0.05	0.05	0.86
		2813	2,741	0.90	0.05	0.05	0.86
		2816	2,848	0.91	0.04	0.05	0.87
		2817	2,917	0.91	0.04	0.05	0.87
		2853	2,820	0.90	0.05	0.05	0.86
		2854	2,741	0.90	0.05	0.05	0.86
End-of-Course	1	2101	4,960	0.93	0.03	0.04	0.90
		2103	4,965	0.93	0.03	0.04	0.90
		2106	4,882	0.93	0.03	0.04	0.91
		2108	4,813	0.93	0.03	0.04	0.90
		2116	4,956	0.93	0.03	0.04	0.90
		2124	4,962	0.93	0.03	0.04	0.90
	2	2101	4,659	0.94	0.03	0.03	0.91
		2103	4,671	0.94	0.03	0.03	0.91
		2106	4,626	0.94	0.03	0.04	0.91
		2108	4,697	0.94	0.03	0.03	0.91
		2116	4,668	0.94	0.03	0.03	0.91
		2124	4,721	0.94	0.03	0.03	0.92
	3	2101	2,539	0.93	0.03	0.04	0.90
		2103	2,570	0.93	0.03	0.04	0.89
		2106	2,558	0.93	0.03	0.04	0.90
		2108	2,570	0.92	0.03	0.04	0.89
		2116	2,513	0.93	0.03	0.04	0.90
		2124	2,585	0.93	0.03	0.04	0.90

2.5 CAT Pool Information

2.5.1 Number of Items in each CAT Pool Overall and by Reporting Category

Table 2.5.1.1 provides a breakdown of the size of the item pool—which varies across the four Virginia CAT pools. It also shows how the items are distributed across reporting categories. The SOL and plain English pools are much larger than the SOL audio and plain English audio pools by design. A much larger number of students take the SOL and Plain English tests, so larger pools are needed.

Table 2.5.1.1 Number of Items in the Mathematics CAT Pools Overall and by Reporting Category

		SOL Pool		Plain English Pool		SOL Audio Pool		Plain English Audio Pool	
Grade	Reporting Category	N Items	% Pool	N Items	% Pool	N Items	% Pool	N Items	% Pool
6	1	165	19%	162	20%	60	21%	58	21%
	2	154	18%	125	16%	53	18%	47	17%
	3	245	28%	232	29%	76	26%	73	27%
	4	306	35%	284	35%	98	34%	92	34%
	Overall	870	100%	803	100%	287	100%	270	100%

2.5.2 Rasch Item Difficulty Distribution Overall and by Reporting Category

Tables 2.5.2.1 through 2.5.2.4 provide information about the distribution of item difficulty in the four CAT pools, both for the total pool, and for each of the reporting categories. Figures 2.5.2.1 through 2.5.2.4 provide frequency distributions of Rasch values for each of the CAT item pools.

Table 2.5.2.1 Rasch Item Difficulty for Items in the Mathematics CAT SOL Pool Overall and by Reporting Category

Grade	Reporting Category	Mean	Standard Deviation	Minimum	Maximum
6	1	-0.292	1.332	-3.112	2.874
	2	0.179	0.951	-2.557	2.718
	3	0.025	1.193	-3.375	4.654
	4	0.180	1.282	-3.788	3.807
	Overall	0.046	1.226	-3.788	4.654

Table 2.5.2.2 Rasch Item Difficulty for Items in the Mathematics CAT Plain English Pool Overall and by Reporting Category

Grade	Reporting Category	Mean	Standard Deviation	Minimum	Maximum
6	1	-0.316	1.331	-3.112	2.874
	2	0.250	0.923	-2.331	2.718
	3	0.007	1.191	-3.375	4.654
	4	0.145	1.288	-3.788	3.807
	Overall	0.029	1.232	-3.788	4.654

Table 2.5.2.3 Rasch Item Difficulty for Items in the Mathematics CAT SOL Audio Pool Overall and by Reporting Category

Grade	Reporting Category	Mean	Standard Deviation	Minimum	Maximum
6	1	0.022	1.201	-2.739	2.874
	2	0.090	0.950	-2.557	1.895
	3	0.165	0.974	-1.881	2.786
	4	0.389	1.111	-2.337	3.551
	Overall	0.197	1.073	-2.739	3.551

Table 2.5.2.4 Rasch Item Difficulty for Items in the Mathematics CAT Plain English Audio Pool Overall and by Reporting Category

Grade	Reporting Category	Mean	Standard Deviation	Minimum	Maximum
6	1	-0.011	1.204	-2.739	2.874
	2	0.236	0.838	-1.513	1.895
	3	0.144	0.986	-1.881	2.786
	4	0.377	1.116	-2.337	3.551
	Overall	0.206	1.063	-2.739	3.551

Figure 2.5.2.1 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT SOL Pool

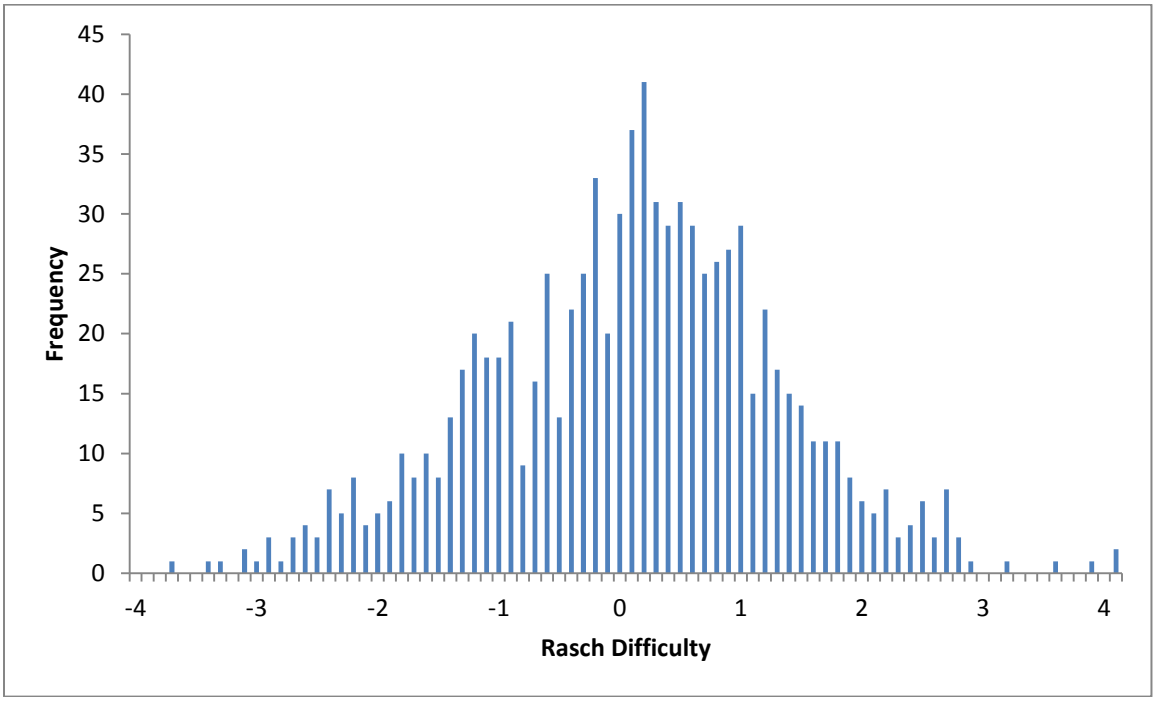


Figure 2.5.2.2 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Plain English Pool

