NCSC 15

National Center and State Collaborative 2015 Operational Assessment Technical Manual

March 2016

National Center and State Collaborative
2015 Operational Assessment
Technical Manual

The federally funded NCSC project will come to an end in the fall of 2016. Project partners are supporting the transition of all resources to the Multi-State Alternate Assessment (MSAA/NCSC) states. For more information on the content of this report, please contact either the MSAA states at MSAA@AZED or, for more general information on the NCSC project and reports, contact the National Center on Educational Outcomes at the University of Minnesota at NCEO@umn.edu.

Many of the design, development, and evaluation activities leading to this technical report were supported by a grant from the U.S. Department of Education, Office of Special Education Programs (H373X100002, Project Officer: Susan.Weigert@ed.gov). The contents do not necessarily represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

National Center and State Collaborative
2015 Operational Assessment
Technical Manual

Table of Contents

See Appendices in accompanying PDFs and PDF Portfolio Binders.
TIPS: The Portfolio Binders are the easiest ways to find a specific Appendix. If you have difficulty opening the binders, check to be sure you have the most current version of Adobe Flash and all necessary plug-ins. The Chrome browser will substitute its own pdf software unless you disable it or use another browser. You may also request one file pdf options from msaa@azed.gov or nceo@umn.edu

Chapter 1-A PDF Portfolio Binder: NCSC Policy Briefs 1-9

Chapter 2 PDF Portfolio Binder:

CHAPTER 1: INTRODUCTION TO THE NCSC SYSTEM

In late 2010, the National Center and State Collaborative (NCSC)[1] began development of the NCSC Alternate Assessments based on Alternate Achievement Standards (AA-AAS) for students with the most significant cognitive disabilities. The foundations for the NCSC AA-AAS had been laid the previous decade, through a series of collaborative research to practice projects that allowed the collaborating states and national centers to understand better how to measure academic achievement for students with significant cognitive disabilities. Research done in these previous projects meant that the state leaders and national center experts who shaped the NCSC project design process and work plan were able to build on a research-based foundation, but were driven by issues and concerns that were still unanswered in 2010. These issues and concerns, and the answers derived over the course of the NCSC AA-AAS design, development, and implementation, are important context to the evidence of the technical quality of the innovative system that is presented in this report.

The end point of the NCSC AA-AAS is straightforward. The 2015 administration of the NCSC AA-AAS included:

- Assessments in Mathematics and English language arts (ELA), which includes both reading and writing, for grades 3-8 and 11;
- Around 30-35 operational items for each subject, mostly selected response;
- Direct student interaction with online testing program or the teacher may print out testing materials and enter student responses into the computer;
- Approximately 1.5 – 2 hours for each assessment (math and ELA), permitting smaller time slots over a 6-8 week period to meet the student's needs.

However, the choices that needed to be made in the design and development of the assessment system to arrive at that end point were complex.

This introduction to the NCSC AA-AAS 2015 Technical Manual provides an overview of the:

- Foundational understanding on which the project was built;
- Shared philosophy and guiding principles for project design, development, and implementation;
- Theoretical and practical tools as context to the technical documentation;
- Shared purposes of the assessment system and uses of the assessment information.

Although NCSC partner states intend to continue development into the future, this report is limited the status of the NCSC AA-AAS system as operationalized in Spring 2015.[2]

[1] The NCSC state partners participating in the spring 2015 NCSC operational assessment are: Arizona, Arkansas, Connecticut, District of Columbia, Idaho, Indiana, Pacific Assessment Consortium (Commonwealth of the Northern Mariana Islands and Guam), Maine, Montana, New Mexico, Rhode Island, South Carolina, South Dakota, and US Virgin Islands.

[2] The shared purposes and uses are further explicated in a companion Brief, including longer term outcomes and connections to both assessment and instruction, along with overarching content claims and assessment specific claims (See Appendix 1-A, NCSC Brief 9, *NCSC's Theory of Action and Validity Evaluation Approach*).

FOUNDATIONAL UNDERSTANDING

State practice in alternate assessment had been well documented through state surveys of practices in the early 2000s, but many issues and concerns remained unresolved. Based on documentation of practices, state practitioners and their expert measurement, special education, and curriculum partners remained concerned about several issues identified over the previous decade:

- the transparency of methods and results in academic instruction and in assessment;
- the integrity of design options to protect the inferences from alternate assessment while providing a balance of flexibility and standardization as required for the population of student;
- the limited existence of validity studies to systematically understand where inferences are supported; and
- the need for planned improvement over time, as both instructional improvements and assessment design improvements are implemented (Quenemoen, 2009).

The partners involved in the NCSC project had already collaborated on a series of projects to address these needs, and Table 1-1 shows the primary projects and the key outcomes of these projects completed prior to development of the NCSC project.

Table 1-1. Outcomes of Previous Collaborative Projects Addressing Technical Adequacy of AA-AAS

Project	Key Outcomes
New Hampshire Enhanced Assessment Initiative (NHEAI) 2004-2008	• Model framework to document the technical characteristics of alternate assessments (Marion & Pellegrino, 2006) with companion workbook format with key questions and content to be addressed as the test is developed, implemented, analyzed, and continuously improved. • Using the assessment triangle of cognition, observation, and interpretation as the foundational conceptual framework, NHEAI and NAAC researchers, experts, and partner states developed and began testing this technical framework (NHEAI, NAAC, & NCIEA, 2006).
National Alternate Assessment Center (NAAC) 2005-2011	• Developed tools and instruments to describe the characteristics of the student population and how they achieve competence in academic domains, the Learner Characteristics Inventory (Kearns, Towles-Reeves, Kleinert, & Kleinert 2006). • Addressed importance of communicative competence for students to access the general curriculum, transition outcomes, and college and career readiness (Browder, Flowers, & Wakeman, 2008). • Outlined a Learning Progression Framework (LPF), a path of concept and skill acquisition across grades for typical learners to guide curriculum, instruction, and assessment development for this population of students, since no common path existed for students with significant cognitive disabilities (Hess, 2010; Hess, 2011).
NAAC Validity	• Expert panelists and evaluators expanded validity analyses based on NAAC/NHEAI technical documentation tools, including making meaning of

GSEG 2007-2011	the technical characteristics of AA-AAS in the interpretive arguments through a theory embedded process (Marion & Perie, 2009).

The technical understanding and tools that resulted from these projects provided guidance on best practice procedures to document the iterative design, development, and implementation of AA-AAS, but stopped short of advising on specific design characteristics. In 2010, the organizational and state partners once again came together to articulate how they would use these procedures to build a defensible AA-AAS from the ground up that would address remaining issues and concerns. They committed to a research-to-practice approach, building in iterative design-test-redesign features in all major design and development activities. The partners also were committed to following the data as choices were made – whether extant research, small focused try-outs, or systematic piloting of resources prior to use – in a collaborative consensus-building model.

The development plan included five major phases, generally sequential but with iterative research-based feedback loops across phases:

- Content Model Phase – define a model of domain learning in mathematics and English language arts for these students; identify prioritized content for assessment.
- Principled Design Phase – develop a set of Design Patterns and Task Templates for prioritized content; develop and pilot curriculum, instruction, and professional development resources based on the same content model; design technology architecture.
- Item and Test Development Phase – try out and revise the Task Templates; item specifications and item development; item reviews and revisions; draft performance level descriptors; finalize pilot/field test design; build technology platform.
- Pilot, Field, Research Phase – conduct a two-phase pilot/field test to generate item statistics and refine items, along with item evaluation studies, student interaction studies (cognitive labs); finalize blueprints, finalize administration procedures, training, and supports.
- Operational Phase – conduct the first operational administration, standard-setting, reports, and technical reporting.

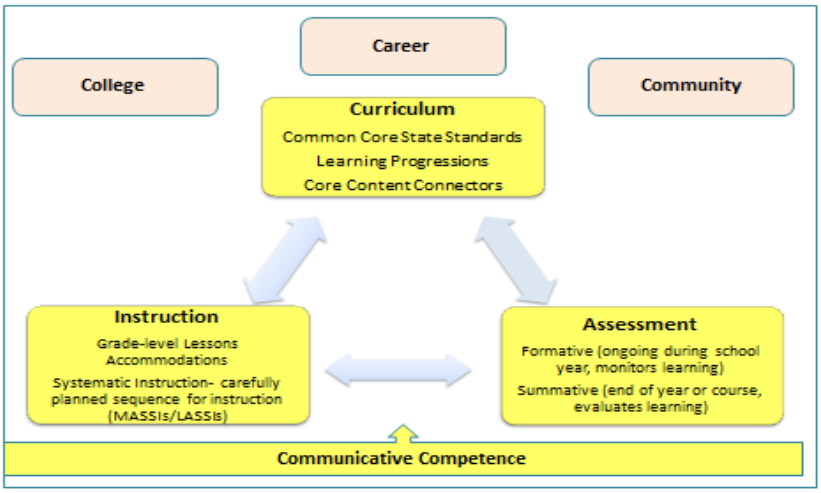
As states and organizational partners implemented the NCSC development plan, they found they had to come to consensus on topics that were a mix of practice and theory in the comprehensive context of teaching and learning for the students. They required a blend of policy, educational, and technical solutions. Through policy discussions and in iterative research and design steps, the partners arrived at a shared philosophy and guiding principles that are reflected in the overall project resources, including the comprehensive system of curriculum, instruction, classroom assessment, and professional development, as well as in the operational assessment design.

SHARED PHILOSOPHY AND GUIDING PRINCIPLES

The expert and state partners articulated a common philosophy to guide project work. It is based in a comprehensive system of curriculum, instruction, and assessment built on a common model of learning in the academic domains. Although the project Theory of Action (TOA, See NCSC Brief 9, *NCSC's Theory of Action and Validity Evaluation Approach*, Appendix 1-A) shows the theoretical interconnections of the comprehensive system, Figure 1.1 shows that the practical connections are based

on an assumption of student communicative competence, with concrete resources to support increased capacity in the field, leading to improved student outcomes in college, career, and community.

Figure 1-1. NCSC Comprehensive Systems Approach



NCSC partners believe that accessibility is central to the validity argument of the AA-AAS, and that accessibility to the academic content based on college- and career-ready academic standards begins with rigorous curriculum and instruction resources and training to teachers. Given the limited availability of these resources, the project committed to developing them. The design of NCSC curriculum and instruction resources was informed by extant research and iterative small studies to ensure inclusive accessibility and appropriately high expectations for learning.[3] Then, the NCSC assessments were based on the same model of learning as reflected in classroom resources.[4] Finally, the NCSC project provided resources for intervention on communicative competence to ensure all students have a way to first learn, and then show what they know on the NCSC assessment.[5]

THEORETICAL AND PRACTICAL TOOLS AS CONTEXT TO NCSC TECHNICAL DOCUMENTATION

To help clarify the NCSC foundation and vision, the NCSC partners developed their TOA to explain the bases for the NCSC resources and how they were intended to relate to one another, to college and career ready academic standards, and ultimately to the goals of having all students with significant cognitive disabilities leave high school ready to participate in college, careers, and their communities. The NCSC TOA and the NCSC approach to a comprehensive validity evaluation are further described in NCSC Brief 9 presented in Appendix 1-A. The NCSC validity evaluation has been developed as a

[3] See https://wiki.ncscpartners.org/index.php/Main_Page for project publicly available curriculum, instruction, and professional development resources.
[4] See NCSC Brief 3, *AA-AAS: How Do Our Students Learn and Show What They Know?* and NCSC Brief 6, *NCSC's Age- and Grade-Appropriate Assessment of Student Learning* in Appendix 1-A.
[5] See NCSC Brief 4, *Promoting Communication Skills in Students with Significant Cognitive Disabilities* in Appendix 1-A.

working document, with ongoing and systematic analyses of outcomes observed over the initial years of the operational program, as comprehensive evidence becomes available.

Practice-focused summaries of the foundational components reflected in the AA-AAS design are available to orient readers to the larger context of the comprehensive NCSC system of curriculum, instruction, assessment, and professional development. These are the NCSC Brief series, listed below, included in Appendix 1-A, and referenced as appropriate throughout the technical manual. They include:

NCSC Brief 1: AA-AAS: Standards That Are the "Same but Different." This Brief addresses what it means to have the assessment based on the same grade-level content standards but different achievement standards from those on which the general assessments are based and provides definitions and examples of each.

NCSC Brief 2: AA-AAS: Defining High Expectations for Students with Significant Cognitive Disabilities. This Brief shows state data that highlight the low expectations defined for AA-AAS in the past, and presents recent evidence from educators that highlights the need to define higher expectations for students with significant cognitive disabilities.

NCSC Brief 3: AA-AAS: How Do Our Students Learn and Show What They Know? This Brief presents the conceptual model of learning and understanding that was the basis for the development of the NCSC mathematics and English language arts resources.

NCSC Brief 4: Promoting Communication Skills in Students with Significant Cognitive Disabilities. The lessons learned during NCSC system development work, combined with research, demonstrate the importance of promoting communication skills in students with significant cognitive disabilities. This Brief provides an overview of the lessons learned and describes publicly available NCSC training and resources for improving communication.

NCSC Brief 5: Standards-based Individualized Education Programs (IEPs) for Students Who Participate in AA-AAS. Decision making at the IEP level determines who participates in AA-AAS, and how these student access the grade-level curriculum. This Brief provides guidelines and examples from the NCSC resources to use when creating standards-based IEPs for students in AA-AAS.

NCSC Brief 6: NCSC's Age- and Grade-Appropriate Assessment of Student Learning. This Brief highlights the systematic approach taken by NCSC to develop an assessment of learning appropriate for students with significant cognitive disabilities. It describes how its items were created to provide an age- and grade-appropriate assessment of student learning.

NCSC Brief 7: NCSC's Content Model for Grade-Aligned Instruction and Assessment: The Same Curriculum for All Students. The purpose of this Brief is to highlight the path NCSC followed to produce an assessment and models of curriculum and instruction that are grade-aligned with alternate achievement of enrolled grade content.

NCSC Brief 8: Characteristics of Students with Significant Cognitive Disabilities: Data from NCSC's 2015 Assessment. The purpose of this Brief is to summarize the Learner Characteristics Inventory (LCI) data collected by NCSC during its operational assessment in Spring, 2015. LCI data show how teachers currently describe their students. LCI data can flag unusual patterns in the numbers which may indicate the need for additional investigation and to document change over time as educators better understand how to present grade-aligned instruction and assessment and communication intervention to these students.

NCSC Brief 9: NCSC's Theory of Action and Validity Evaluation Approach. This Brief explicates the NCSC Theory of Action and validity evaluation approach, developed for communicating about the system and its components, obtaining feedback during the development process to allow continuous improvements, and evaluating the system. NCSC's Theory of Action is an essential part of the NCSC system. It helps answer fundamental questions about how the NCSC system is meant to work. Into the future, it can guide evaluation and understanding of how well the system is achieving its ultimate goal as well as identify needed revisions and improvements as the system continues to evolve.

The remainder of this technical manual will describe the technical characteristics of the 2015 NCSC Operational Assessments, as part of the comprehensive NCSC system:

Chapter 2: Test Development

Chapter 2 describes the NCSC Principled Design approach to assessment development along with a description of development research and decision-points; a description of the technology platform; and a focus on defining the measurement construct through prioritization of grade-level content, developing the item model, and iterative development of Design Patterns and Task Templates and related studies. Item development process and results, item reviews for content, bias and sensitivity, and universal design/accessibility procedures are described. Special studies to address needs of the lowest incidence students are included, as well as security procedures in place to protect integrity of items during development and review. The family of studies to augment initial pilot testing, including cognitive labs (called Student Interaction Studies) are included. The two-phased pilot approach to refining the system is described, including procedures and outcomes of item data reviews. Finally, this chapter addresses form assembly, quality control, alignment studies in addition to those covered in Chapter 3, and embedded field test items.

Chapter 3: Alignment and System Coherence

NCSC's approach to system coherence focuses on understanding the connections among components of the NCSC system throughout design and development. There are key points in the design and development process for checking the assumption that the NCSC AA-AAS will yield scores that support intended interpretations and uses. To this end, NCSC developed processes and designed studies to provide evidence supporting the formative and summative evaluation of this key assumption. Chapter 3 includes information collected across these studies to support evaluation of the relationship between (a) the larger learning and instructional context through which students have the opportunity to learn and (b) the current assessment context as defined through the academic grade-level content prioritized for assessment. More specifically, Chapter 3 includes evidence about the relationship between the academic

grade-level content prioritized for the NCSC assessment and the college and career ready standards intended as the reference for NCSC's system.

Chapter 4: Test Administration
Chapter 4 on Test Administration provides an overview of the focus and general parameters guiding test administration for the operational test. This includes administration procedures and guidelines, test security, administration procedures and manual, accommodations procedures and manual, administration support, and support services. It addresses training of administrators and coordinators, certification procedures, monitoring and quality control, and a description of the results, including irregularities and implications and analyses of how challenges were addressed.

Chapter 5: Scoring
Chapter 5 provides an overview of scoring processes and rules for items in mathematics and reading (selected response, multi-part, open response, constructed response). It includes processes and rules by item type, along with training and monitoring procedures.

Chapter 6: Psychometric Analyses
This chapter provides an overview of the psychometric analyses of the operational test data. It includes classical item analyses and analyses based on the application of item response theory (IRT) modeling techniques. The chapter also includes the process used to establish the NCSC scale and the methods used for equating multiple forms used in each grade and content area
.

Chapter 7: Standard Setting
Chapter 7 covers the rationale for selecting the standard setting method, as well as a full description of the standard setting process. It includes the standard setting PLDs, panel composition and structure, reporting forms and standard setting forms, rating process, round by round implementation, and panelist evaluations. Ratings and results, vertical articulation, cross content area coherence, and recommendations are included. Policy decisions and rationales are provided. External evaluation data are also provided.

Chapter 8: Studies of Reliability and Construct-Related Validity
Chapter 8 includes an in-depth look at reliability, including test-level reliability using classical test reliability and the standard error of measurement. Dimensionality and generalizability analyses are provided, as are post-hoc bias and sensitivity analyses.

Chapter 9: Reporting Interpretation and Use
Chapter 9 describes the process used to guide report development, and to shape the interpretation and use of all reports. It includes detailed discussion on how state partners and the vendor worked together to meet unique state needs while also adhering to core NCSC requirements for reporting, and ultimately for interpretation and use. It describes the primary reports (Student Reports; School Roster Reports; School, District, and State Summary Reports), along with state by state variations and amendments to the common reports. Quality assurance procedures are also described. The Guide to Score Report Interpretation is included in a chapter appendix, along with other supporting documentation.

SHARED STATE PURPOSES AND USES OF THE NCSC AA-AAS

The NCSC assessment is intended to support comprehensive long-term project goals in several ways.

- In order to permit educators and parents to track student progress toward college, career, and community readiness, it measures students' academic achievement.
- It yields defensible scores that can be used for school accountability decisions and program evaluation.
- It provides reporting structures that support appropriate interpretation and use of data in support of improving practices that will result in higher student achievement.

The primary NCSC claim is that the NCSC scores provide information that reflects what students know and can do in relation to the academic expectations defined in state academic content and achievement standards.

Additionally, through the process of administering the assessment, teachers have the opportunity to improve their skills in communicating with and instructing their students as well as learning more about what their students know and can do academically. Unlike the assessment administration process for general assessments, alternate assessments for students with significant cognitive disabilities typically involve interactions between an individual student and the teacher/test administrator for item presentation and for recording responses. Thus, from the assessment administration, teachers have the opportunity to gain deeper insight into the academic expectations for their students and into their students' knowledge and skills, which enhances and clarifies what they may learn from the assessment scores.

Uses as part of federal and state accountability and stakes associated with assessment results

Student scores on NCSC assessments can be interpreted as reflecting the knowledge and skills defined by college, career, and community readiness standards and are intended to provide useful information for tracking student progress toward achieving the knowledge and skill in their enrolled-grade standards. They may be used to provide information to teachers to guide instruction toward these standards and to allow educators and parents to track student progress toward college, career, and community readiness. They may also be informative within school accountability systems as well as part of a program evaluation.

Limitations

Since there are so few students who take alternate assessments and because these students vary widely in both characteristics and academic skill level, care should particularly be taken when using student NCSC scores in making high-stakes decisions. This is especially true when these scores are used in isolation to make decisions. When aggregating data within schools and districts, care should be made to ensure that a sufficient number of students' scores are included prior to making any decisions, high-stakes or not.

CHAPTER 2: TEST DEVELOPMENT

Test development encompasses the processes used to design and create test items as well as the methods used to combine items into the test forms that students and teachers see when the teacher administers the test with a student.[6] This chapter addresses the following questions from the National Center and State Collaborative (NCSC) Interpretative Argument (IA) (see NCSC Brief 9, *NCSC's Theory of Action and Validity Evaluation Approach,* Appendix 1-A):

- Do the content and skills assessed by the test items reflect the grade- level college and career ready standards?
- Do students engage in the thinking processes considered necessary to solve test items while they are answering those items?
- Are administration procedures and data-capture methods standardized in ways that support comparability across students, schools, and time?
- Are the administration procedures and data-capture methods flexible enough to allow students to demonstrate what they know and can do?
- Do the decisions about how student responses are scored result in accurate and meaningful differences across performance levels?

To begin answering these questions, this chapter provides a wide range of information on the test development process. This information includes descriptions of the test development steps for English language arts (ELA) and mathematics in grades 3–8 and 11. It describes how the NCSC consortium addressed the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014) in ways that best support the research-based perspective on how students with the most significant cognitive disabilities think, learn, and demonstrate what they know and can do. This chapter explicates the test development processes and activities by presenting detailed information regarding the following:

- The general approach to test design;
- Defining the measurement construct;
- Developing the item model;
- Passage and item/task development process and results;
- Design and implementation of Pilot Phase 1 and Phase 2 testing;
- Design of the technology platform;
- Pilot Item Data Review process and results;
- Report of additional studies and surveys completed as part of iterative design work;
- Test structure and form assembly for the spring 2015 operational tests.

[6] NCSC partners engaged multiple vendors as partners to the work. They included: several item writing vendors (principled design/task template development and actual item bank development), technology architecture design vendor, pilot development vendor, operational administration vendor, and assessment technology vendor. They are referred to generically in this document by their role.

General Approach to Test Design

Introduction

This section addresses the IA questions focused on (a) whether the content and skills assessed by the test items represent college and career ready standards, and (b) whether students responding to test items engage in the thinking processes considered necessary to solve the test items.

Expectations for students on alternate assessments developed and administered in the late 1990s and early 2000s reflected a prevalent belief that students with significant cognitive disabilities could not learn academic content or could only learn very basic skills (See NCSC Brief 2: *AA-AAS: Defining High Expectations for Students With Significant Cognitive Disabilities,* Appendix 1-A.). Research over the past two decades has countered this belief by providing powerful insights into how students with significant cognitive disabilities represent knowledge and develop competence in specific domains (See NCSC Brief 3: *AA-AAS: How Do Our Students Learn and Show What They Know?,* Appendix 1-A*)*. In addition, experts in the fields of special education and communication science have provided models to support how learning opportunities and assessment tasks can be designed to provide evidence for inferences about what students know and can do across a full range of performance (Kleinert et al., 2009).

NCSC approached designing the NCSC alternate assessments based on alternate achievement standards (AA-AAS) to reflect these insights and began with an informed understanding of the characteristics of the population of students with the most significant cognitive disabilities and how they represent knowledge and develop competence in specific domains. The NCSC assessment constructs and content-model definitions assume students can learn (a) when given the opportunity to learn the expressed academic knowledge of the ELA and mathematics college and career ready standards of their enrolled grade and chronological age, and (b) when the prioritized assessment constructs focus on the critical content for progressing from grade to grade and use structured scaffolds and supports that do not interfere with the measurement of the content.

The Assessment Triangle (National Research Council, 2001), serving as a foundation for the development of educational assessments for all students, provides a process of reasoning from evidence about what students know and how they know it. The NCSC test design applied the elements of the Assessment Triangle to ensure that the NCSC AA-AAS reflects reliably and validly what students with the most significant cognitive disabilities have learned in the presence of evidence-based academic instruction. The design of the NCSC assessment began with a specific understanding not only of which knowledge and skills were to be assessed, but also of how students with the most significant cognitive disabilities develop competence in mathematics and English language arts.

The student learning model rests in the cognition vertex of the Assessment Triangle. This is defined as a "theory or set of beliefs about how students represent knowledge and develop competence in a subject domain" (National Research Council, 2001, p. 44). The systematic approach taken by NCSC to develop an assessment of learning that is appropriate for these students was implemented through the application of the observation vertex. NCSC item development for an assessment of age- and grade-appropriate student learning required that specific items were designed to show what students know and can do based on the precise aspects of cognition of prioritized grade-level academic content.

The application of the cognition and observation elements resulted in an assessment development process that was based on (a) what we have learned about how these students build competence in academic content and (b) evidence of that learning across a range of performance collected by assessment items that combine a gradual increase in complexity of content and a gradual decrease in the types of built-in scaffolds and supports. The NCSC assessments engage students to think in ways that mirror developers' understanding of how these students learn academic content and show what they know (See NCSC Brief 6: *NCSC's Age- and Grade-Appropriate Assessment of Student Learning,* Appendix 1-A).

The NCSC AA-AAS serves three main purposes: (1) to measure student achievement; (2) to provide defensible scores for state accountability systems; and (3) to provide reporting structures that support appropriate interpretation and use of data. The NCSC partnership designed the NCSC AA-AAS to capture student performance at different levels of skill acquisition. The assessment items incorporate important aspects of item design related to both varying levels of content complexity and the degree and type of scaffolds and supports. NCSC's intentional assessment development process addressed the targeted grade-level academic content linked to evidence-based curricular and instructional materials, and resulted in useful information for educators and families.

NCSC underscored the assessment development process for ELA and mathematics with two fundamental premises: (a) assessment is part of a broader, comprehensive system that includes curriculum, instruction, and professional development; and (b) assessments for students with the most significant cognitive disabilities rely on a foundation of communicative competence (see Chapter 1, Figure 1-1.). To demonstrate what they know and can do, students must have an established and clear method of communication that allows them to indicate clear answers to test questions. Integrated into the development of the NCSC assessment system are components that help teachers and other educators identify students who need intervention/support. As described in NCSC Brief 4, *Promoting Communication Skills in Students with Significant Cognitive Disabilities* (Appendix 1-A), partners also created a comprehensive Communication Tool Kit to support teachers' capacity to build students' communication skills and to provide teachers and other educators with suggestions and resources for communication intervention.

NCSC content and special education state and center experts focused on defining the constructs of reading, writing, and mathematics to reflect an appropriate expectation of instruction and learning throughout a student's educational experience and to make those constructs adaptable to the way in which students with significant cognitive disabilities demonstrate acquired knowledge and skills. NCSC established overarching content definitions by examining (a) existing content definitions in general education; (b) the content, concepts, terminology, and tools of each domain; (c) a body of extant research; and (d) the Common Core State Standards (CCSS; see NCSC briefs 3, 6, and 7 for further details, Appendix 1-A). These content definitions became central to the development of assessment items.

NCSC developers revised and refined the NCSC AA-AAS design using cycles of continuous feedback from state and center partners. Developers evaluated proposed designs through iterative item and test development steps, special studies, and pilot testing, all of which were central to the final NCSC assessment model implemented through the first administration of the operational test in spring 2015.