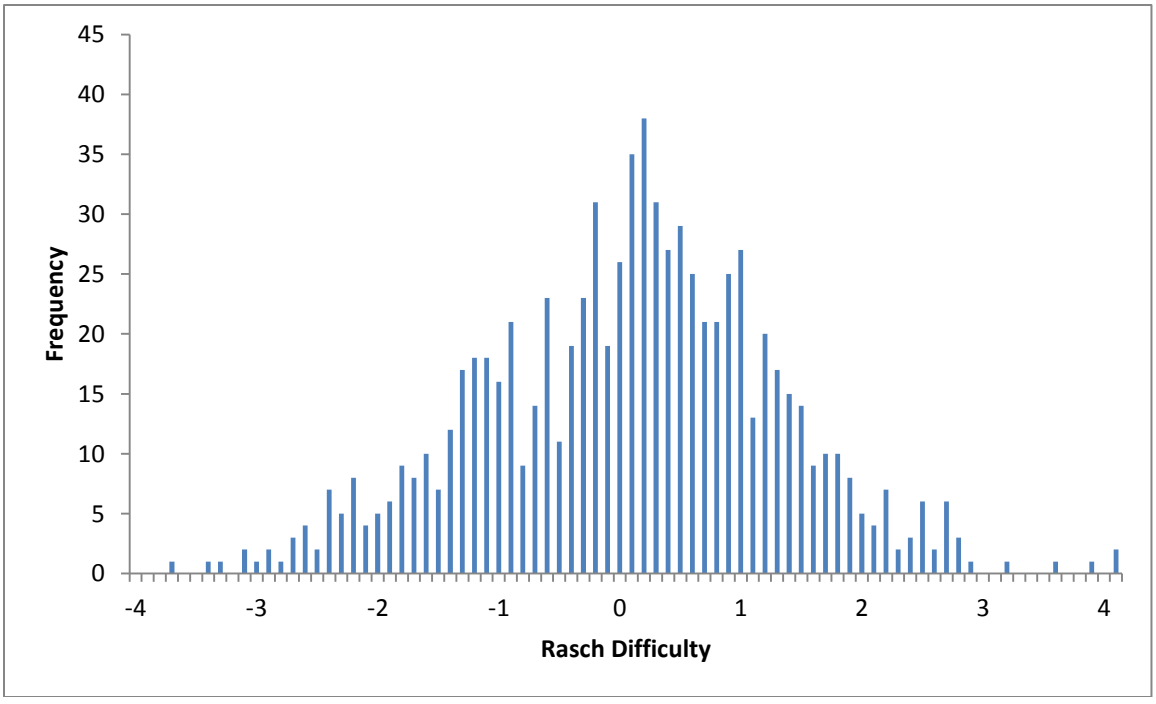


Figure 2.5.2.3 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Audio Pool

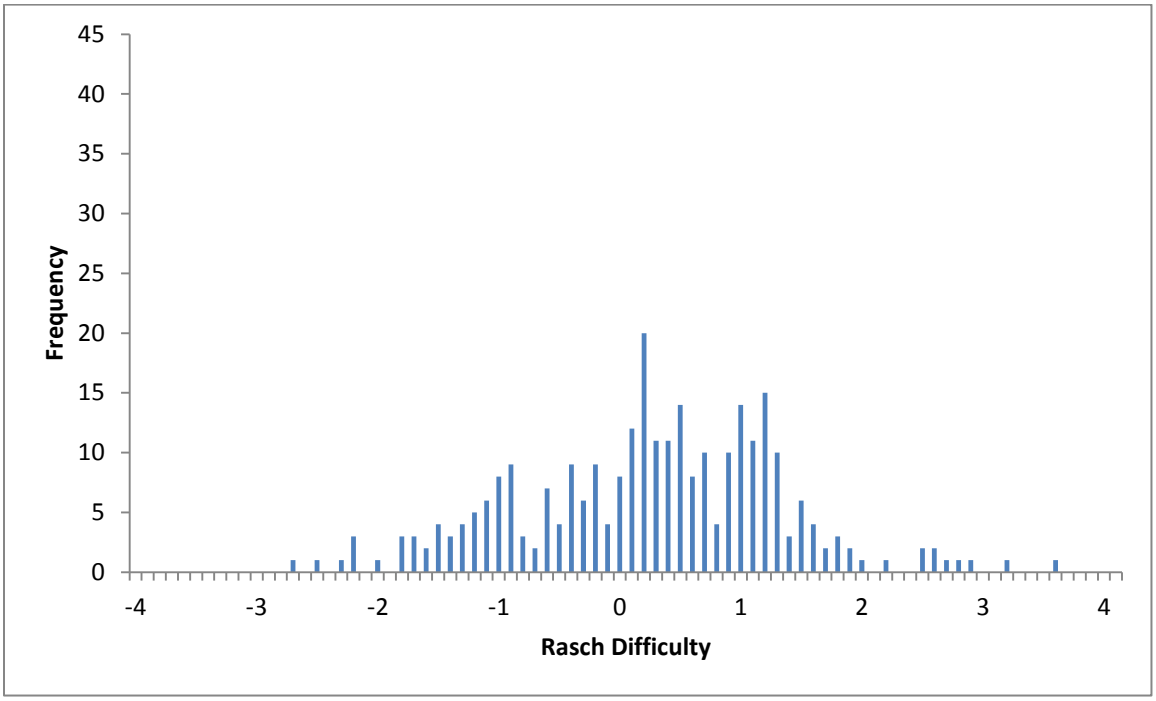
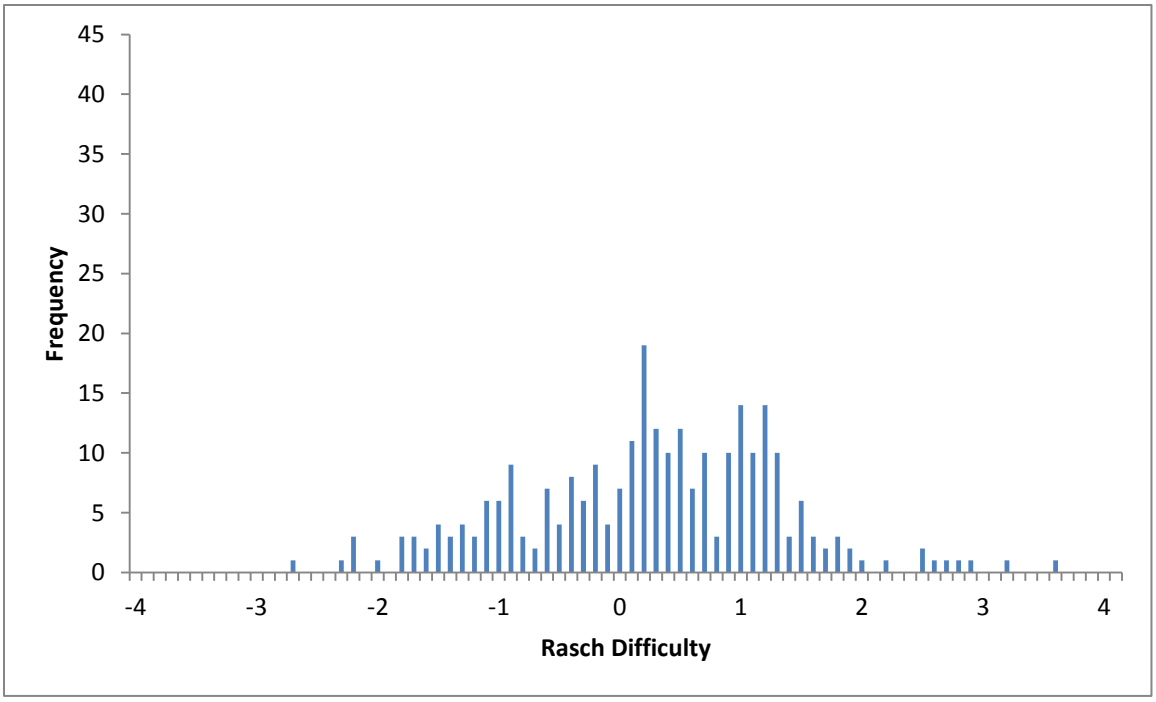


Figure 2.5.2.4 Rasch Item Difficulty Distribution for All Items in the Grade 6 Mathematics CAT Plain English Audio Pool