Principled Approach to Assessment Development

To create the framework for assessment item development, NCSC and a NCSC item development partner used a principled design approach that incorporated components of evidence-centered design (ECD) and aligned with college and career ready standards. They applied this approach to the development of items in ELA and mathematics with a continual focus on developing a system that was flexible enough to meet the needs of students with the most significant cognitive disabilities but that protected content to be assessed. Their implementation design was framed by the following:

- An articulated model of student learning that informed the full range of performance observed;
- The characteristics of the students to be tested;
- Attention to system coherence.

NCSC and its item development partners used components of ECD to translate academic content into test items and to develop design patterns and task templates aligned with the college and career ready standards. The design patterns and task templates identified the constructs and prerequisite knowledge and skills needed to perform successfully on assessment items that addressed prioritized grade-level academic targets (Mislevy & Haertel, 2006). More specifically, design patterns incorporated a variety of approaches to obtaining evidence of targeted knowledge or skills and supported development of task templates that provided templates for tasks or items that defined evidence, task models, and item specifications. (See Appendix 2-A Item Specifications Reflected in Example Annotated Design Pattern and Task Template.)

The NCSC item development partner employed its Principled Assessment Design for Inquiry (PADI) online assessment design system, which was built on a conceptual framework that is based on the tenets of ECD. PADI draws on new understandings in cognitive psychology and recent advances in measurement theory and technology to create the framework used to design assessments. It includes design components implemented by NCSC (e.g., design patterns and task templates) as well as the articulation of student, evidence, and task models, which were used to guide the development of the design patterns, task templates, and, ultimately, the NCSC item bank.

Universal Design for Learning Principles

According to the *Standards for Educational and Psychological Testing*, tests should be designed to minimize construct-irrelevant barriers for all test takers in the target population (AERA, APA, & NCME 2014, pp.6–7). Universal Design for Learning (UDL) seeks to optimize the accessibility of educational materials and assessments while minimizing separate-but-equal situations. To allow the widest possible range of students to demonstrate what they know and can do and to be able to make valid inferences about the performance of all students who participate in an assessment, universally designed assessments are developed from the beginning with an eye toward maximizing fairness (Johnstone et al., 2006). NCSC development partners applied their understanding of the characteristics of this student population and UDL principles to inform the design of each item. Their focus was to ensure that any necessary additional adaptations and accommodations did not interfere with the measured construct. A strength of the NCSC AA-AAS ECD-based approach was the support it provided for the development of items that (a) focused on construct-relevant content (the knowledge, skills, and abilities intended to be

measured), (b) minimized the impact of construct-irrelevant skills (e.g., inability to read text due to size of print, inability to access items due to absence of assistive device, inability to engage with the items), and (c) considered appropriate accessibility options (Cameto, Haertel, Morrison, & Russell, 2010, p. 1). In addition, NCSC provided flexible materials, techniques, and strategies for instruction and assessment to address the needs of students with the most significant cognitive disabilities (Dolan, Rose, Burling, Harms, & Way, 2007).

Accessibility and Item Features

NCSC created and adopted policies for accessibility and item features that resulted in flexible assessment design and delivery (i.e., computer-based or paper-based) and provided opportunities for all students to show what they know, while incorporating other important aspects of item design such as depth of knowledge, text complexity, and degree and type of scaffolds and supports.

Using a principled design approach, NCSC strove to minimize accessibility challenges by taking into account test characteristics, such as the choice of content, response processes, and testing procedures, that may impede test takers' access to the construct. Specifically, each design pattern included information that informed the task model by specifying design features of tasks. Examples of such information include a description of features that must be present to elicit the focal knowledge, skills, and abilities (focal KSAs) as well as additional KSAs and nonconstruct-relevant KSAs that may be required for successful performance on tasks associated with each design pattern.

For example, the variable features section of some mathematics design patterns included options for guiding exploration and information processing. These features included, but were not limited to, modeled prompts. (A modeled prompt demonstrates the process or procedures needed to complete an item successfully, but it does not provide the correct answer for the item on the test.) In addition, the design patterns described (a) task features used to support cognitive background knowledge and (b) student abilities associated with executive functioning, engagement, perceiving task stimuli, expressing responses to tasks, comprehending linguistic components of tasks, and processing information. These features were used to inform specific accessibility strategies that were built into item by item directions for test administrators, as well as overall accommodations policies. (See Appendix 2-B Accessibility By Design.)

NCSC used the final design patterns as the mechanism to implement the varying levels of content difficulty in the resulting task templates, which included items measuring a particular aspect of the prioritized academic content in mathematics, reading, and writing. Assessment developers used the task templates to develop the NCSC item bank.

The small but very diverse population with multiple and complex disabilities presents those designing alternate assessments with a range of student accessibility challenges (Kearns, Kleinert, Towles-Reeves, Kleinert, & Thomas, 2011). To address these challenges, NCSC formed an accessibility committee to ensure that the accessibility features students receive on NCSC assessments provide a valid reflection of what students who are blind, deaf, deaf/blind, nonverbal, and/or who use augmentative and alternative communication (AAC) know and can do without altering the measured constructs. This committee used an iterative process that provided specific guidance to inform the development of

enhanced item protocols, special forms, and item presentation (e.g., braille). Input and feedback was elicited from NCSC state and organizational partners, special education teachers, disability and communication science experts, the American Printing House for the Blind (APH), and the Perkins School for the Blind.

NCSC then developed the *Procedures for Assessing Students Who Are Blind, Deaf, or Deaf-Blind: Additional Guidance for Test Administration,* which outlined accessibility guidelines and testing procedures. To assess reading foundational skills (e.g., word recognition), NCSC developers created verbal, nonverbal, and braille versions of the foundational skills items. They also developed the *Augmentative and Alternative Communication Guidelines for Writing Constructed-Responses* to provide test administrators (TAs) with assessment administration protocols to use with other testing materials to ensure accurate administration of the field-tested writing constructed-response (CR) items to students who respond using AAC.[7]

By using a rigorous and replicable assessment design process that carefully considered how content, task, and knowledge about learner characteristics interact in the creation of assessment tasks, the NCSC consortium directly addressed the most critical issues related to the varied characteristics of students with significant cognitive disabilities and how to measure their achievement accurately.

Graduated Complexity

NCSC uses the term "graduated complexity" in the NCSC AA-AAS to refer to the relationship among four items (a family of items) associated with a single focal KSA. The four items purposely vary from most to least difficult along two dimensions: (1) the difficulty of the content presented to the student (e.g., magnitude of numbers used in math, and length and complexity of text in ELA reading); and (2) the number and type of features applied to support student performance (e.g., definitions, demonstrations, graphic organizers in both math and ELA).

Students have the opportunity to attempt a full range of item complexities in the NCSC test design. The assessment includes items with multiple levels of complexity and varying degrees of scaffolds and supports to provide opportunities for students to show what they know and can do at varying levels of understanding. NCSC systematically varied item complexity, in accordance with design principles, to create item families that allow students to demonstrate what they know and can do across the range of learning observed in classroom instruction with respect to NCSC grade-level academic targets, called Core Content Connectors (CCCs, see NCSC Brief 7: *NCSC's Content Model for Grade-Aligned Instruction and Assessment: The Same Curriculum for All Students,* Appendix 1-A), which align to college and career ready standards.

To ensure greater accessibility for all assessed students, NCSC's item development partner selected variable features to vary levels of difficulty across an item family and to mitigate construct-irrelevant variance. More specifically, developers linked variable features to learner needs (e.g., perceptual, expressive, and cognitive) to support performance without changing the measured construct. Student cognition experts and content experts participated throughout item development to ensure the

[7] These documents include item specific information and are secure testing materials.

appropriate identification and application of variable features. Content experts, in particular, identified variable features that reduced construct-irrelevant variance and supported the production of item families with graduated complexity. NCSC developers used the variable features to create item families such that the first three levels graduate in terms of complexity and all levels provide greater support relative to the highest level 4 item.

For each grade-level academic content target prioritized for assessment, a unique design pattern defined the kinds of observations that provide evidence about acquisition of the focal KSA as well as the features of task situations that allow students to provide this evidence. Developing the exemplar items for each task template required the identification and articulation of additional KSAs for each item. This combination of focal KSAs and additional KSAs comprised the item development model associated with each particular item family.

Definition of Comprehension of Text and Reading

Reading requires at least two components: access to text and comprehension of that text. For many years, reading interventions for students with significant cognitive disabilities focused primarily on accessing text through sight-reading of functional sight words (Browder, Wakeman, Spooner, Ahlgrim-Delzell, & Algozzine, 2006). While these sight words might help individuals navigate some aspects of daily living (e.g., reading menu words), they provide no access to literature and informational text—which requires a reader to manage passages of text. Browder et al. (2008) proposed that the conceptual model for literacy for students with significant cognitive disabilities focus primarily on listening comprehension while also building the capacity for as many students as possible to learn to access text through decoding. The authors noted that because text has little purpose unless the student can gain meaning from it, decoding without comprehension will not be useful. Being able to understand a passage of text, whether the passage is independently read or accessed through technology or a human reader, is the most important goal of literacy. They proposed that for students with the most significant cognitive disabilities, the assessment of standards related to gaining meaning from text should be separated from the demands of decoding. That is, the concept of "reading" for students with the most significant cognitive disabilities should be expanded to include listening comprehension.

The application of listening comprehension as a mechanism for students to gain increasing levels of understanding and even engage with literature and informational text from the grade that matches their chronological age (i.e., grade and age-appropriate text) is supported by a body of research (see Browder, Trela, & Jimenez, 2007; Hudson & Test, 2011;Koppenhaver, Erickson, & Stotko, 2001; Mims, Hudson, & Browder, 2012; Wood, Browder, & Flynn, 2015).

NCSC adopted this perspective to inform the development of the NCSC reading assessment items and established definitions of reading that are consistent with the goal for all students – that reading is a meaning-making process. NCSC expanded the definition of literacy to include listening comprehension. NCSC state and center partners established the following definitions of reading:

Reading (foundational): Knowledge of concepts about print (e.g., reading left to right, reading top to bottom, parts of a book, identify the title), the alphabetic principle (i.e., words are composed of letters that make sounds), and basic conventions of the English writing system to pronounce or identify

words (decode text) including deciphering symbols (letters, pictures, braille); identification of sight words/symbols or irregularly spelled words.

Reading (literature and informational text): Making meaning from texts (may be adapted with picture supports) and a variety of print media (including but not limited to picture symbols) and non-print media. Text and media may be presented in conjunction with read-aloud as an accommodation unless item/instruction is designated as decoding text.

Varying Degrees of Difficulty for Literary and Informational Passages

Current college and career ready standards (e.g., CCSS and various state standards) highlight the growing complexity of the texts students must comprehend to be ready for the demands of college, career, and life. The standards call for a staircase of increasing complexity, so that all students are ready for the demands of college- and career-level reading no later than the end of high school. The standards also outline a progressive development of reading comprehension, so that students advancing through the grades are able to gain more from what they read.

NCSC content experts recognized that a variety of factors influence text complexity and that the complexity (or the reading demands) of a particular text is the result of combinations and interactions between quantitative and qualitative dimensions. For example, a text with a low Lexile score (word frequency and sentence length) may contain mature themes and complex ideas. This type of interaction coincides with descriptions included in college and career ready standards documentation. For example, the CCSS describes a three-part model for measuring text complexity: (1) qualitative evaluation of the text, (2) quantitative evaluation of the text, and (3) matching reader to text and task (from *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*).

The NCSC ELA literacy model for comprehension of text focuses on understanding age- and grade-appropriate text read aloud to the student by a technological device or a human reader. Texts written across a range of complexity levels provide an opportunity for students with varying ranges of acquired reading skills to answer text-based questions.

NCSC partners established an expectation that the reading literacy acquired by students with the most significant cognitive disabilities would be observed across the full range of expected performance. To meet this expectation, developers created grade- and age-appropriate sets of four passages (one passage associated with each item in an item family) ranging from most complex to least complex to assess students' attainment of comprehension and vocabulary skills. Developers designed the most complex passages at a grade-level to approximate the qualitative and quantitative expectations for complexity for that grade-level. Conversely, passages designed as the least complex allowed students who are just beginning to interact with the academic content presented in text to show what they know with simplified text that linked to the assessed reading concepts and skills.

NCSC developers designed the context of each passage to match the grade of the student and to be age-appropriate; therefore, context was consistent across instructional materials and assessment items for that grade. In addition, passage themes were similar to themes found in literature for a particular grade

band, and the main ideas in the informational texts were similar to those found in informational texts for a particular grade band. NCSC specified quantitative dimensions that described the traditional measures of text complexity, including the range of reading demand (i.e., Lexile), word count, approximate number of sentences, and mean sentence length. For both literary and informational texts, NCSC used qualitative dimensions of text complexity (e.g., level of inference, sentence structure, and vocabulary) to describe criteria central to a student's comprehension of text. NCSC applied the assumption that students had experience reading text described as appropriate for their grade levels as well as those for earlier grade spans.

DEFINING THE MEASUREMENT CONSTRUCT

Introduction

NCSC focused on defining the constructs of reading, writing, and mathematics to reflect an appropriate expectation of instruction and learning throughout a student's educational experience. This section presents the processes NCSC employed to define the measurement construct, and provides evidence relative to the IA questions that focused on the following:

(a) Whether the content and skills assessed by the test items represents the college and career ready standards and
(b) Whether students responding to test items engage in the thinking processes considered necessary to solve the test items.

The evidence is drawn from how NCSC prioritized academic grade-level content for assessment, compared the NCSC content claims and the initial test blueprints to the prioritized content, and the requirements expressed through the policy Performance Level Descriptors.

Prioritization of Academic Grade-Level Content

To reflect high expectations for students with the most significant cognitive disabilities, NCSC prioritized academic grade-level content for the alternate assessment. This prioritization process produced content assessed by the NCSC AA-AAS that adequately represents and appropriately measures KSAs with respect to college- and career-ready standards. In addition, NCSC prioritized the range of assessed grade-level content to reflect the expectation that students learn and make progress in the general curriculum toward more complex learning, while at the same time reducing their need for adaptations, scaffolds, and supports.

The prioritization of the assessed content and the NCSC assessment development process was based on how these students build competence in academic content (the cognition vertex of the Assessment Triangle) and the relationship between how they build competence and the development of the assessment framework, items, and tasks (the observation vertex of the Assessment Triangle). NCSC employed an intentional process to define the measurement construct (prioritized targets from NCSC defined academic grade-level content, or Core Content Connectors (CCCs, see NCSC Brief 7, Appendix 1-A, cited earlier) based on the student model of learning and influenced by the NCSC content claims; the test blueprints which reflect the development and emphasis of content in the college and career ready standards; and the NCSC policy Performance Level Descriptions (PLDs).

Using an iterative review process, NCSC state and organizational partners evaluated the ELA and mathematics academic grade-level content (see NCSC Briefs 3, 6, 7 Appendix 1-A) to determine a challenging but achievable number of grade-level KSAs that would reflect how these students move from grade to grade content with their peers without disabilities in meaningful, naturally occurring pathways. Criteria for selection of prioritized content included the importance of the content to be assessed with respect to NCSC's content-specific claims, the distribution of and alignment to the mathematics domains and ELA strands in college- and career-ready standards consistent with general assessments under development by NCSC state partners, and how that content would provide a degree of flexibility in developing test items at differing complexity levels.

NCSC content experts and severe disabilities experts guided the work of state partners to consider the following questions during the prioritization of NCSC's academic grade-level content:

- Why is this learning important? (rationale)
- How can the knowledge and skills (that have been emphasized or prioritized) collectively inform interpretations about what a student knows and can do? (claims to be interpreted and reported)
- What evidence do we need to collect to enable us to make the intended claims?
- How will we obtain that evidence from students in this population?

This process supported the selection of representative and measurable targets for the development of the NCSC assessment system aligned to grade-level academic content and the way in which students with the most significant cognitive disabilities learn and make progress in the general curriculum. The next section explicates this process with respect to a review of the mathematics and ELA claims.

Review of Claims in Relation to the Prioritized Academic Grade-level Content

The selection of the prioritized assessment content included an examination of the NCSC content claims (i.e., what constitutes readiness for post-secondary options) and an evaluation of the eligible academic content and skills to produce sufficient evidence against each claim. NCSC state and organizational partners, including measurement and content experts, examined the content claims to inform the adoption of the prioritized grade-level academic content as presented and discussed during full project meetings, assessment development team meetings, and NCSC Ad Hoc state and organizational weekly conference calls. NCSC used a review and adoption process that involved evaluating the coherence and vertical articulation of the prioritized content against the appropriate NCSC content claim(s) by examining of the following:

- The alignment between the claims and the identified prioritized content;
- The degree of representation of content coverage to support the claims;
- The progression of expectations for student performance across grades and content areas.

Review of Test Blueprints in Relation to Prioritization

In mathematics and ELA, NCSC developed preliminary test blueprints early in the development process to indicate the targeted percent distribution of content as related to the college and career ready standards in mathematics domains and ELA strands. NCSC used the preliminary blueprints as part of the

prioritization process to ensure that the targets for measurement adequately addressed and represented the intended percent distribution.

In mathematics, the college and career ready standards define the content knowledge and skills all students should have (as domains and clusters) by the end of a specific grade. Domains are the larger content strands that frame the clusters across grades. The values in Table 2-1 represent the percent distribution of items to inform prioritization for grades 3–8 and 11 as related to the mathematics domains.

Table 2-1. Target Percent Distribution of Mathematics Content

Grade level						
3	**4**	**5**	**6**	**7**	**8**	**11**
			Ratio and Proportions		**Functions**	**Algebra And Functions**
			30	40	20	
Operations and Algebraic Thinking			**Expressions and Equations**			50
30	30	10	20	10	20	
Number and Operations Base Ten			**The Number System**			**Number and Quantity**
20	10	40				
Number and Operations Fractions			30	20	10	20
20	30	20				
Measurement and Data			**Statistics and Probability**			**Statistics and Probability**
20	20	20	10	10	20	20
Geometry						**Geometry**
10	10	10	10	20	30	10

In ELA, the grades K–5 college and career ready standards for reading follow the National Assessment of Educational Progress' model of balancing the reading of literature with the reading of informational texts. NCSC used the same percentages of literature and informational texts suggested by these standards; however, for grade 8, NCSC placed a slightly greater emphasis on informational texts to support a smooth transition from grade 8 to grade 11. Thus, NCSC prioritized an increasing proportion of informational text as the prioritization process advanced through the grades.

Early in the development of the prioritization process for ELA, project partners proposed a 70/30 distribution of reading to writing targets for assessment. NCSC partners based the prioritization of the writing on an approximately even distribution of each of the three writing text types (narrative, explanatory, argument) across all grades, with a slight increase in the emphasis on argument in the higher grades.

The intended percent distribution of ELA skills allows students to exhibit their acquisition of expected knowledge and skills for reading and writing across a range of reading and writing texts. The

values in Tables 2-2 to 2-6, on the target percent distribution of ELA content, represent the percent distribution of items to inform prioritization for grades 3-8 and grade 11 as related to the ELA strands.

Table 2-2. Grades 3 and 4 Target Percent Distribution of ELA Content

Overall test form percent distribution of reading and writing						
Reading 70				**Writing 30**		
Distribution by reading context types				**Distribution by writing text**		
Text genre 70-75		Vocabu-lary 15-20	Foundational 10	Narrative 70	Explanatory 30	Argument 0
Distribution by Genre						
Literary 50	Informational 50					

Table 2-3. Grade 5 Target Percent Distribution of ELA Content

Overall test form percent distribution of reading and writing					
Reading 70			**Writing 30**		
Distribution by Reading Context Types			**Distribution by Writing Text**		
Text genre 80-85		Vocabulary 15-20	Narrative 70	Explanatory 30	Argument 0
Distribution by Genre					
Literary 50	Informational 50				

Table 2-4. Grades 6 and 7 Target Percent Distribution of ELA Content

Overall test form percent distribution of reading and writing					
Reading 70			**Writing 30**		
Distribution by Reading Context Types			**Distribution by Writing Text**		
Text genre 80-85		Vocabulary 15-20	Narrative 30	Explanatory 70	Argument 0
Distribution by Genre					
Literary 40	Informational 60				

Table 2-5. Grade 8 Target Percent Distribution of ELA Content

Overall test form percent distribution of reading and writing					
Reading 70			**Writing 30**		
Distribution by Reading Context Types			**Distribution by Writing Text**		
Text genre 80		Vocabulary 20	Narrative 0	Explanatory 70	Argument 30
Distribution by Genre					
Literary 40	Informational 60				

Table 2-6. Grade 11 Target Percent Distribution of ELA Content

Overall test form percent distribution of reading and writing	
Reading 70	**Writing 30**

<table>
<tr><th colspan="3">Distribution by Reading Context Types</th><th colspan="3">Distribution by Writing Text</th></tr>
<tr><td colspan="2">Text genre
80</td><td>Vocabulary
20</td><td rowspan="3">Narrative
0</td><td rowspan="3">Explanatory
30</td><td rowspan="3">Argument
70</td></tr>
<tr><td colspan="2">Distribution by Genre</td><td rowspan="2"></td></tr>
<tr><td>Literary
30</td><td>Informational
70</td></tr>
</table>

Use of Policy PLDs

Prioritization of NCSC's academic grade-level content targets for the NCSC AA-AAS involved partner review and approval of policy PLDs developed by NCSC content and special education experts. Policy PLDs define the overall expectations for how much and how well students perform in relation to expectations at each of four performance levels. The following considerations guided development of both mathematics and ELA policy PLDs and informed the selection of prioritized assessment content:

- A focus on the levels of complexity, depth and breadth, and the accuracy of understanding needed at each performance level;
- The need for scaffolds and some supports for students with the most significant cognitive disabilities to permit independent demonstration (avoiding changing the content being assessed);
- The level of support, interrelated with the content, helped distinguish the performance levels. For example, a student performing at the advanced level may only need minimal scaffolding, whereas a student performing at the proficient level may need to see a model, and a student performing at the basic level may require step-by-step instructions while being required to select a correct response independently at each step;
- The need to reflect content alignment;
- The need to avoid references to communication and consistency of performance; although the communication level is a significant factor influencing performance for this population, students must be able to communicate to participate in meaningful instruction or assessment in order to be able to show what they know and can do.

Adoption of Prioritized Academic Grade-Level Content

The NCSC partnership examined the coherence of evidence collected across comparisons with the claims, draft blueprints, and draft policy PLDs to ensure that the prioritized academic grade-level content and skills for assessment provided adequate representation of grade-level college and career ready standards.

The multi-step review process resulted in NCSC partner approval of prioritized assessment academic grade-level content targets, which produced sufficient evidence (i.e., content and situations) against each NCSC content claim. NCSC partners approved 10 mathematics targets per grade level, 7–9 reading targets per grade level, and 3 writing targets per grade level.

In mathematics, the final set of prioritized targets for assessment consisted of mathematics KSAs requiring students to demonstrate the following:

- The ability to carry out mathematical procedures;
- An understanding of mathematical concepts;

- The ability to model quantitative relationships;
- The ability to solve problems based on real-world situations.

In reading, the final set of prioritized targets for assessment consisted of reading KSAs requiring students to demonstrate the following:

- The use of key details to describe the central idea or theme from literary texts;
- The use of evidence to summarize or make inferences from literary texts;
- The use of key details and evidence to summarize or support the main idea from informational texts;
- The location of relevant information using text features to answer questions from informational texts;
- The determination of comparability of key ideas when making connections across informational texts (grades 5 through high school);
- The use of context to determine the meaning of general academic words or phrases or domain-specific vocabulary;
- The identification of words (grades 3 and 4).

In writing, the final set of prioritized targets for assessment consisted of the writing KSAs requiring students to demonstrate the following:

- The ability to generate a permanent product to represent and/or organize ideas or thoughts so that messages can be interpreted by someone else when the writer is not present—that is, when responding to a writing prompt, the ability to produce a Literary/Narrative, Informational/Explanatory, or Persuasive/Argument permanent product;
- The ability to include in a written product grade-specific writing skills related to organization, language and vocabulary, idea development, and conventions that are specific to a text type;
- The ability to apply writing skills to develop a narrative, informative/explanatory, or argument text.

Each academic content target prioritized for the NCSC assessment represented critical content for progressing from grade to grade. NCSC developed an array of items to address each prioritized target, and this array of items gave students an opportunity to show what they know and can do, whether they are just beginning instruction on the content or have already made notable progress.

Developing the Item Model

NCSC used a systematic process to develop an assessment of learning appropriate for students with the most significant cognitive disabilities. The item model includes assessment content designed with multiple levels of complexity and degrees of support that mirror evidence-based classroom practices; this characteristic of the model allows all students the opportunity to attempt the full range of items. The term “graduated complexity” refers to a relationship among four items (a family of items) associated with a single content target. The four items purposely vary from most to least difficult.

NCSC designed its AA-AAS to capture student performance through two item-design features: (1) levels of content complexity, and (2) degrees and types of scaffolds and supports to permit

independent performance. Using these features, the assessment design was based on the same model of learning used to develop the NCSC curriculum, instruction, and professional development resources described in Brief 7, Appendix 1-A. Incorporating structured scaffolds and supports helped ensure that NCSC's summative assessment provided opportunities for students to show *independently* what they know at varying levels of understanding.

Using this systematic process, the NCSC item development team selected the variable features associated with the mathematics or ELA content that was used to vary levels of difficulty in the item family. They also selected variable features associated with the UDL supports that were used to mitigate construct-irrelevant variance. As a result, the NCSC items provide greater item accessibility for all students. In particular, item developers linked variable features to learners' needs (e.g., perceptual, expressive, cognitive) in an effort to support student performances in non-construct relevant ways. Experts in student cognition and academic content participated throughout the item development process to ensure the appropriate identification and application of the variable features.

For each prioritized content target, a unique design pattern defined the types of observations NCSC could use to provide evidence about students' acquisition of the focal KSAs as well as the features (e.g., response mode options, graphic organizers, use of scaffolds and supports) of task situations that allowed students to provide this evidence. Each design pattern included a set of focal KSAs aligned with a prioritized academic grade-level content target. NCSC state partners and content and disabilities experts reviewed and approved a single focal KSA selected from each design pattern upon which to construct a task template.

Each task template included example items, in an item family, that served as a model for item development. The example items articulated the variable features of the design patterns applied to the content prioritized by the focal KSA. Development of the example items for each task template required the identification and articulation of additional KSAs for each item, or additional knowledge, skills, and abilities that might also be required in a task that addresses the focal KSAs. These include skills that may be required by tasks from a particular design pattern, some of which can be supported by UDL and accommodations.

In addition to the focal KSA, each item family included one example item linked to an essential understanding. The term essential understanding refers to the necessary knowledge and skills required for students to engage in the content identified by the academic grade-level content target. For example, in mathematics, the essential understandings refer to the fundamental concepts and skills essential to a student just beginning to learn the content, and include the specific symbols or referents related to the learning of these specific concepts and skills (e.g., mathematical operations of addition, subtraction, multiplication, and division).

As a result, the principled design approach was well suited to support the development of items integrating UDL to vary complexity across items in a family. In addition, NCSC's principled approach resulted in a process of systematic documentation of item development, which supported efficient future item development for the operational NCSC assessment. (See Appendices 2-A and 2-B).