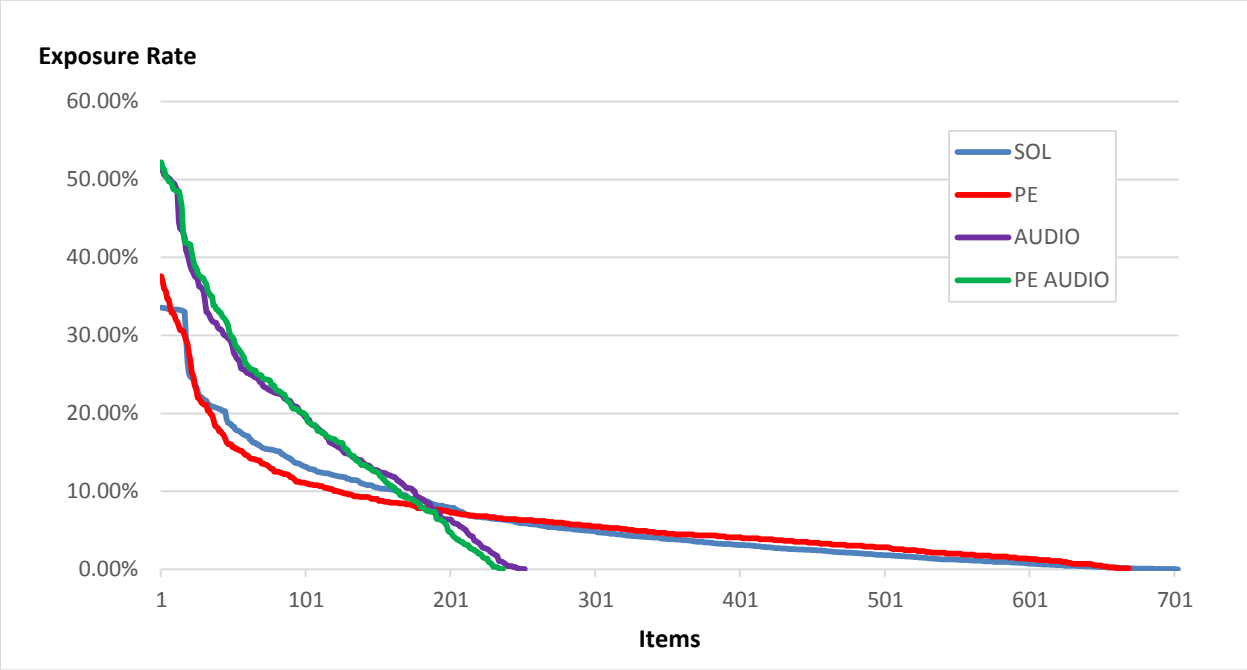
2.5.3 Item Exposure Rates

The target maximum item exposure rate for the Grade 6 Mathematics SOL and plain English (PE) CAT pools is 0.33. For the smaller audio CAT pools, the target maximum item exposure rate is 0.50. Table 2.5.3.1 shows that the observed maximum exposure rates were very close to the target maximum exposure rates. Very few items in the CAT pool have item exposures as high as the maximum exposure rate. As shown in Figure 2.5.3.1, the item exposure rate is much lower for many items. In the larger CAT SOL and plain English pools, most of the item pool is exposed to less than 10% of the student population. Even for the smaller audio pools, more than half of the item pool is exposed to less than 20% of the student population.

Table 2.5.3.1 Maximum Item Exposure Rate by Mathematics CAT Pool

Grade	SOL Pool	Plain English Pool	SOL Audio Pool	Plain English Audio Pool
6	0.3358	0.3759	0.5133	0.5219

Figure 2.5.3.1 Item Exposure Rates for Grade 6 Mathematics CAT Pools



2.5.4 Pool Usage

Not every item in the CAT pools is administered to students. Some of the items did not help to meet the content and statistical constraints given the characteristics of the student population. These items would have a 0% exposure rate. However, using a large proportion of the CAT pool is desirable to keep item exposure across all items low. Table 2.5.4.1 provides the percentage of items within each CAT pool that were administered during the spring 2015 administration. For

all CAT pools the percentage of items used is above 80%. For the smaller audio pools, a greater percentage of items were administered.

Table 2.5.4.1 Percentage of Items Used within each Mathematics CAT Pool

Grade	SOL Pool	Plain English Pool	SOL Audio Pool	Plain English Audio Pool
6	81.38%	83.31%	87.80%	87.78%

2.5.5 On Target Rates for Meeting Content and Statistical Criteria

Each of the four Grade 6 Mathematics CAT pools had 129 content and statistical constraints that were incorporated into the adaptive algorithm. For the SOL and Plain English tests, 127 of the constraints (98%) were met 100% of the time. The other two constraints were met 96% of the time or more. For the SOL audio pool, 124 constraints (96%) were met 100% of the time. The other 5 constraints were met at least 94% of the time. Finally, the plain English audio tests met 121 constraints (94%) and met the other 8 constraints at least 94% of the time. Constraints that were not met included having exactly the number of items specified for each of 45 SOL categories and being slightly off between the balance of MC items and TEIs.

2.5.6 CSEM

Table 2.5.6.1 Average CSEM for each Mathematics CAT Pool

Grade	Pool	Number of Students	Mean	Standard Deviation	Minimum	Maximum
6	SOL	80,230	0.3511	0.0750	0.3040	1.0249
	Plain English	854	0.3313	0.0285	0.3054	0.5420
	Audio	6,448	0.3378	0.0387	0.3041	1.0356
	Plain English Audio	2,872	0.3149	0.0438	0.3065	1.0242

Table 2.5.6.2 Average CSEM at the Basic, Proficient, and Advanced Cut Scores for each Mathematics CAT Pool

Grade	Pool	Performance Level	Mean	Standard Deviation	Minimum	Maximum
6	SOL	Basic	0.330	0.006	0.318	0.366
		Proficient	0.312	0.004	0.304	0.330
		Advanced	0.388	0.008	0.363	0.416
	Plain English	Basic	0.330	0.007	0.320	0.351
		Proficient	0.312	0.005	0.308	0.329
		Advanced	0.384	0.012	0.371	0.395
	Audio	Basic	0.339	0.006	0.327	0.363
		Proficient	0.315	0.004	0.306	0.328
		Advanced	0.439	0.012	0.420	0.462

Grade	Pool	Performance Level	Mean	Standard Deviation	Minimum	Maximum
	Plain English Audio	Basic	0.340	0.006	0.328	0.357
		Proficient	0.315	0.004	0.307	0.327
		Advanced	0.442	0.016	0.424	0.461

Figure 2.5.6.1 Average CSEM across the Score Scale for the Grade 6 Mathematics SOL CAT Pool

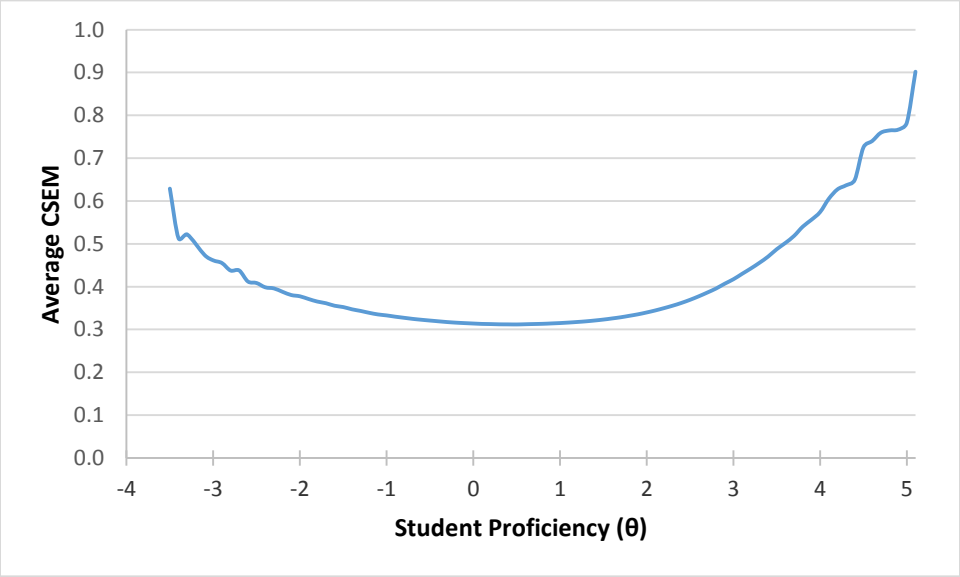


Figure 2.5.6.2 Average CSEM across the Score Scale for the Grade 6 Mathematics Plain English CAT Pool

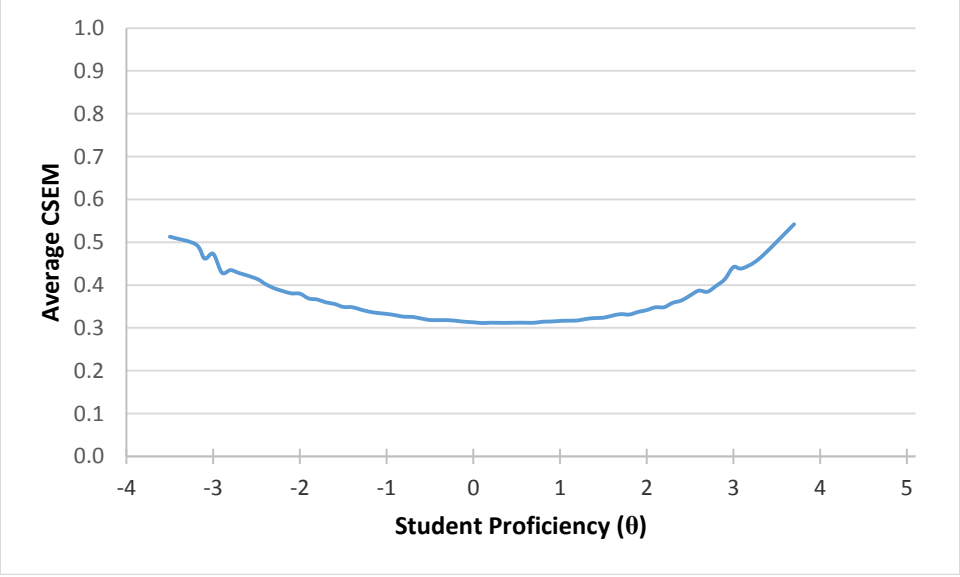


Figure 2.5.6.3 Average CSEM across the Score Scale for the Grade 6 Mathematics SOL Audio CAT Pool