Iterative Process for Finalizing the Design Patterns and Task Templates

NCSC content leads employed a highly structured iterative review process with state partners to finalize the design patterns and task templates. With direction from NCSC organizational partners, state partners reviewed each design pattern and task template with regard to (a) selection of focal KSAs, (b) identification of situations or potential observations to collect evidence of student acquisition of the focal KSAs, (c) guidelines for creating a range of difficulty of items across the four items in a family, and (d) the qualitative and quantitative criteria for the given themes for reading passages.

The NCSC iterative review process included several cycles of review used to inform multiple decision points in the development process. The evaluation at each of these decision points ensured that the design patterns and task templates reflected the intended application of UDL principles for the assessed population as well as NCSC's ongoing focus on evidence-based practices in curriculum, instruction, and assessment. This iterative model included the following elements:

- State partner feedback throughout the development process;
- Survey research and focus groups of key stakeholders;
- Content reviews of design patterns and task templates;
- Reviews of selected focal KSAs, essential understandings, passage topics, text types, passages, and item types;
- Content and bias reviews of the exemplar items;
- Student interaction studies (cognitive labs);
- Task template tryouts with partner teachers ;
- An evaluation study of the writing items.

According to the *Standards for Educational and Psychological Testing*, "validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA, APA, and NCME, 2014). NCSC used the feedback from the iterative review process to confirm that the resulting tests would adequately represent the intended content domain.

PASSAGE AND ITEM/TASK DEVELOPMENT PROCESS AND RESULTS

Introduction

This section of the chapter provides a description of the process undertaken by the NCSC partnership to develop items that elicit intended thinking processes. NCSC ensured that the content and skills in the items represent an adequate and appropriate sample of the grade-level college and career ready standards. As part of the test development process, item development followed the test specifications and included screening using criteria appropriate to the intended uses of the test (AERA, APA, and NCME, 2014). Using a principled design approach, NCSC delineated the format and content of the items; the response format or conditions for responding; and the type of scoring procedures.

Development of the NCSC item bank began in the winter of 2012 with the initial creation of item specifications, style and alternative text guidelines, and item quantities required for a 2015 operational assessment. Key decisions for art development provided the framework for both mathematics and ELA graphic development specific to the grade level and level of complexity of the item. These key decisions also provided the framework for the use of photographs to accompany informational passages.

The NCSC item specifications accomplished two purposes: (1) they provided both general and specific guidelines for developing all test items at the grade levels assessed by NCSC AA-AAS and (2) they described the test items and prompt types needed for the NCSC assessments. Within the specifications documents are sections dedicated to information about item contexts, variable features, cognitive task levels, use of graphics, item style and format, and general content limits by academic grade-level content target. NCSC drafted item specifications for each of the prioritized academic content targets at each grade and content area, which is documented below. (More detailed information on this work is documented in the vendor's *Item Writing Project Technical Report March 2014*, which is available on request.)

Item Development

A conceptual framework of principled design and UDL underpins the NCSC design of the example items and example item families that form the basis of the assessment. Each prioritized grade-level academic content target in the NCSC assessment represents the critical curriculum and instruction content for progressing from grade to grade. By selecting the prioritized academic content targets and identifying the essential understandings, NCSC staff and state partners determined the focal and additional KSAs, first in the design patterns and second as addressed in the task templates. NCSC then finalized each task template developed for a single focal KSA. The task template operationalizes the constructs to be measured and details the types of scoring to be used (including task specific rubrics) and the logic and presentation of the tasks. (See Appendix 2-A for example Design Pattern and Task Template.)

Each task template facilitates the development of item families with four items that vary systematically in complexity and degrees and types of scaffolds and supports. These four items in an item family focus on providing an age- and grade-appropriate assessment of a range of student learning relative to a specific grade-level academic content target. The items developed to address each content target give students an opportunity to show what they know and can do, whether they are just beginning instruction on the content or have already made a lot of progress. The least complex items in a family provide extensive adaptations, scaffolds, and supports. Other items for the same prioritized academic content target include content that is increasingly complex, and the items have fewer adaptations, scaffolds, and supports.

NCSC content specialists worked with the item development partner to develop item specifications based on the final design patterns and task templates. Each set of specifications identified a specific college and career ready standard, the prioritized academic content target, the focal KSA, and the essential understanding derived from the grade-level target. The item specifications also narrowed the item context and measurement focus of each assessment item. The specifications provided the item writers with the elements items include that are specific to the focal KSA and essential understanding, and are needed to assess different levels of student understanding. NCSC utilized an iterative process to review and finalize the item specifications. Once approved by NCSC, item writers received these specifications to guide item development. NCSC gave item writers access to and required them to use the design patterns and task templates as a reference throughout item development.

For mathematics, there were 10 academic content targets prioritized per grade level. Four item families (i.e., 16 items) were developed for each prioritized target. To ensure that 160 items would be viable for piloting after all rounds of review, 168 items were developed per grade for mathematics. Given the intensive principled development of design patterns and task templates, and tryouts of the task templates, this overage proved to be appropriate. Table 2-7 outlines the mathematics item development distribution.

Table 2-7. Mathematics Item Family Development

Grade	Mathematics families	Mathematics items
3	42	168
4	42	168
5	42	168
6	42	168
7	42	168
8	42	168
11	42	168
TOTAL	294	1176

Factors such as the balance of reading versus writing and the balance of literary versus informational texts interacted to create different demands to address content coverage for ELA item development. Thus, the number of prioritized content targets varied by grade level for ELA. The ELA item development requirements resulted in a greater overall number of ELA items to ensure adequate numbers of items for piloting after all review rounds. Table 2-8 outlines the ELA item development distribution.

Table 2-8. ELA Item Family Development

Grade	Reading families	Reading items	Writing families	Writing items
3	40	160	10	40
4	40	160	10	40
5	40	160	10	40
6	38*	152	10	40
7	40	160	10	40
8	40	160	10	40
11	40	160	10	40
Total	278	1112	70	280
	Total ELA Families: 348		Total ELA Items: 1392	

* Given that nine Academic Grade-level Content targets were prioritized at grade 6, fewer unique items families were required to address the targeted breadth of content coverage.

NCSC content and measurement experts created and presented intensive training for the test development partners' lead item writers. This training addressed the requirements of item development necessary to create items that are valid measures of grade- and age-appropriate student learning of the general curriculum by students with the most significant cognitive disabilities. Trainers also emphasized NCSC's focus on high expectations for students in this population, including the need to have items reflect the model of learning used to develop resources to support curriculum and instruction.

To ensure consistent preparation for item development, the same NCSC experts provided an intensive training program for all item writers assigned to the item writing leads. The training, delivered via an online platform, addressed the following objectives:

- Understanding the learner characteristics, communication modes, and varying levels of expressive and receptive communicative competence of students with the most significant cognitive disabilities;
- Producing items accessible to diverse groups of students;
- Understanding how to apply UDL principles to writing test items;
- Understanding cognitive complexity levels, paying specific attention to the descriptions of depth of knowledge used by NCSC;
- Interpreting college and career ready standards and NCSC's prioritized academic content targets;
- Using item specifications and test blueprints;
- Interpreting the NCSC design patterns and task templates;
- Understanding NCSC passage writing criteria for literary and informational text;
- Using NCSC-developed graphic and editorial style guides and guidelines for item writing, use of visual supports, and use of alternative text.

According to AERA, APA, & NCME (2014), the nature of the item and response formats appropriate for a test depends on the purposes of the test, the defined domain of the test, and the testing platform (pp. 75-77). NCSC used the design patterns and task templates to determine the item and response formats appropriate for the NCSC AA-AAS, which included selected-response (SR), multiple part selected-response (MSR), constructed-response (CR), and open-response (OR). Table 2-9 provides the item and response formats and scoring for the NCSC assessed content areas.

The item development partner developed passage sets for the NCSC reading assessment. Each passage set required four passages representing a range of complexity concerning reading level, length, and vocabulary. The items developed for a passage set also included a range of provided supports and scaffolds (e.g., an introduction to the text, a rereading option, pictures, prompts for what to listen for, and definitions). Thus, students who comprehend simplified text, students who comprehend longer, adapted grade-level texts, and students who need a blend of text across the range have the opportunity to demonstrate their understanding of text.

Prior to the start of passage writing, NCSC reviewed and approved passage set topics provided by the development partner and aligned with NCSC's required themes at each grade. The topics presented a diversity of characters of different genders and ethnicities and a variety of sub-genres. NCSC content experts and state partners employed an iterative process to review each passage in a family using NCSC-

generated passage development criteria (e.g., passage length) along with literary and informational complexity guidelines. This review process provided valuable feedback that the item development partner incorporated.

The test blueprint required four passages per test form so that each test form would include two literary passages and two informational passages. The approved passages were the basis for development of the ELA reading items. As with passage development, NCSC employed an iterative review process to review and revise the items. NCSC partners had multiple opportunities to provide feedback and approval during the review process. The item development partner supplied an initial round of items to NCSC for feedback prior to preparation for state partner content committee and bias committee reviews. For further information regarding the item writer selection and training, passage development specifications, item content and bias reviews, and application of alternative text, please refer to the *NCSC Item Writing Project Technical Report* produced by the item development partner in March 2014, referenced above.

Table 2-9. Item and Response Formats and Scoring

Item Type		Mathematics	Reading	Writing
Selected-Response	SR	• Connects to a single academic grade-level content target. • The student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1. • A correct response is worth 1 point at each of the four complexity levels.		
Multiple Part Selected-Response	MSR		• Connects to a single academic grade-level content target • In reading: o a cluster of two or three items is presented intentionally in a standardized, scripted sequence; o the student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1; o a correct response for each item is worth 1 point at each of the four complexity levels. • In writing: o a cluster of four to six items is presented in a standardized, scripted sequence of steps at complexity level 1 only; o the student independently selects a response from two options; o four to six correct responses are worth 2 points; one to three correct responses are worth 1 point (see Chapter 5, Scoring).	
Constructed-Response	CR	• Connects to a single academic grade-level content target. • Student interacts with presented materials to construct a response. • A correct response is worth 1 point at each of the four complexity levels.		• Connects to a single academic grade-level content target • Student interacts with presented materials to construct a response. • Response is scored using a 3 trait rubric.

Item Type		Mathematics	Reading	Writing
		• Response is scored using a 0-1 point rubric.		
Open-Response	OR		• In grades 3 and 4 only, connects to a single academic grade-level content target. • A set of five items/words at complexity levels 4, 3, or 2 or a set of three items/words at complexity level 1 are presented in a standardized, scripted sequence of steps. • Four to five correct responses at levels 4, 3, or 2 are worth 1 point; three correct responses at level 1 are worth 1 point (see Chapter 5, Scoring).	

Item Reviews

Item development was a collaborative effort between NCSC and the item development partner to review and revise items until they were consistent with development criteria and thus eligible to be pilot tested. NCSC conducted a total of seven mathematics and ELA item review meetings attended by approximately 270 content and special education experts from 23 NCSC state partners. According to AERA, APA, & NCME (2014), the review process should include expert judges to review items, qualifications, and relevant experiences; in addition, demographic characteristics, item review instructions, and reviewers' training should be documented (pp. 87-88). NCSC's item development partner collected evidence in support of these requirements.

NCSC sent out applications to every partner state prior to the recruitment phase of each review. Each group of potential panelists included subject matter experts with special education and/or content-area experience, educators with expertise using college and career ready standards, and state/district-level policy makers. For all reviews, NCSC formed two panels for each of four grade spans (grades 3–4, 5–6, –8, and grade 11). One panel focused on a review of item content, and the other panel focused on a bias and sensitivity review. Each panel consisted of two representatives from states and four representatives from local education agencies. NCSC organizational and state partners approved the criteria and process used to select panel members and asked potential participants to agree to participate in the following:

1. Initial and final reviews consisting of individual, independent reviews conducted remotely, followed by conference calls to share and collect feedback; and/or
2. On-site, face-to-face review meetings, including the following:

(a) An item review core team comprising state education agency (SEA) representatives from each state for both ELA and mathematics; reviewers were assigned to one of the review panels (content or bias and sensitivity) for each of the item reviews; and also participated in reconciliation meetings following item reviews;

(b) Other state and local education agency representatives who participated in one or more reviews as part of a content panel or a bias and sensitivity panel.

To review and revise items until they were deemed consistent with development criteria, NCSC and the item development partner collaborated to design and deliver training for web-based and face-to-face reviews. The training included an overview of the project's philosophy and goals, a recap of the work accomplished to date, and criteria for evaluation of each item by type and level of complexity.

NCSC content experts conducted reconciliation calls with the item development partner's content specialists following each review. The item development partner then revised the items to match the final NCSC-approved revisions, created a "Final for Review" PDF of each passage/item set, printed the PDF, and brought copies of the PDF to item review face-to-face meetings.

Through a series of face-to-face meetings, NCSC SEA and Local Education Agency (LEA) panels reviewed items against NCSC-established criteria. Training included strategies for examining the overall technical qualities of all items, such as language clarity, readability, plausibility of options, parallel structure of response options, significance and suitability of subject content, lack of bias, one correct answer, proper level of difficulty, and alignment to the focal KSA. The item review panels also

provided input on accessibility, addressing the needs of low-incidence special populations, and potential bias and/or sensitivity in the test content. With regard to fairness and content, panelists suggested revising or deleting items if necessary. NCSC and the item development partner then collaborated to evaluate and implement proposed revisions to the items as appropriate. At the close of the review schedule, all reviewers convened on a call with NCSC staff to reconcile item revisions. Lastly, NCSC staff convened on a call to reconcile further revisions or concerns with the item development partner's content specialists.

The final step of the review process included an examination of the applied Accessible Portable Item Protocol (APIP) Standard. The APIP standard provides assessment programs and question item developers with a data model for standardizing the interchange file format for digital test items. When applied properly, the APIP standard accomplishes two important goals. First, the standard allows digital assessments and items to be ported across APIP-compliant test item banks. Second, it provides a test delivery interface with all the information and resources required to make a test and an item accessible to students with a variety of disabilities and special needs. The item development partner applied the APIP standard and provided a test delivery interface with all the information and resources required to make the NCSC AA-AAS and each item accessible to students with a variety of disabilities and special needs. The APIP standard provided spoken or read-aloud text and graphics to address accessibility needs.