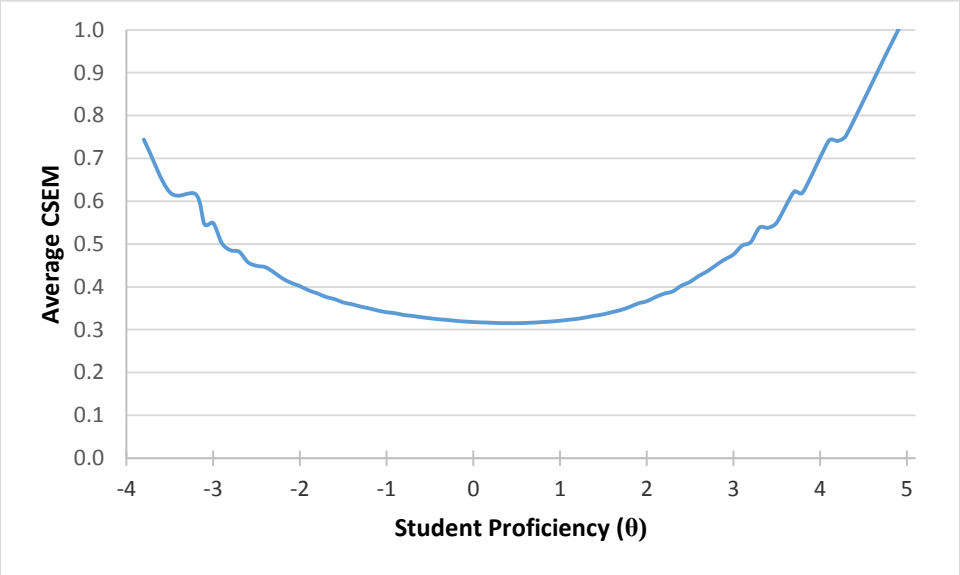
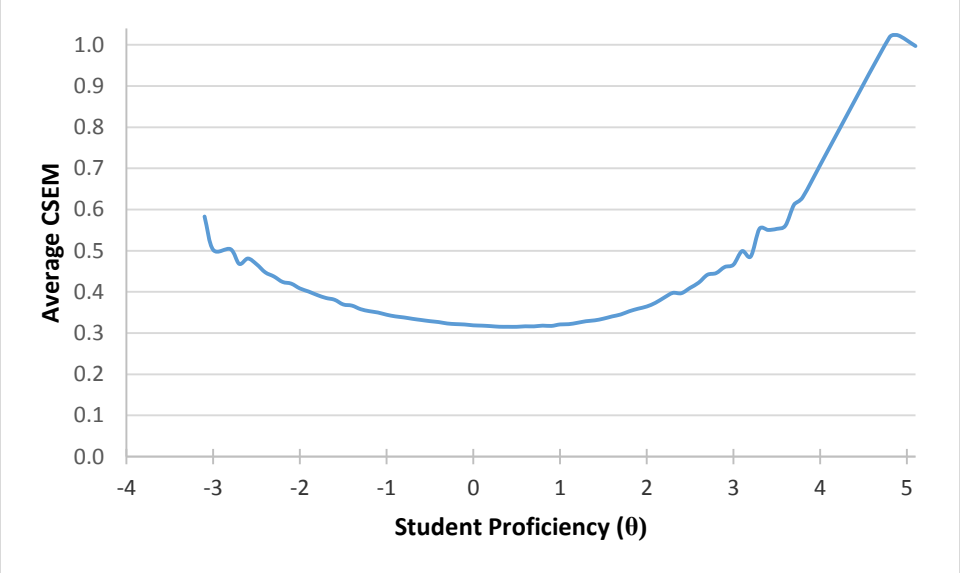


Figure 2.5.6.4 Average CSEM across the Score Scale for the Grade 6 Mathematics Plain English Audio CAT Pool



2.6 Raw Score to Scale Score (RSSS) Conversion Tables and Conditional SEM

Table 2.6.1 RSSS Conversions for Grade 3 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	117	64	115	64
2	163	46	161	46
3	191	38	189	38
4	211	34	210	34
5	228	31	226	31
6	242	28	240	29
7	254	27	253	27
8	265	26	264	26
9	275	25	274	25
10	284	24	284	24
11	293	23	293	23
12	302	23	301	23
13	310	22	309	22
14	317	22	317	22
15	325	22	325	22
16	332	21	332	21
17	340	21	340	21
18	347	21	347	21
19	354	21	354	21
20	361	21	361	21
21	368	21	368	21
22	375	21	375	21
23	382	21	382	21
24	390	21	390	21
25	397	22	397	22
26	405	22	405	22
27	412	22	413	22
28	421	23	421	23
29	429	23	430	23
30	438	24	439	24
31	448	25	448	25
32	458	26	459	26
33	469	27	470	27
34	481	29	482	29
35	495	31	497	31
36	512	34	514	34
37	532	38	535	39
38	560	46	563	46
39	600		600	
40	600		600	

Table 2.6.2 RSSS Conversions for Grade 4 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	126	62	126	62
2	171	45	172	45
3	198	37	200	38
4	218	33	220	33
5	234	30	237	30
6	247	28	251	28
7	259	26	264	27
8	270	25	275	26
9	280	24	285	25
10	289	23	295	24
11	298	22	304	23
12	306	22	313	22
13	313	22	321	22
14	321	21	329	22
15	328	21	336	21
16	335	21	344	21
17	342	20	351	21
18	349	20	358	21
19	356	20	365	21
20	363	20	372	21
21	370	20	379	21
22	376	20	386	21
23	383	21	393	21
24	390	21	400	21
25	398	21	407	21
26	405	21	414	21
27	413	22	422	22
28	420	22	430	22
29	429	23	438	22
30	437	23	446	23
31	447	24	455	24
32	457	25	465	25
33	468	26	476	26
34	480	28	488	28
35	494	30	501	30
36	511	33	517	33
37	531	38	537	37
38	559	45	564	44
39	600		600	
40	600		600	

Table 2.6.3 RSSS Conversions for Grade 5 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	106	66	114	66
2	154	48	162	48
3	184	40	191	40
4	205	35	213	35
5	223	32	231	32
6	238	30	246	30
7	251	28	259	28
8	262	27	271	27
9	273	26	282	26
10	283	25	292	25
11	292	24	301	24
12	301	24	310	24
13	310	23	319	23
14	318	23	327	23
15	326	22	335	23
16	334	22	343	22
17	341	22	350	22
18	349	22	358	22
19	356	22	365	22
20	364	22	373	22
21	371	22	380	22
22	378	22	388	22
23	386	22	395	22
24	393	22	403	22
25	401	22	411	22
26	409	23	418	23
27	417	23	427	23
28	425	23	435	24
29	434	24	444	24
30	443	25	453	25
31	453	26	463	26
32	463	27	473	27
33	475	28	485	28
34	487	30	498	30
35	502	32	512	32
36	519	35	529	35
37	540	39	551	39
38	569	47	579	47
39	600		600	
40	600		600	

Table 2.6.4 RSSS Conversions for Grade 5 Science

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	129	60	139	60
2	173	43	183	43
3	199	36	209	36
4	219	32	228	32
5	234	29	244	29
6	248	27	257	27
7	260	26	269	25
8	270	25	279	24
9	280	24	289	23
10	289	23	298	23
11	298	22	306	22
12	306	22	314	21
13	314	21	322	21
14	322	21	329	21
15	329	21	336	20
16	336	21	343	20
17	343	20	350	20
18	350	20	357	20
19	357	20	364	20
20	364	20	370	20
21	371	20	377	20
22	378	20	384	20
23	386	21	391	20
24	393	21	398	20
25	400	21	405	21
26	408	21	412	21
27	415	22	419	21
28	424	22	427	22
29	432	23	435	22
30	441	23	444	23
31	450	24	453	24
32	460	25	463	25
33	471	26	473	26
34	483	27	485	27
35	497	29	499	29
36	513	32	515	32
37	533	36	535	36
38	560	44	562	44
39	600		600	
40	600		600	

Table 2.6.5 RSSS Conversions for Grade 6 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	124	59	124	59
2	166	42	166	42
3	192	35	192	35
4	211	31	211	31
5	226	28	226	28
6	239	26	239	26
7	250	25	250	25
8	260	24	261	24
9	269	23	270	23
10	278	22	278	22
11	286	21	286	21
12	293	21	294	21
13	300	20	301	20
14	307	20	308	20
15	314	19	314	19
16	320	19	321	19
17	327	19	327	19
18	333	19	333	19
19	339	19	339	19
20	345	18	345	18
21	350	18	351	18
22	356	18	357	18
23	362	18	362	18
24	368	18	368	18
25	374	18	374	18
26	380	19	380	18
27	386	19	386	19
28	392	19	392	19
29	398	19	398	19
30	404	19	404	19
31	411	20	411	19
32	418	20	417	20
33	425	20	424	20
34	432	21	432	21
35	440	22	439	21
36	448	22	448	22
37	457	23	457	23
38	467	25	466	24
39	478	26	477	26
40	491	28	490	28
41	506	31	505	31
42	525	35	523	35
43	550	42	549	42
44	592	59	591	59
45	600		600	

Table 2.6.6 RSSS Conversions for Grade 7 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	132	59	141	59
2	175	42	183	42
3	200	35	209	35
4	219	31	228	31
5	235	28	243	28
6	248	26	256	26
7	259	25	267	25
8	269	24	277	23
9	278	23	286	22
10	287	22	295	22
11	295	21	303	21
12	303	21	310	20
13	310	20	317	20
14	317	20	324	20
15	323	19	330	19
16	330	19	337	19
17	336	19	343	19
18	342	19	349	19
19	348	19	355	18
20	354	18	361	18
21	360	18	366	18
22	366	18	372	18
23	372	18	378	18
24	378	18	384	18
25	383	18	390	18
26	389	18	395	18
27	395	19	401	18
28	401	19	407	19
29	408	19	413	19
30	414	19	420	19
31	421	20	426	19
32	427	20	433	20
33	434	20	440	20
34	442	21	447	21
35	450	22	455	21
36	458	22	463	22
37	467	23	472	23
38	477	24	482	24
39	488	26	493	26
40	501	28	506	28
41	515	31	521	31
42	534	35	539	35
43	559	42	565	42
44	600		600	
45	600		600	

Table 2.6.7 RSSS Conversions for Grade 8 Reading

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	139	56	142	56
2	179	41	183	41
3	204	34	207	34
4	222	30	226	30
5	237	27	240	27
6	249	25	253	25
7	260	24	263	24
8	270	23	273	23
9	279	22	282	22
10	287	21	290	21
11	295	21	298	20
12	302	20	305	20
13	309	20	312	19
14	316	19	318	19
15	323	19	325	19
16	329	19	331	18
17	336	19	337	18
18	342	18	343	18
19	348	18	348	18
20	354	18	354	18
21	360	18	360	18
22	366	18	365	18
23	372	18	371	18
24	378	18	377	18
25	384	18	382	18
26	390	18	388	18
27	396	19	394	18
28	403	19	400	18
29	409	19	406	18
30	416	19	412	19
31	422	19	419	19
32	429	20	425	19
33	437	20	432	20
34	444	21	440	20
35	452	21	448	21
36	461	22	456	22
37	470	23	465	23
38	480	24	475	24
39	492	26	486	26
40	505	28	499	28
41	520	30	514	30
42	539	34	533	35
43	564	41	559	41
44	600		600	
45	600		600	

Table 2.6.8 RSSS Conversions for Grade 8 Science

Raw Score	Online			
	Core 1		Core 2	
	SS	SEM	SS	SEM
0	0		0	
1	200	46	197	46
2	233	33	230	34
3	253	28	251	28
4	267	24	265	25
5	279	22	278	22
6	289	20	288	21
7	298	19	297	20
8	305	18	305	19
9	313	18	312	18
10	319	17	319	17
11	325	16	325	17
12	331	16	331	16
13	336	16	337	16
14	342	15	342	16
15	347	15	347	15
16	351	15	352	15
17	356	15	357	15
18	361	14	362	15
19	365	14	367	15
20	370	14	371	14
21	374	14	376	14
22	378	14	380	14
23	383	14	385	14
24	387	14	389	14
25	391	14	394	14
26	395	14	398	14
27	400	14	403	14
28	404	14	407	14
29	408	14	412	14
30	412	14	416	14
31	417	14	421	15
32	421	14	425	15
33	426	15	430	15
34	431	15	435	15
35	436	15	440	15
36	441	15	445	16
37	446	16	451	16
38	451	16	456	16
39	457	16	462	17
40	463	17	469	17
41	470	18	475	18
42	477	18	483	19
43	485	19	490	20
44	493	21	499	21
45	503	22	509	22
46	515	24	521	25
47	530	28	536	28
48	550	33	557	33
49	583	46	590	46
50	600		600	

Table 2.6.9 RSSS Conversions for EOC Reading

Raw Score	Online					
	Core 1		Core 2		Core 3	
	SS	SEM	SS	SEM	SS	SEM
0	0		0		0	
1	188	48	186	48	186	48
2	222	34	220	34	220	34
3	243	28	241	28	240	28
4	258	25	256	25	255	25
5	270	23	268	23	267	23
6	280	21	278	21	277	21
7	289	20	287	20	286	20
8	297	19	295	19	294	19
9	304	18	302	18	301	18
10	310	17	308	17	307	17
11	316	17	314	17	313	17
12	322	16	320	16	319	16
13	328	16	326	16	324	16
14	333	15	331	15	329	15
15	338	15	336	15	334	15
16	342	15	340	15	339	15
17	347	15	345	15	343	14
18	351	14	350	14	347	14
19	356	14	354	14	352	14
20	360	14	358	14	356	14
21	364	14	362	14	360	14
22	368	14	367	14	364	14
23	372	14	371	14	368	14
24	376	14	375	14	372	14
25	380	14	379	14	376	14
26	384	14	383	14	380	13
27	388	14	387	14	383	13
28	392	14	391	14	387	13
29	396	14	395	14	391	13
30	400	14	398	14	395	14
31	404	14	402	14	399	14
32	408	14	407	14	403	14
33	412	14	411	14	407	14
34	416	14	415	14	411	14
35	420	14	419	14	415	14
36	425	14	423	14	419	14
37	429	14	428	14	423	14
38	433	15	432	15	428	14
39	438	15	437	15	432	15
40	443	15	441	15	437	15
41	448	15	446	15	442	15
42	453	16	452	16	447	16
43	458	16	457	16	452	16
44	464	17	463	17	458	17
45	470	17	469	17	464	17
46	476	18	475	18	470	18
47	483	19	483	19	477	19
48	491	20	490	20	485	20
49	500	21	499	21	493	21
50	510	23	509	23	503	23
51	522	25	521	25	515	25
52	537	28	536	28	530	28
53	558	34	557	34	551	34
54	592	48	591	48	585	48
55	600		600		600	

Table 2.6.10 RSSS Conversions for EOC Earth Science

Raw Score	Online			
	Core 1		Core 3	
	SS	SEM	SS	SEM
0	0		0	
1	217	43	227	43
2	248	31	258	31
3	267	26	277	26
4	280	23	290	23
5	291	21	301	21
6	300	19	310	19
7	308	18	318	18
8	316	17	326	17
9	322	16	332	16
10	328	16	338	16
11	334	15	344	15
12	339	15	349	15
13	344	14	354	14
14	349	14	359	14
15	354	14	363	14
16	358	14	368	14
17	362	14	372	13
18	367	13	376	13
19	371	13	380	13
20	375	13	384	13
21	379	13	388	13
22	383	13	392	13
23	387	13	396	13
24	391	13	400	13
25	395	13	404	13
26	399	13	407	13
27	403	13	411	13
28	407	13	415	13
29	411	13	419	13
30	415	13	423	13
31	419	13	427	13
32	423	13	431	13
33	428	14	435	13
34	432	14	440	14
35	437	14	444	14
36	441	14	449	14
37	446	15	453	14
38	451	15	458	15
39	457	15	464	15
40	463	16	469	16
41	469	16	475	16
42	475	17	482	17
43	483	18	489	18
44	491	19	497	19
45	500	21	507	21
46	511	23	518	23
47	525	26	532	26
48	544	31	551	31
49	575	43	582	43
50	600		600	

Table 2.6.11 RSSS Conversions for EOC Biology

Raw Score	Online			
	Core 1		Core 3	
	SS	SEM	SS	SEM
0	0		0	
1	207	46	216	46
2	239	33	248	33
3	259	27	268	27
4	273	24	282	24
5	285	22	294	22
6	295	20	304	20
7	303	19	312	19
8	311	18	320	18
9	318	17	326	17
10	324	17	333	17
11	330	16	339	16
12	335	16	344	16
13	341	15	349	15
14	346	15	354	15
15	350	15	359	15
16	355	14	364	14
17	360	14	368	14
18	364	14	373	14
19	368	14	377	14
20	373	14	381	14
21	377	14	385	14
22	381	14	389	13
23	385	14	393	13
24	389	14	397	13
25	393	14	401	13
26	397	14	405	13
27	401	14	409	13
28	405	14	414	13
29	410	14	418	14
30	414	14	422	14
31	418	14	426	14
32	422	14	430	14
33	427	14	435	14
34	432	15	439	14
35	436	15	444	15
36	441	15	448	15
37	446	15	453	15
38	452	16	459	16
39	458	16	464	16
40	464	17	470	17
41	471	18	477	17
42	478	19	484	18
43	486	20	491	19
44	495	21	500	20
45	506	23	510	22
46	518	25	521	24
47	535	29	536	28
48	557	35	556	33
49	593	48	590	46
50	600		600	