The item development partner facilitated grade-span panels for an APIP review of mathematics items and ELA passages and items; implementing the same number of rounds used for item review. Panels were comprised of NCSC SEA and LEA panelists with appropriate expertise. The APIP reviews of all final and approved items were conducted via day-long webinars. The APIP reviews focused on correct pronunciation of words given the applied APIP elements. Reviewers viewed the items in the item development partner's APIP services tool, which presented the items in the way that a student would see and hear them, based on the applied APIP element. The panelists saw and heard the items read aloud for a spoken nonvisual audience. Any items that survived this rigorous examination became part of the item pool used to develop test forms for piloting. (For a more detailed description of the item bank development, see the *Item Writing Project Technical Report*.)

NCSC maintained protocols to ensure the security of all reviewed items. Item development partners required reviewers to sign a nondisclosure form prior to distributing any secure materials. The item development partner's facilitators collected the signed form before providing reviewers with numbered item booklets. The facilitators maintained a tracking sheet for the item booklets and required sign-in and sign-out of materials each day of the meeting. In addition, the panel rooms were locked when the facilitator was not in the room. NCSC prohibited reviewers from removing any of the materials from the room; the materials stayed in the locked room used by a given panel. Daily, as well as at the completion of the meetings, each facilitator collaborated with the NCSC partner representatives to account for all materials. At the close of each review meeting, the item development partner collected for secure transportation and storage all item booklets containing reviewer comments and placed all other materials in locked bins for shredding via secure methods. As part of the item reviews, NCSC researchers conducted focus groups with LEA and SEA representatives. Summaries of the reports of each focus group follow.

Mathematics Item Content Review Focus Group Report: Local and State Education Agency Representatives' Perspectives

Following the NCSC mathematics item content review, NCSC researchers conducted focus groups with 16 LEA and 8 SEA representatives who participated in the NCSC mathematics item content review. NCSC conducted the focus groups in July and August 2013, following the content and bias reviews of the NCSC mathematics assessment items. NCSC researchers conducted the focus groups to gather LEA and SEA perspectives regarding item content, instruction, materials and supports, and process consistency. The focus group facilitators encouraged participants to share their suggestions and voice their concerns. The focus groups were structured to collect LEA and SEA feedback separately, and questions focused on the expertise of each group.

NCSC researchers condensed the patterns identified from focus group participants' feedback to determine broad themes across the data. Independent researchers then verified these themes to ensure that they accurately represented the input and perspectives of the participants. NCSC researchers found common themes in the feedback provided by the two groups, but researchers also recognized that representatives provided some feedback specific to their LEA or SEA roles.

Through the focus groups, LEA and SEA representatives had the opportunity to interact with each other and share input regarding the varying aspects of the content and items they reviewed. Overall, focus group participants indicated that many of the items were too difficult for the students who would be taking the NCSC AA-AAS and that these items would not allow those students to demonstrate what they know and can do. In addition, representatives consistently indicated that a paradigm shift is needed to ensure that students in this population have the learning opportunities necessary to support their success. NCSC staff used the information collected from the focus groups to understand the representatives' feedback and to inform feedback implementation. Results were analyzed first by focus group type, followed by cross-group themes. With respect to the identified themes, the focus group participants indicated that teachers face a shift in the content they teach and need support and professional development to implement the new assessment system. Representatives indicated that changes are needed at both the classroom and district levels to prepare teachers and students for the NCSC assessment, and they made recommendations to improve the NCSC mathematics items.

Reading Item Content Review Focus Group Report: Local and State Education Agency Representatives' Perspectives

Following the NCSC reading item content review, NCSC researchers conducted focus groups with 16 LEA and 8 SEA representatives who participated in the NCSC third-round reading item content review. NCSC conducted the focus groups in September 2013, following the content and bias reviews of the NCSC ELA reading assessment items. NCSC researchers conducted the focus groups to gather LEA and SEA perspectives regarding item content, instruction, materials and supports, and process consistency. The focus group facilitators encouraged participants to share their suggestions and to voice their concerns. The focus groups were structured to collect LEA and SEA feedback separately, and questions were focused on the expertise of each group.

Researchers condensed the patterns identified from focus group participants' feedback to determine broad themes across the data. Independent researchers then verified these themes to ensure that

they accurately represented the input and perspectives of the participants. Researchers found common themes in the feedback provided by the two groups, but they also recognized that representatives provided some feedback specific to their LEA or SEA roles. Results were analyzed first by focus group type, followed by cross-group themes. With respect to the themes, the focus group participants indicated that teachers face a shift in the content they teach and need support and professional development to implement the new assessment system. Representatives indicated that changes are needed at both the classroom and district levels to prepare teachers and students for the NCSC assessment, and they made recommendations to improve the NCSC reading items.

PILOT PHASE 1 AND PHASE 2 TESTING OVERVIEW

Introduction

This section addresses the Interpretive Argument (IA, see NCSC Brief 9, *NCSC's Theory of Action and Validity Evaluation Approach*, Appendix 1-A) questions focused on (1) an AA-AAS that will allow students to demonstrate their knowledge and skills, (2) items that elicit the intended cognitive processes, and (3) administration procedures and data-capture methods that are standardized in ways that support comparability across students, schools, and time. In addressing these questions, the chapter also describes how the information collected supports the observation and interpretation vertices of the Assessment Triangle.

Overview of Pilot Phase 1 and Phase 2 Approach

Over the course of item and test development, the NCSC summative project maintained a consistent focus on the goal of developing a defensible and innovative online summative assessment program for students with the most significant cognitive disabilities. To this end, the field test design developed in 2013 included a two-phased pilot approach. NCSC conducted Pilot Phase 1 in spring 2014 and Pilot Phase 2 in fall 2014. NCSC operated under the assumption that a two-phase process supported the collection of evidence at multiple stages of assessment development and that these research efforts would provide evidence to evaluate and inform the further development of the NCSC summative assessment. Using the phased approach, NCSC obtained critical data and evidence of student performance at early stages of item and test development; this data and evidence informed further revision of items as well as the development of the operational tests.

According to AERA, APA, and NCME (2014), the analyses carried out using pilot and field-testing data should seek to detect aspects of test design, content, and format that might distort test score interpretations for the intended use of the test scores for particular groups and individuals (p. 64). NCSC's analysis of the collection of evidence obtained through Pilot Phase 1 and Phase 2 provided opportunities to ensure that all aspects of assessment development contributed to interpreting appropriately the scores of students participating in the NCSC AA-AAS. The pilot phases were completed in conjunction with a series of studies and surveys to collect additional data, and collectively were considered a multiple method field test of the NCSC assessment system and items.

In Pilot Phase 1, students from 17 NCSC partner states and territories in grades 3–8 and 11 were administered test forms in either reading or mathematics. Students from 19 NCSC partner states and territories in grades 4, 5, 6, 7, 8, 9 and 12 participated in Pilot Phase 2; these students were assessed on academic content in mathematics and ELA from grades 3, 4, 5, 6, 7, 8, and 11, respectively (i.e., one

grade below the students' current grade). NCSC recruited student participants for Pilot Phase 1 and Phase 2 from the population of students who previously participated in a state's AA-AAS, which represented approximately 1% of the total population assessed through general and alternate assessment. Both phases included a diverse student population including students of all disability categories, though the majority of students identified with three disability categories (intellectual disabilities, multiple disabilities, and autism). Students had highly varied levels of expressive/receptive language use, and most students were identified as using symbolic communication.

NCSC designed the policies and administration procedures for the Pilot Phase 1 and Phase 2 to adhere to best practices (AERA, APA, & NCME 2014, pp. 63–65, 67–70). These best practices ensure that there are opportunities for all assessed students to demonstrate what they know and can do and that interpretable, accurate, and actionable information about academic performance can be obtained.

Technical Platform for NCSC Assessment Pilot

NCSC's technology platform development partner, working with NCSC partners, created the NCSC system to administer the NCSC AA-AAS to students participating in pilot phases implemented in the spring and fall of 2014. Students and teachers interacted with the NCSC assessment system as the TAO-based online technology platform (an open source e-testing platform) to facilitate computer-based testing of linear, fixed-length assessment forms consisting of SR and CR items. The system was supported on several browsers including Chrome, Firefox, and Internet Explorer and worked only on Windows or Mac operating systems.(For additional information about the NCSC technology platform, see Appendix 2-C: Design for Technology Platform for NCSC Assessment System.)

During recruitment and student registration for each pilot phase, state partners registered Test Administrators (TAs) and test coordinators (TCs) into the online platform. This allowed TAs and TCs to enter the system, register their students, and complete online training. Once TAs and TCs were registered in the system, they received login information to access the NCSC assessment system.

The online platform was home to the NCSC online test administration training modules for TCs and TAs. Prior to testing, TAs completed the training modules and were required to complete the end-of-training quiz with a qualifying score of 80% correct answers in order to access the testing materials and administer a pilot test. TAs had the opportunity to retake the test if they did not meet the qualifying criterion and could access the training modules throughout the testing process. TCs were required to complete the training modules but were not required to pass an end-of-training quiz.

The training modules provide critical information to equip TAs with the necessary knowledge to navigate the system and administer the test in accordance with established guidelines. The training modules included information about the following:

- Accessing and logging into the system;
- Understanding the platform dashboard/home page and the navigation panel;
- Test roster and directions on test actions, including downloading and closing the test;
- Examples of what the assessment looks like for students.

NCSC partners provided System User Guides and a Test Administrator Manual (TAM) to support TAs' understanding and use of the system. This documentation included information regarding various tools and capabilities of the assessment system to support TA understanding and use of system components.

Before administering a pilot test, NCSC required that TAs confirm student demographic information, Learner Characteristics Inventory (LCI) data, and accommodations information within the assessment system. By completing a student response check (see the Test Administration Manual) in the NCSC assessment system prior to testing, TAs also documented whether the student had an observable way to communicate responses to the items. TAs were also asked to document any accommodation the student used during the pilot test. The system also supported interaction with sample items, so TAs and students could practice using appropriate assessment features and accommodations.

TAs administered the test individually to each student following the scripted text provided in the Directions for Test Administration. If the student did not use a mouse to enter an item response directly into the NCSC assessment system, the TA entered the student's answer choice as indicated by the student's verbalization, pointing, eye gaze, use of assistive technologies, and so on. NCSC permitted students to use a variety of assistive technology devices for responding to items; including text-to-speech, alternate keyboards, switch-based navigation, and eye gaze. The system was also able to create a paper-based version for pilot test administration.

The NCSC assessment platform for pilot testing included tool bar buttons, which allowed TAs to select the following options: "read again," "previous," "next," "question list," "toggle," "student name," "name of test/session," "current question number of the total number of questions," "bookmark," "upload evidence," and "save and exit," which was used to pause and resume the test. There were also arrows to move forward and backward through the test. The TA was able to pause and resume the test to accommodate the student's needs.

For math CR items, TAs checked a radio button to indicate a student's response. For ELA writing CR items, the TA or student recorded the student's response to a writing prompt on response templates that were part of the online NCSC assessment system. TAs could either use the computer's webcam to capture an image of the evidence or scan the evidence using a scanner and then upload the file as an attachment. NCSC required TAs to upload the evidence of a student's response to the writing CR items if the student was unable to enter the response directly into the online template.

Platform Assessment Features

NCSC partners incorporated flexibility into the assessment design by implementing various technology features in the platform for Pilot Phases 1 and 2. The student or TA could enable platform features in the test delivery system at the time of testing. Some of the features were unique to the NCSC assessment system and some were computer-based. The NCSC assessment system User Guide included detailed descriptions of the platform features and directions for enabling the features.

All students during Pilot Phase 1 had access to a limited number of platform features, including the ability to increase the size of text and graphics. Additional features were available for use during Pilot Phase 2:

- Answer masking (hide answers and focus on stem);
- Audio player (to read all passages, items, and response options to the student);
- Alternate color themes (background and font color);
- Increase or decrease size of text and graphics;
- Increase volume;
- Line reader tool;
- Read aloud and reread item directions, response options, passage.

Purpose of Pilot Phase 1 and Phase 2

NCSC and the pilot development partner used evidence from the pilot research at both the item and form levels to support decisions related to the development of the NCSC item pool, the design and delivery of tests, and the methods used for linking and scaling.

Pilot Phase 1

NCSC used Pilot Phase 1 to (a) generate student performance data, (b) investigate administrative conditions, (c) understand item functioning, and (d) investigate the proposed item scoring processes and procedures. This initial phase of the NCSC pilot represented a large-scale effort to test all of the developed items with students.

Pilot Phase 2

NCSC used Pilot Phase 2 to (a) continue to investigate item functioning; (b) investigate test structure, including the possibility of supporting an adaptive algorithm in future administrations; and (c) address specific NCSC Technical Advisory Committee (TAC) recommendations they developed using evidence from Pilot Phase 1. TAC recommendations supported NCSC's Pilot Phase 2 focus on representing the population of interest in the student sample, clearly communicating the use of the student response check, and minimizing the impact of scrolling on student performance. Evidence gleaned from research during both phases of the pilot supported the evaluation of the items' statistical properties and addressed whether the item formats functioned as intended (AERA, APA, & NCME 2014, Standard 4.10).

Student Participation in Pilot Phase 1 and Phase 2

Students who were determined eligible by their Individualized Education Program (IEP) teams for participation in their state's AA-AAS were eligible to participate in the NCSC Pilot Phase 1 and Phase 2. *Guidance for IEP Teams on Participation Decisions for the NCSC Alternate Assessment* describes in detail the NCSC AA-AAS participation criteria and is found at the following link: www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC_Participation_Guidance-Nov-2013.pdf.