Table 2.6.12 RSSS Conversions for EOC Chemistry

Raw Score	Online			
	Core 1		Core 3	
	SS	SEM	SS	SEM
0	0		0	
1	208	45	213	45
2	241	33	245	32
3	261	27	265	27
4	275	24	279	24
5	287	22	291	22
6	297	20	301	20
7	306	19	309	19
8	313	18	317	18
9	320	17	324	17
10	327	17	330	17
11	333	16	336	16
12	339	16	342	16
13	344	15	347	15
14	350	15	352	15
15	355	15	357	15
16	359	15	362	14
17	364	14	367	14
18	369	14	371	14
19	373	14	375	14
20	378	14	380	14
21	382	14	384	14
22	386	14	388	14
23	390	14	392	14
24	395	14	396	13
25	399	14	401	13
26	403	14	405	13
27	407	14	409	14
28	412	14	413	14
29	416	14	417	14
30	420	14	421	14
31	425	14	426	14
32	429	14	430	14
33	434	14	435	14
34	438	15	439	14
35	443	15	444	15
36	448	15	449	15
37	453	15	454	15
38	459	16	460	16
39	465	16	465	16
40	471	17	471	17
41	477	17	478	17
42	484	18	485	18
43	492	19	493	19
44	501	20	502	20
45	511	22	511	22
46	523	24	523	24
47	537	27	538	27
48	557	33	558	33
49	590	45	590	45
50	600		600	

Table 2.6.13 RSSS Conversions for EOC Algebra I

Raw Score	Core 3	
	SS	SEM
0	0	
1	240	39
2	268	28
3	285	23
4	298	21
5	308	19
6	316	17
7	323	16
8	330	15
9	336	15
10	341	14
11	346	14
12	351	13
13	356	13
14	360	13
15	364	13
16	368	12
17	372	12
18	376	12
19	380	12
20	383	12
21	387	12
22	391	12
23	394	12
24	398	12
25	401	12
26	404	12
27	408	12
28	411	12
29	415	12
30	419	12
31	422	12
32	426	12
33	430	12
34	433	12
35	437	12
36	441	13
37	446	13
38	450	13
39	455	14
40	460	14
41	465	15
42	471	15
43	478	16
44	485	17
45	493	19
46	503	20
47	515	23
48	532	28
49	560	39
50	600	

Table 2.6.14 RSSS Conversions for EOC Geometry

Raw Score	Core 3	
	SS	SEM
0	0	
1	216	44
2	248	32
3	267	26
4	282	23
5	293	21
6	303	20
7	311	19
8	319	18
9	326	17
10	332	16
11	338	16
12	344	15
13	349	15
14	354	15
15	359	14
16	364	14
17	368	14
18	373	14
19	377	14
20	381	14
21	386	13
22	390	13
23	394	13
24	398	13
25	402	13
26	406	13
27	410	13
28	415	13
29	419	14
30	423	14
31	427	14
32	432	14
33	436	14
34	441	14
35	446	14
36	451	15
37	456	15
38	461	15
39	467	16
40	473	16
41	480	17
42	487	18
43	494	19
44	503	20
45	513	21
46	524	23
47	539	27
48	558	32
49	590	44
50	600	

Table 2.6.15 RSSS Conversions for EOC Algebra II

Raw Score	Core 3	
	SS	SEM
0	0	
1	163	55
2	203	40
3	228	33
4	246	29
5	260	27
6	273	25
7	283	23
8	293	22
9	302	21
10	310	21
11	318	20
12	325	19
13	332	19
14	338	19
15	344	18
16	350	18
17	356	18
18	362	17
19	367	17
20	373	17
21	378	17
22	383	17
23	389	17
24	394	17
25	399	17
26	404	17
27	409	17
28	415	17
29	420	17
30	425	17
31	430	17
32	436	17
33	441	17
34	447	18
35	453	18
36	459	18
37	465	19
38	472	19
39	479	20
40	486	20
41	494	21
42	503	22
43	512	23
44	522	24
45	534	26
46	548	29
47	566	33
48	590	40
49	600	
50	600	

Table 2.6.16 RSSS Conversions for EOC World Geography

Raw Score	Core 2	
	SS	SEM
0	0	
1	188	45
2	220	33
3	240	27
4	254	24
5	266	22
6	275	20
7	283	19
8	291	18
9	298	17
10	304	16
11	310	16
12	315	15
13	320	15
14	325	15
15	329	14
16	334	14
17	338	14
18	342	14
19	346	13
20	350	13
21	354	13
22	358	13
23	362	13
24	365	13
25	369	13
26	373	13
27	376	13
28	380	13
29	383	13
30	387	12
31	390	12
32	394	13
33	397	13
34	401	13
35	404	13
36	408	13
37	411	13
38	415	13
39	419	13
40	423	13
41	427	13
42	430	13
43	435	14

44	439	14
45	443	14
46	448	14
47	452	15
48	457	15
49	463	16
50	468	16
51	475	17
52	481	18
53	488	19
54	497	20
55	506	21
56	517	24
57	531	27
58	551	32
59	583	45
60	600	

Table 2.6.17 RSSS Conversions for Grade 8 Writing Core 1

Raw Score	Core 1											
	Prompt 2809		Prompt 2813		Prompt 2816		Prompt 2817		Prompt 2853		Prompt 2854	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	14		13		14		14		14		14	
2	27		25		28		28		28		28	
3	41		38		42		42		43		42	
4	55		51		56		56		57		55	
5	68		63		70		70		71		69	
6	82	113	76	112	84	112	84	113	85	113	83	113
7	156	55	148	54	157	54	158	54	159	54	157	55
8	192	38	183	38	192	38	193	38	194	38	192	38
9	213	32	203	32	212	31	213	31	214	31	213	31
10	228	28	219	28	228	28	228	28	229	28	228	28
11	240	26	231	26	240	26	241	26	242	26	241	26
12	251	24	243	25	251	24	252	24	252	24	251	24
13	261	24	253	24	261	23	262	23	262	23	261	23
14	271	23	263	24	270	23	271	23	271	23	271	23
15	280	23	272	23	279	22	280	22	280	22	279	22
16	288	22	282	23	287	22	288	22	288	22	288	22
17	297	22	290	22	296	22	297	22	296	22	296	22
18	305	22	299	22	304	21	305	22	304	21	304	21
19	313	21	307	21	312	21	313	21	312	21	312	21
20	321	21	315	21	319	21	321	21	320	21	319	21
21	329	21	322	21	327	21	328	21	327	21	327	21
22	336	21	329	20	334	21	336	21	335	20	334	20
23	343	20	337	20	342	20	343	20	342	20	341	20
24	351	20	344	20	349	20	350	20	349	20	349	20
25	358	20	351	20	356	20	357	20	356	20	356	20
26	365	20	358	20	363	20	364	20	363	20	363	20
27	372	20	365	20	370	20	371	20	370	20	370	20
28	379	20	372	20	378	21	378	20	377	20	377	20
29	386	21	379	20	385	21	385	20	384	21	384	20
30	394	21	386	21	393	21	393	21	392	21	391	21
31	401	21	394	21	400	21	400	21	399	21	399	21
32	409	22	402	21	408	22	408	21	407	22	407	21
33	417	22	410	22	417	22	416	22	416	22	415	22
34	426	23	418	22	426	23	425	22	424	23	423	22
35	435	23	426	22	435	23	434	23	433	23	432	23
36	445	24	435	23	445	24	443	23	443	24	441	23
37	455	25	444	23	455	25	453	24	453	24	451	24
38	466	25	453	23	466	26	463	25	463	25	461	24
39	477	26	463	24	478	26	474	26	474	25	471	25
40	490	27	473	25	491	27	486	26	486	26	482	26
41	503	28	484	25	504	28	498	27	498	27	494	26
42	517	29	496	26	519	30	512	28	511	28	506	27
43	533	31	509	28	535	31	526	30	525	29	520	29
44	550	33	523	30	553	33	543	32	541	31	535	31
45	570	36	541	34	573	36	562	35	559	34	554	34
46	596	42	565	41	599	42	586	41	583	41	578	41
47	600		600		600		600		600		600	
48	600		600		600		600		600		600	