Reflecting the pervasive nature of a significant cognitive disability, the criteria for student participation in the NCSC Pilot Phase 1 and Phase 2 included the requirement that all content areas should be considered when determining who should participate. The criterion was applied in an overall sense for the pilot work in that, to participate, a student had to be eligible for both content areas. For the

purpose of NCSC Pilot Phase 1, students participated either in the ELA (reading only) test or the mathematics test. For Pilot Phase 2, students participated in either the ELA (reading and writing) test or the separate reading and mathematics tests.

For both Pilot Phase 1 and Phase 2, the provision of accommodations included in a student's IEP was required. (For information regarding the student sample in Pilot Phase 1 and 2, see Appendix 2-D: Pilot Phase 1 Prioritized Sample Characteristics and Student Demographics and Appendix 2-E: Pilot Phase 2 Prioritized Sample Characteristics and Student Demographics).

Test Security for Pilot Phase 1 and Phase 2

All persons associated with Pilot Phase 1 and Phase 2 administrations had assigned responsibilities with respect to test security, which is critical to ensuring that information about student academic performance is accurate, valid, reliable, and timely. Security procedures often include procedures for tracking and storing materials. To ensure that inappropriate test practices did not undermine efforts to improve student achievement, in Pilot Phase 1 and 2, NCSC provided District Test Coordinators, School Test Coordinators, and TAs with a TAM containing implementation policies related to security, integrity, and appropriate and inappropriate test practices. Also included were policies concerning compliance with a state's test security protocols and procedures, and signing and submitting state-specific required test security agreements, as outlined in the various states' laws and policies.

Pre-Administration Activities for Pilot Phase 1 and Phase 2

Training and Certification of TAs and TCs

According to AERA, APA, and NCME (2014), equity in treatment requires that all test takers have qualified TAs with whom they feel comfortable and can communicate to a practical extent; equity in treatment also requires that TAs follow carefully the standardized procedures for administration and scoring specified by the test developer (pp. 50, 114). NCSC used TA training and the provision of ancillary testing materials to ensure the fidelity of test implementation and the validity of the assessment results and to prevent, detect, and respond to irregularities in academic testing and testing integrity practices for technology-based assessments. NCSC required that each TA was a certified, licensed educator and that each was familiar with the tested students and the students' accommodations. In addition, to reduce threats to fair and valid interpretation of test scores, NCSC required that TAs and TCs adhere to all aspects of the testing process and testing conditions. Training modules customized for the specific responsibilities of TAs and TCs for Pilot Phase 1 and Pilot Phase 2 were self-paced, narrated, closed-captioned modules that elaborated upon the information in the TAM and the Directions for Test Administration (DTA). NCSC required that all TAs and TCs complete the training, which is described in more detail in the sections that follow.

Pilot Phase 1 Training and Certification

According to AERA, APA, and NCME (2014), standardization of administration conditions and scoring procedures helps ensure that test takers have comparable contexts (p. 65). Therefore, NCSC required all TAs and TCs to meet training requirements and obtain qualification by completing eight comprehensive training modules posted in March of 2014 prior to the spring 2014 Pilot Phase 1 testing window. TAs completed a quiz after each module and a comprehensive quiz after completing all modules and were required to obtain a score of 80% on the comprehensive quiz. After meeting this requirement, a

TA could administer the NCSC Pilot Phase 1 tests. In addition, NCSC provided ancillary testing materials that outlined specific practices and policies, including the TAM; NCSC Online Test Administration Training; and grade- and content-specific DTAs.

Upon successful completion of the required training, NCSC directed TAs to enter Learner Characteristic Inventory (LCI, See NCSC Brief 8, *Characteristics of Students with Significant Cognitive Disabilities: Data from NCSC's 2015 Assessment,* Appendix 1-A) data for each participating student, using the NCSC AA-AAS delivery system. In addition, NCSC requested that each TA submit an End of Test Survey (EOTS) for each student assessed by a TA. The EOTS responses provided information regarding (1) the demographics and learner characteristics of students participating in the NCSC pilot; (2) the provision of opportunity to learn grade-level academic content; (3) TA perceptions of the administration procedures for the test; and (4) general feedback on the test, items, format, and accommodations.

Pilot Phase 2 Training and Certification

NCSC required all TAs and TCs to meet training requirements and obtain qualification by completing 13 NCSC Online Test Administration Training modules (for TAs) and 5 modules (for TCs) available from September 29 through November 14, 2014. TAs completed a quiz after each module and a comprehensive quiz after completing all modules and were required to obtain a score of 80% on the comprehensive quiz. After meeting these requirements, NCSC permitted TAs access to (1) DTA for the tests assigned to students; and (2) test forms assigned to the students.

In addition, NCSC provided ancillary testing materials that outlined specific practices and policies, including the TAM; NCSC Online Test Administration Training; and grade- and content-specific DTAs. The purpose of the training and ancillary testing materials was to ensure fidelity of implementation and the validity of the assessment results and to prevent, detect, and respond to irregularities in academic testing and testing integrity practices for technology-based assessments.

Upon successful completion of the required training, NCSC directed TAs to enter LCI data for each participating student, using the NCSC AA-AAS delivery system. In addition, NCSC requested that each TA submit an EOTS for each student assessed by a TA.

Student Response Check for Pilot Phase 1 and Phase 2

The purpose of the student response check (SRC) was to ensure that each student had the opportunity to demonstrate an independent, consistent, and observable response to the items on the Pilot Phase 1 and Phase 2 tests and that the TA could clearly identify which answer a student indicated in response to an SR test item. Entry of the student's response by the TA in the NCSC assessment system required that a student's response to a test item be observable.

The SRC was a three-question, content-neutral task presented by the TA to individual students in either a computer or paper administration so that a student could demonstrate his or her preferred mode(s) of communication. The NCSC TAM for both Pilot Phase 1 and Phase 2 included administration guidance on the SRC. Guidance in the Pilot Phase 1 TAM required that the TA administer a response check to

observe a student's. The TAM included a PDF version of the SRC to familiarize the TA with the SRC's structure and content.

In Pilot Phase 2, the NCSC TAM required the TA to conduct an SRC if the TA was not certain that the student's response to a test item would be observable; or it required the TA to not conduct an SRC if the TA was certain that the student had an established and clear method of communication and would clearly indicate their answers to the test questions by responding verbally, using a nonverbal communication mode, or using assistive technology (AT). The TAM provided additional guidance on administering the SRC using the computer or paper.

For both Pilot Phase 1 and Phase 2, the TAM specified that the use of teacher-directed hand-over-hand was not considered to be a consistent and observable response on the part of the student (i.e., using this method, the student was not indicating an answer choice independently). In addition, for both Pilot Phase 1 and Phase 2, the guidance stated that a consistent and observable response did not mean that the student must use the same response mode for every item. Examples of consistent and observable responses include but are not limited to the following:

- Using the mouse to select the answer;
- Verbalizing the answer;
- Gesturing or pointing to the answer;
- Using AT to indicate the answer;
- Using eye gaze chart to select the answer;
- Circling or marking the answers on a paper copy of the test.

Administration of Pilot Phase 1 and Phase 2

NCSC's pilot development partner administered the Pilot Phase 1 and Phase 2 tests using the NCSC online platform. According to the *Standards for Educational and Psychological Testing*, a principle of standardization includes orienting test takers prior to the testing experience to materials and accommodations with which they may not be familiar (AERA, APA, & NCME 2014, pp. 90-91, 116). NCSC understood that providing instruction and sample items allowed the test taker and TA to gain an understanding of the item type, assessment features, and operation of the equipment or software. Thus, NCSC allowed students and TAs to access sample ELA and mathematics items located on the NCSC online portal. Each content area had three grades of sample items. The sample items provided students and their teachers with opportunities to familiarize themselves with the software, navigational tools, and item types used in Pilot Phase 1 and Pilot Phase 2.

An option to print the tests in all grades for administration promoted accessibility because it provided equal opportunity for test takers to demonstrate their knowledge of the assessed content utilizing a paper/pencil vs. online presentation of the test. Untimed testing conditions during the administration of Pilot Phase 1 and Phase 2 provided flexibility for the TA to consider the characteristics and needs of the students with respect to engagement and fatigue, for example. The NCSC TAM provided district and school administrators and TAs with guidelines for planning, preparing, and managing Pilot Phase 1 and Phase 2 administrations; and the grade- and content-specific DTAs provided TAs with specific directions for administering different item types following specific protocols, preparing ancillary testing materials, and using item-specific scripts and directives.

Accommodations for Pilot Phase 1 and Phase 2

Accommodations are changes in the materials or procedures of an assessment that do not alter the item and what is being measured. Without accommodations, the assessment may not accurately measure the student's knowledge and skills (Thurlow, Lazarus, & Christensen, 2013). During the fall of 2013 and the early part of 2014, the NCSC Accommodation Committee determined the category and type of accommodation that would be allowed on the NCSC Pilot Phase 1 and Pilot Phase 2 tests following an iterative review process of the design patterns by organizational and state partners. The committee analyzed the variable features and UDL components of the design patterns, which defined the scaffolds and supports that did not interfere with the measured constructs and optimized accessibility. The focus was to ensure that any necessary additional accommodations did not interfere with the measured construct and were necessary for students to demonstrate their KSAs through valid measurement. Accordingly, the built-in scaffolds, supports, and universally designed assessment minimized the number of identified and allowed accommodations. See Appendix 2-B for a description of the committee's work.

Accessibility Considerations for Pilot Phase 1 and Phase 2

For students who, in addition to having significant cognitive disabilities, are blind, deaf, or deaf/blind, the Pilot Phase 1 TAM included reference to the Accessibility Addendum. The Phase 2 TAM included reference to the Procedures for Assessing Students Who Are Blind, Deaf, or Deaf/Blind for additional administration guidance. The guidelines described in the addendums provided background information to guide teachers' administration of the Pilot Phase 1 and Phase 2 tests to students requiring accessibility solutions. These addendums did not replace the TAM and NCSC online test administration training, but rather provided additional information to TAs to support the administration of the test to the aforementioned students. NCSC directed TAs to use these guidelines in conjunction with the DTA, but told them that the guidelines did not supersede those in the TAM. The addendums were secure documents, and the TAM directed TAs and TCs to apply the security procedures outlined in the TAM. Each state NCSC coordinator securely transmitted the Pilot Phase 1 or Phase 2 addendum to the TAs of students identified as having one of the listed disabilities or identified as using braille.

The Pilot Phase 1 addendum provided information for TAs about preparing and administering the test to students with the most significant cognitive disabilities who are blind, deaf, or deaf/blind. The addendum provided information on the use of tactile symbols, object replacement, sign language interpretation, and additional considerations for these students. Upon completion of each student's assessment, the addendum directed TAs to complete a unique EOTS, which included specific questions related to increasing accessibility. Information from this EOTS was used to inform revisions and expand guidance for TAs in Pilot Phase 2.

The updated Pilot Phase 2 addendum, Procedures for Assessing Students Who Are Blind, Deaf, or Deaf/Blind, included guidance related to (1) tasks to complete before, during, and after the assessment; (2) strategies, with definitions and examples, that could be used by the TA with individual students, as appropriate, to enhance access to the NCSC AA-AAS; and (3) DTAs that must be used to administer open-response foundational reading items in grades 3 and 4.

Item Types Utilized in Pilot Phase 1 and Phase 2

The Pilot Phase 1 and Phase 2 tests utilized and scored a variety of item types that included response formats appropriate for the ELA and mathematics tests. The item and response formats included in the Pilot Phase 1 and Phase 2 tests were selected-response (SR), multiple part selected-response (MSR), constructed-response (CR), and open-response (OR).

Pilot 1 Item Types and Scoring

Pilot Phase 1 included SR, MSR, CR, and OR items (see Table 2-10. Pilot Phase 1 Item and Response Formats and Scoring). For the reading SR and MSR items and the mathematics SR and CR items, correct answers were assigned a score of 1 and incorrect answers were assigned a score of 0. In the SR and MSR items, students independently selected a response from a list of available answer options using appropriate response modes. For the mathematics CR items and the reading OR items, the TA scored the student's independent response as directed in the grade-specific DTAs.

To administer and score the mathematics CR items, the TA presented each item using a standardized, scripted sequence of steps that culminated in a TA's scoring of the student performance against the mathematics scoring rubrics. The mathematics scoring rubrics provided scoring standards to evaluate student responses. The TA entered the student CR score into the NCSC test platform. All directions and materials needed for administering CR items were in the DTA that accompanied each test form.

The TA scored the OR items (i.e., reading foundational items) as administered in grades 3 and 4 only. The DTA that accompanied each test form detailed specific directions for administering and scoring the OR items. After each student response, the TA entered a student's score into the NCSC test platform. Students with clear and consistent oral speech were administered the OR reading foundational items. Students who used means of communication other than oral speech, such as an AAC device, American Sign Language or eye gaze, were administered the SR reading foundational items.

For both the CR mathematics items and the OR reading items, the answer options selected by the TA were "student responded correctly" or "student did not correctly respond." SR items had two or three response options and did not use "none of the above" or "all of the above" as response options.

Table 2-10. Pilot Phase 1 Item and Response Formats and Scoring

Item Type		Mathematics	Reading	(No Writing items in Pilot Phase 1)
Selected Response	SR	• Connects to a single academic grade-level content target. • The student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1. • A correct response is worth 1 point at each of the four complexity levels.		
Multiple Part Selected-Response	MSR		• Connects to a single academic grade-level content target • In reading: o a cluster of two or three items is	

Item Type		Mathematics	Reading	(No Writing items in Pilot Phase 1)
			presented intentionally in a standardized, scripted sequence; o the student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1; o a correct response for each item is worth 1 point at each of the four complexity levels	
Constructed-Response	CR	• Connects to a single academic grade-level content target. • Student interacts with presented materials to construct a response. • A correct response is worth 1 point at each of the four complexity levels. • Response is scored using a 0–1 point rubric.		
Open-Response	OR		• In grades 3 and 4 only, connects to a single academic grade-level content target. • A set of five items/words at complexity levels 4, 3, or 2 or a set of three items/words at complexity level 1 are presented in a standardized, scripted sequence of steps. • Four to five correct responses at levels 4, 3, or 2 are worth 1 point;	

Item Type		Mathematics	Reading	(No Writing items in Pilot Phase 1)
			three correct responses at level 1 are worth 1 point (see Chapter 5, Scoring).	