Table 2.6.18 RSSS Conversions for Grade 8 Writing Core 2

Raw Score	Core 2											
	Prompt 2809		Prompt 2813		Prompt 2816		Prompt 2817		Prompt 2853		Prompt 2854	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	12		11		12		12		13		12	
2	24		22		25		25		25		24	
3	37		34		37		37		38		37	
4	49		45		49		49		50		49	
5	61		56		62		62		63		61	
6	73	114	67	113	74	114	74	114	75	114	73	114
7	149	56	142	55	150	55	150	56	151	56	149	56
8	186	39	177	38	186	39	187	39	188	39	186	39
9	207	32	199	32	207	32	208	32	209	32	208	32
10	223	28	214	28	223	28	224	28	224	28	223	28
11	236	26	227	26	235	26	236	26	237	26	236	26
12	247	25	238	25	246	24	247	24	248	24	247	24
13	257	24	249	24	256	23	257	23	258	23	257	23
14	266	23	258	23	265	23	266	23	267	23	266	23
15	275	22	268	23	274	22	275	22	275	22	275	22
16	284	22	277	23	283	22	284	22	284	22	283	22
17	292	22	286	22	291	22	292	22	292	22	291	22
18	300	22	294	22	299	21	300	22	300	21	300	21
19	308	21	302	22	307	21	308	21	308	21	307	21
20	316	21	310	21	315	21	316	21	316	21	315	21
21	324	21	318	21	323	21	324	21	323	21	323	21
22	332	21	325	21	330	21	331	21	331	21	330	21
23	339	21	333	20	338	21	339	21	338	20	337	20
24	347	21	340	20	345	21	346	20	345	20	345	20
25	354	20	347	20	352	21	353	20	352	20	352	20
26	361	20	354	20	360	21	360	20	359	20	359	20
27	369	20	361	20	367	21	368	20	367	20	366	20
28	376	21	369	20	374	21	375	21	374	21	373	21
29	383	21	376	21	382	21	382	21	381	21	381	21
30	391	21	383	21	390	21	390	21	389	21	388	21
31	399	21	391	21	398	22	398	21	397	21	396	21
32	407	22	399	22	406	22	406	22	405	22	404	22
33	415	22	407	22	415	23	414	22	413	22	413	22
34	424	23	416	22	424	23	423	23	422	23	421	23
35	433	23	424	23	433	24	432	23	431	23	430	23
36	443	24	433	23	443	24	441	24	441	24	440	24
37	454	25	443	23	454	25	451	24	451	25	449	24
38	465	26	452	24	465	26	462	25	462	25	460	25
39	477	26	462	24	477	27	473	26	474	26	471	25
40	489	27	473	25	490	28	485	27	485	26	482	26
41	503	28	484	26	504	29	498	27	498	27	494	27
42	517	30	495	27	519	30	512	29	511	28	506	28
43	533	31	509	28	535	31	526	30	525	29	520	29
44	551	33	523	31	553	33	543	32	541	31	536	31
45	571	36	541	34	574	36	562	35	560	35	554	34
46	596	42	566	41	600		587	41	584	41	578	41
47	600		600		600		600		600		600	
48	600		600		600		600		600		600	

Table 2.6.19 RSSS Conversions for Grade 8 Writing Core 3

Raw Score	Core 3											
	Prompt 2809		Prompt 2813		Prompt 2816		Prompt 2817		Prompt 2853		Prompt 2854	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	12		11		12		12		12		12	
2	24		22		24		24		24		24	
3	36		33		36		36		37		36	
4	47		43		48		48		49		47	
5	59		54		60		60		61		59	
6	71	114	65	113	72	114	72	114	73	114	71	114
7	147	56	140	55	148	55	148	56	149	56	148	56
8	184	39	176	39	184	39	185	39	186	39	185	39
9	206	32	197	32	206	32	207	32	208	32	206	32
10	222	28	213	28	221	28	222	28	223	28	222	28
11	234	26	226	26	234	26	235	26	236	26	235	26
12	245	25	237	25	245	24	246	24	246	24	246	24
13	255	24	247	24	255	23	256	23	256	23	255	23
14	265	23	257	23	264	23	265	23	265	23	265	23
15	274	22	266	23	273	22	274	22	274	22	273	22
16	282	22	275	23	281	22	282	22	282	22	282	22
17	291	22	284	22	289	22	290	22	290	22	290	22
18	299	22	292	22	297	21	299	22	298	21	298	21
19	307	21	301	22	305	21	307	21	306	21	306	21
20	315	21	308	21	313	21	314	21	314	21	313	21
21	322	21	316	21	321	21	322	21	321	21	321	21
22	330	21	324	21	328	21	330	21	329	21	328	21
23	338	21	331	20	336	21	337	21	336	20	336	20
24	345	21	338	20	343	21	344	20	343	20	343	20
25	352	20	345	20	351	21	352	20	351	20	350	20
26	360	20	353	20	358	21	359	20	358	20	357	20
27	367	21	360	20	365	21	366	20	365	20	365	20
28	374	21	367	20	373	21	373	21	372	21	372	21
29	382	21	374	21	380	21	381	21	380	21	379	21
30	389	21	382	21	388	21	388	21	387	21	387	21
31	397	21	390	21	396	22	396	21	395	21	395	21
32	405	22	397	22	404	22	404	22	403	22	403	22
33	414	22	406	22	413	23	412	22	412	22	411	22
34	423	23	414	22	422	23	421	23	421	23	420	23
35	432	23	423	23	432	24	430	23	430	23	429	23
36	442	24	432	23	442	24	440	24	440	24	438	24
37	452	25	441	23	452	25	450	24	450	25	448	24
38	463	26	451	24	464	26	461	25	461	25	458	25
39	475	27	461	24	476	27	472	26	472	26	469	25
40	488	27	471	25	489	28	484	27	484	27	481	26
41	501	28	482	26	503	29	497	28	497	27	493	27
42	516	30	494	27	518	30	510	29	510	28	505	28
43	532	31	507	28	534	31	525	30	524	29	519	29
44	550	33	522	31	552	33	542	32	540	31	535	31
45	570	36	540	34	573	36	561	35	559	35	553	34
46	596	42	565	41	599	42	586	41	583	41	577	41
47	600		600		600		600		600		600	
48	600		600		600		600		600		600	

Table 2.6.20 RSSS Conversions for EOC Writing Core 1

Raw Score	Core 1											
	Prompt 2101		Prompt 2103		Prompt 2106		Prompt 2108		Prompt 2116		Prompt 2124	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	17		17		16		17		17		17	
2	34		33		33		33		34		34	
3	51		50		49		50		52		51	
4	68		67		65		67		69		68	
5	85		83		82		83		86		85	
6	102	106	100	106	98	106	100	106	103	106	102	105
7	172	52	170	51	169	52	170	51	173	51	171	51
8	206	36	203	36	203	36	203	36	206	36	204	35
9	226	30	222	29	223	30	222	29	226	29	223	29
10	240	26	236	26	238	26	237	26	240	26	236	26
11	252	24	247	24	249	24	248	24	251	24	248	23
12	261	22	257	22	259	22	258	22	261	22	257	22
13	270	21	266	21	268	21	267	21	269	21	266	21
14	278	21	274	21	276	21	275	21	278	21	274	20
15	286	20	282	20	284	20	282	20	285	20	281	20
16	294	20	289	20	291	20	290	20	293	20	289	20
17	301	20	296	19	298	19	297	19	300	20	296	19
18	308	19	303	19	305	19	303	19	307	19	303	19
19	315	19	310	19	312	19	310	19	314	19	309	19
20	321	19	316	19	319	19	316	18	320	19	316	19
21	328	19	322	18	325	18	323	18	327	19	322	18
22	334	18	328	18	331	18	329	18	333	18	329	18
23	340	18	335	18	337	18	335	18	339	18	335	18
24	346	18	340	18	343	18	341	18	346	18	341	18
25	352	18	346	18	349	18	346	18	352	18	346	18
26	358	18	352	18	355	18	352	18	357	18	352	18
27	363	17	358	17	361	18	358	18	363	18	358	18
28	369	17	363	17	366	18	364	18	369	17	364	17
29	375	17	369	17	372	18	369	18	375	17	369	18
30	380	17	375	18	378	18	375	18	380	17	375	18
31	386	17	380	18	384	18	381	18	386	18	381	18
32	391	17	386	18	389	18	387	18	392	18	387	18
33	397	18	392	18	395	18	393	18	398	18	393	18
34	403	18	398	18	401	18	399	18	404	18	399	18
35	409	18	404	18	408	18	405	19	410	18	405	19
36	415	18	411	19	414	19	412	19	416	19	411	19
37	421	19	417	19	421	19	419	19	423	19	418	19
38	428	19	424	19	428	20	426	20	429	19	425	20
39	435	19	431	20	435	20	433	20	437	20	432	20
40	442	20	438	20	443	21	441	20	444	20	440	20
41	449	20	446	20	451	21	449	21	452	21	448	21
42	457	21	454	21	459	22	457	22	461	22	456	21
43	465	21	462	21	468	22	466	22	470	22	465	22
44	473	22	470	22	478	23	475	23	479	23	474	23
45	482	22	479	22	489	24	485	24	490	24	484	23
46	491	23	489	23	500	25	496	24	501	25	495	24
47	501	23	499	24	512	26	508	25	512	25	506	25
48	512	24	510	25	525	27	520	26	525	26	518	26
49	523	26	522	26	540	29	534	28	538	28	531	27
50	537	28	536	29	556	31	549	29	553	29	546	29
51	553	31	553	32	575	34	566	32	571	32	564	33
52	575	37	576	38	600		588	38	593	38	587	38
53	600		600		600		600		600		600	
54	600		600		600		600		600		600	

Table 2.6.21 RSSS Conversions for EOC Writing Core 2

Raw Score	Core 2											
	Prompt 2101		Prompt 2103		Prompt 2106		Prompt 2108		Prompt 2116		Prompt 2124	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	17		17		17		17		17		17	
2	34		33		33		33		34		34	
3	51		50		50		50		52		51	
4	68		67		66		67		69		68	
5	85		83		83		83		86		85	
6	102	106	100	106	99	106	100	106	103	106	102	105
7	173	52	170	51	170	52	170	51	173	51	172	51
8	207	36	203	36	204	36	204	36	207	36	204	35
9	227	30	223	29	224	30	223	29	226	29	223	29
10	241	26	237	26	238	26	237	26	240	26	237	26
11	252	24	248	24	250	24	249	24	252	24	248	24
12	262	22	258	22	260	23	259	22	262	22	258	22
13	271	21	267	21	269	22	268	21	271	21	267	21
14	280	21	275	21	277	21	276	21	279	21	275	21
15	288	20	283	20	285	20	284	20	287	20	283	20
16	295	20	291	20	293	20	291	20	294	20	290	20
17	303	20	298	20	300	20	298	19	302	20	297	20
18	310	20	305	19	307	19	305	19	309	20	305	19
19	317	19	312	19	314	19	312	19	316	19	311	19
20	324	19	318	19	321	19	318	19	323	19	318	19
21	330	19	325	19	327	19	325	18	329	19	325	19
22	337	19	331	18	334	18	331	18	336	19	331	18
23	343	18	337	18	340	18	337	18	342	18	337	18
24	349	18	343	18	346	18	343	18	349	18	343	18
25	355	18	349	18	352	18	349	18	355	18	349	18
26	361	18	355	18	358	18	355	18	361	18	355	18
27	367	18	361	18	364	18	361	18	367	18	361	18
28	372	17	367	18	370	18	367	18	372	18	367	18
29	378	17	372	18	376	18	373	18	378	18	373	18
30	384	17	378	18	381	18	379	18	384	18	379	18
31	389	17	384	18	387	18	385	18	390	18	385	18
32	395	18	390	18	393	18	391	18	396	18	391	18
33	401	18	396	18	399	18	397	18	402	18	397	18
34	407	18	402	18	406	18	403	18	408	18	403	18
35	413	18	409	19	412	19	410	19	414	18	409	19
36	419	18	415	19	419	19	416	19	420	19	416	19
37	425	19	422	19	425	19	423	19	427	19	423	19
38	432	19	428	19	432	20	430	20	434	20	430	20
39	439	19	435	20	440	20	438	20	441	20	437	20
40	446	20	443	20	447	21	445	20	449	20	445	20
41	453	20	450	20	455	21	453	21	457	21	453	21
42	461	20	458	21	464	22	462	21	465	22	461	21
43	469	21	466	21	473	22	470	22	474	22	470	22
44	477	21	474	21	483	23	480	23	483	23	479	22
45	486	22	483	22	493	24	490	23	494	24	488	23
46	495	22	493	23	504	25	500	24	504	24	498	24
47	504	23	502	24	516	26	511	25	516	25	509	25
48	515	24	513	25	528	27	523	26	528	26	521	26
49	526	26	525	26	543	28	536	27	541	27	534	27
50	540	28	539	28	559	30	551	29	556	29	549	29
51	556	31	556	32	577	33	568	32	573	32	566	32
52	577	37	578	38	600		590	38	595	38	589	38
53	600		600		600		600		600		600	
54	600		600		600		600		600		600	

Table 2.6.22 RSSS Conversions for EOC Writing Core 3

Raw Score	Core 3											
	Prompt 2101		Prompt 2103		Prompt 2106		Prompt 2108		Prompt 2116		Prompt 2124	
	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM	SS	SEM
0	0		0		0		0		0		0	
1	18		17		17		17		18		18	
2	35		34		34		34		35		35	
3	53		52		51		52		53		53	
4	70		69		67		69		71		70	
5	88		86		84		86		88		88	
6	105	106	103	105	101	106	103	105	106	106	105	105
7	175	51	172	51	172	52	172	51	176	51	174	51
8	209	36	205	36	206	36	205	36	209	36	206	35
9	228	29	224	29	226	30	225	29	228	29	225	29
10	242	26	238	26	240	26	239	26	242	26	239	26
11	254	24	250	24	251	24	250	24	253	24	250	24
12	264	22	259	22	261	23	260	22	263	22	259	22
13	273	21	268	21	270	22	269	21	272	21	268	21
14	281	21	277	21	279	21	277	21	280	21	276	21
15	289	20	284	20	287	20	285	20	288	20	284	20
16	297	20	292	20	294	20	292	20	296	20	292	20
17	304	20	299	20	301	20	299	19	303	20	299	20
18	311	20	306	19	309	19	306	19	310	20	306	19
19	318	19	313	19	315	19	313	19	317	19	313	19
20	325	19	319	19	322	19	320	19	324	19	319	19
21	332	19	326	18	329	19	326	18	331	19	326	19
22	338	18	332	18	335	18	332	18	337	19	332	18
23	344	18	338	18	341	18	339	18	344	18	338	18
24	350	18	344	18	347	18	345	18	350	18	344	18
25	356	18	350	18	353	18	351	18	356	18	350	18
26	362	18	356	18	359	18	356	18	362	18	356	18
27	368	17	362	18	365	18	362	18	368	18	362	18
28	373	17	368	18	371	18	368	18	373	18	368	18
29	379	17	373	18	377	18	374	18	379	18	374	18
30	385	17	379	18	382	18	380	18	385	18	380	18
31	390	17	385	18	388	18	386	18	391	18	385	18
32	396	18	391	18	394	18	392	18	396	18	391	18
33	402	18	397	18	400	18	398	18	402	18	397	18
34	408	18	403	18	406	18	404	18	408	18	404	18
35	413	18	409	18	413	19	410	19	414	18	410	19
36	420	18	416	19	419	19	417	19	421	19	417	19
37	426	19	422	19	426	19	424	19	427	19	423	19
38	432	19	429	19	433	20	431	20	434	19	430	19
39	439	19	436	19	440	20	438	20	441	20	437	20
40	446	19	443	20	447	20	445	20	449	20	445	20
41	453	20	450	20	455	21	453	21	457	21	453	21
42	461	20	458	20	464	22	462	21	465	21	461	21
43	468	21	466	21	473	22	470	22	474	22	469	22
44	477	21	474	21	482	23	479	22	483	23	478	22
45	485	22	483	22	492	24	489	23	493	23	488	23
46	494	22	492	23	503	25	499	24	504	24	498	24
47	504	23	502	23	515	26	510	25	515	25	509	25
48	514	24	513	25	528	27	522	26	527	26	520	26
49	526	26	524	26	542	28	536	27	540	27	533	27
50	539	28	538	28	558	30	550	29	555	29	548	29
51	555	31	555	32	576	33	567	32	572	32	565	32
52	576	37	577	38	600		589	38	594	38	588	38
53	600		600		600		600		600		600	
54	600		600		600		600		600		600	

REFERENCES

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

Andrich, A., & Luo, G. (2004). *Modern measurement and analysis in social science*. Perth, Western Australia: Murdoch University.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (CASMA Research Report No. 9). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. (Available from http://www.education.uiowa.edu/casma).

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

Cronbach, L.J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement, 25,* 291–312.

Divgi, D. R. (1980). *Dimensionality of binary items: Use of a mixed model.* Paper presented at the annual meeting of the National Council on Measurement in Education. Boston MA.

Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessments. *Perspectives: International Dyslexia Association, 27*(4), 22–25. http://www.cast.org/udl/Dolan_IDA_Perspectives_2001.htm

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141–151.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd Ed.). New York: Springer-Verlag.

Linacre, J. M. (2006). *WINSTEPS: Rasch measurement computer programs*. Chicago: Winsteps.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* New York: Erlbaum Associates.

Lunz, M. E., & Bergstrom, B. A. (1991). Comparability of decisions for computer adaptive and written examinations. *Journal of Allied Health, 20*, 15–23.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research, & Evaluation, 10*(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence.

Thissen, D. (1990). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized Adaptive Testing: A primer* (pp. 161-185). Hillsdale: NJ: Lawrence Erlbaum.

Webb, N. L. (2005). *Web alignment tool (WAT) training manual*. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.