Pilot Phase 2 Item Types and Scoring

Pilot Phase 2 included SR, MSR, CR, and OR items (see Table 2-11. Pilot Phase 2 Item and Response Formats and Scoring). For the reading SR and MSR items and the mathematics SR and CR items, correct answers were assigned a score of 1 and incorrect answers were assigned a score of 0. In the SR and MSR items, students independently selected a response from a list of available answer options using appropriate response modes. In the mathematics CR items and the reading OR items, the TA scored the student's independent response as directed in the grade-specific DTAs.

In mathematics, the TA presented each item using a standardized, scripted sequence of steps that culminated in a TA's scoring of the student performance against the mathematics scoring rubrics. These rubrics provided scoring standards to evaluate student responses. The TA entered the student CR score into the NCSC test platform. All directions and materials needed for administering CR items were in the DTA that accompanied each test form.

For the reading foundational items, the TA scored the OR administered in grades 3 and 4 only. The DTA that accompanied each test form detailed specific directions for administering and scoring the OR items. The TA entered a student's score after each student's response into the NCSC test platform. Students with clear and consistent oral speech were administered the OR reading foundational items. Students who used means of communication other than oral speech, such as an AAC device, American Sign Language, or eye gaze, were administered the SR reading foundational items.

For both the CR mathematics items and the OR reading items, the answer options selected by the TA were "student responded correctly" or "student did not correctly respond." SR items had two or three response options and did not use "none of the above" or "all of the above" as response options.

Table 2-11. Pilot Phase 2 Item and Response Formats and Scoring

Item Type		Mathematics	Reading	Writing
Selected-response	SR	• Connects to a single academic grade-level content target. • The student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1. • A correct response is worth 1 point at each of the four complexity levels.		
Multiple part selected-response	MSR		• Connects to a single academic grade-level content target. • In reading: o a cluster of two or three items is presented intentionally in a standardized, scripted sequence; o the student independently selects a response from three options at complexity levels 4, 3, and 2, and two options at complexity level 1; o a correct response for each item is worth 1 point at each of the four complexity levels • In writing, o a cluster of four-six items is presented in a standardized, scripted sequence of steps at complexity level 1 only o the student independently selects a response from two options o four – six correct responses are worth 2 points; one-three correct responses are worth 1 point (see Chapter 5, Scoring)	
Constructed-response	CR	• Connects to a single academic grade-level content target. • Student interacts with presented materials to construct a response. • A correct response is worth 1 point at each of the four complexity levels. • Response is scored using a 0-1 point rubric.		• Connects to a single academic grade-level content target. • Student interacts with presented materials to construct a response. • Response is scored using a 3-trait rubric.

Open-response	OR		• In grades 3 and 4 only, connects to a single academic grade-level content target. • A set of five items/words at complexity levels 4, 3, or 2 or a set of three items/words at complexity level 1 are presented in a standardized, scripted sequence of steps. • Four to five correct responses at levels 4, 3, or 2 are worth 1 point; three correct responses at level 1 are worth 1 point (see Chapter 5, Scoring)	

PILOT PHASE 1 AND PHASE 2 TEST DESIGN

Pilot Phase 1 Test Design Overview

For Pilot Phase 1, the NCSC development partner constructed linear, fixed forms in reading and mathematics that included a mix of levels of item complexity. All items developed for reading and mathematics, across all levels of complexity, were included in Pilot Phase 1. (For further information, see Appendix 2-F: Pilot Phase 1 Blueprint and Forms.) NCSC administered all the developed writing items through a separate study conducted concurrent to Pilot Phase 1. (See description of additional studies later in this chapter for more information about the writing evaluation study.)

Pilot Phase 1 Reading Test Design

The Pilot Phase 1 reading test design consisted of two sessions. Each reading Pilot Phase 1 form included four reading passages—two each associated with literary and informational texts with accompanying SR comprehension and vocabulary items. Additional foundational reading items were included in grades 3 and 4 only. In total, the reading tests consisted of approximately 25 SR items.

Reading session 1 began with a literary passage set (a passage and the corresponding set of comprehension and vocabulary items) followed by an informational passage set. Reading forms in grades 3 and 4 included either one three-part or five-part foundational item in Session 1 and Session 2. Reading

session 2 began with an informational passage set (passage and the corresponding set of comprehension and vocabulary items) followed by a literary passage set.

Pilot Phase 1 Mathematics Test Design

The mathematics tests for Pilot Phase 1 consisted of approximately 25 SR and CR items. Each mathematics form included three sessions with 8–9 items administered within each session to ease the testing load on students. The same number of items (25) comprised an overall mathematics form administered for Pilot Phase 1. Within each of the three test sessions, the pilot development partner and NCSC content experts grouped items by domain so that students were not required to shift focus frequently during a session.

Pilot Phase 2 Test Design Overview

Pilot Phase 2 consisted of four forms per grade and content area for the reading and mathematics tests, and two forms per grade for the ELA tests. The forms were built using classical item and test statistics from Pilot Phase 1 results and were based on content distribution and distribution of complexity levels. NCSC's pilot development partner structured Pilot Phase 2 forms at each grade to support piloting 100 items per content area per grade, providing the item pool needed to create forms for the first operational tests. As in Pilot Phase 1, NCSC administered the fixed-test forms through the online NCSC assessment system. Unlike Pilot Phase 1, the purpose of Pilot Phase 2 was not to test every item in the pool; rather, the purpose was to assess specific content, complexity, and statistical specifications that were targeted for each of the two test sessions. The test design included a session of common items (session 1) for all students before branching off to a different and distinct session 2. Developers created the Pilot Phase 2 forms to cover the breadth of the grade-level, content-area blueprints. The Pilot Phase 2 test blueprints were constructed with prioritized content goals at proportions consistent with the content standards and reflected the overall operational design intended for the first operational tests. (For further information, see Appendix 2-G: Pilot Development Partner Report: Pilot Phase 2 Blueprint and Forms.)

Pilot Phase 2 Special Form Development

Session 1 and one of the Session 2 sessions for ELA and mathematics tests were designed to allow maximum accessibility for students who, in addition to having significant cognitive disabilities, are blind, deaf, or deaf/blind, and possibly have other disabilities. The NCSC Accessibility Committee developed accessibility guidelines, in conjunction with low-incidence experts and partner action research teachers in several state schools for the blind and for the deaf, which were then used for the selection of reading passage sets and mathematics items. This was done to ensure that the items selected to construct Form 1 in ELA and mathematics were considered maximally accessible by expert and practitioner advisors.

There were also special form assignment criteria for students who could not respond verbally, who had vision or hearing impairments, or who used braille. Students with vision or hearing impairments, as indicated by variables in their enrollment file, were assigned Form 1 regardless of the test form assignment other students in their school or classroom received. Students who were nonverbal or who used braille received nonverbal or braille versions of the foundational items in reading and ELA. Each state NCSC coordinator securely transmitted the "Procedures for Assessing Students Who Are Blind, Deaf, or Deaf/Blind" to the TAs of students who were identified as having one of the listed disabilities or

who used braille. NCSC directed TAs to use these guidelines in conjunction with the DTA, but told them that the guidelines did not supersede those of the TAM. The "Procedures" guidelines provided background information to guide the administration of the Pilot Phase 2 tests (e.g., reading foundational items) to students requiring accessibility solutions.

Pilot Phase 2 ELA Test Design

Pilot Phase 2 included two ELA test forms at each grade that consisted of both reading and writing content. The first session of the ELA form consisted of reading content, four stand-alone writing SR items, and one writing-prompt-based set of SR items developed at the lowest level of complexity. The reading content on the ELA test forms was the same reading session 1 content used for the reading tests. Session 2 consisted of two writing CR items. The writing CR items required students to produce a permanent product in response to a writing prompt. The presentation of each item included a standardized, scripted sequence of steps. The student, TA, or a scribe (an adult familiar to the student who writes or types exactly what the student communicates by speech, sign language, or AT) recorded the response to the prompt on the response templates that are part of the NCSC assessment system.

Pilot Phase 2 Reading Test Design

The Pilot Phase 2 reading tests consisted of four forms dispersed across two sessions. Session 1 was a common (anchor) session across all forms within a grade. The NCSC reading Pilot Phase 2 tests contained SR items, MSR items, and OR items. The reading forms included 25–29 items across the tested grades.

The session 1 common items (i.e., anchor set) were used to anchor performance on all forms at a grade level to a common scale. The anchor set quantified the relationship between student performance on the anchor set and student performance on all other items. The anchor set was not intended to cover the breadth and depth of the assessed content. Session 1 consisted of two literary passages and one informational passage at grades 3–5, and two informational and one literary passage at grades 6–8 and 11. Session 1 was coupled with one of four session 2 components: session 2A, session 2B, session 2C, or session 2D. Each session 2 included two passages selected to address varying complexity levels and to assess overall literacy skills. (For additional information, see Appendix 2-G: Pilot Phase 2 Blueprint and Forms.)

In grades 3 and 4, the Pilot Phase 2 included eight reading test forms, which included reading foundational items (i.e., word identification). The prioritized academic content for assessment was identical on all eight forms. With respect to the foundational item types, four forms were developed for verbal responding assessed by OR items and four forms were developed for nonverbal response assessed by SR items. The item type varied to allow students to demonstrate their word identification skills in a manner consistent with their form of communication. All other items across the verbal forms (1–4) and nonverbal forms (5–8) were the same for the corresponding form (e.g., verbal form 1 and nonverbal form 5) and were SR items. Information provided in a student's enrollment file determined the assignment of a verbal or nonverbal form. In grades 6–8 and grade 11, four reading test forms were developed.

Pilot Phase 2 ELA Tests

The Pilot Phase 2 ELA test forms consisted of both reading and writing content. In addition to the reading items, each form at all grades included four stand-alone writing SR items, one writing-prompt-based cluster of SR items designed at the lowest level of complexity, and two writing CR items designed at the second and third of the four levels of complexity. The NCSC ELA Pilot Phase 2 tests contained SR items, MSR items, OR items, as described for Pilot Phase 1, and Constructed Response (CR) items as described below and in the summary of the writing evaluation study later in this chapter. The ELA forms included 25–31 items across the tested grades.

The ELA CR items (i.e., writing prompt) asked a student to write a response to a prompt. NCSC designed the writing portion of the assessment to measure college and career readiness standards in writing which require that the writing process (including generate ideas, draft, revise, and edit) is used by students to demonstrate narrative, expository, or argument writing skills. Test development partner scorers rated a student's response on each of three traits—idea development, organization, and conventions—as providing full, partial, limited, or unrelated evidence related to the scoring criteria for each trait.

Test development partner scorers individually read and evaluated student responses to CR writing test items. NCSC leadership and content experts had upfront oversight and control of training materials and audited scorer performance at their discretion to ensure adherence to consistency in applying scoring criteria. Scorers who did not consistently apply the scoring criteria based on daily audits received additional training.

NCSC content and measurement experts collaborated with the pilot development partner to develop a scoring plan describing the procedures for pre-range-finding, range-finding, and scoring of the writing CR items administered in the NCSC Pilot Phase 2 ELA test. (For further information, see Appendix 2- H: Pilot Development Partner Report: Pilot Phase 2 Writing Constructed-Response Hand Scoring Plan.) The implementation of hand-scoring activities from pre-range-finding through scoring occurred through the test development partner's electronic imaging system. This allowed for online scoring of the student responses to writing CR items at the scoring site. The online scoring system maintained a database of actual student responses and the scores associated with those responses. The system also provided continuous, up-to-date monitoring of all scoring activities.

Pilot Phase 2 pre-range-finding activities were designed and implemented to (a) identify 15 representative student responses (with scores and annotations) across a range of trait scores for each item at each grade—for use in level-setting for range-finding participants, (b) identify additional student responses (unscored) for each item at each grade level that would make up the item packets reviewed by range-finding panels, and (c) familiarize the hand-scoring supervisors with the writing-scoring rubrics and their application at each grade for complexity levels 2 and 3. Pre-range-finding resulted in a set of student papers the test development partner used with range-finding panelists to define the application of the scoring rubrics and to provide panelists with an understanding of the complexity level and grade-level expectations.

NCSC held the range-finding meeting for Pilot Phase 2 at the pilot development partner's scoring facility. Scoring panels included both on-site and virtual participation by NCSC state partners assigned to grade-specific panels who examined responses to item prompts at both levels of complexity. The facilitation of the panels mirrored that of pre-range-finding, with the same hand-scoring supervisors leading the grade-specific panel discussions. The objective of this activity was to develop the sets of student papers—anchors, training, qualification, and validity sets— to help ensure that the test development partner's scorer training and scoring was consistent with NCSC standards and guidelines. (For further information, see Appendix 2-H: Pilot Phase 2 Test Writing Constructed-Response Hand Scoring Plan.)

Following the range-finding meetings and NCSC approval of all scorer-training sets, hand-scoring teams finalized the scorer training sets in the electronic scoring system based on the documented instructions from the range-finding panels. Scorer training materials for each item included the following:

- NCSC writing-scoring rubric;
- Anchor sets;
- Training sets;
- Qualification sets (2 rounds);
- Validity sets.

Scoring activities took place at the test development partner's facility immediately following pre-range-finding. NCSC team members remained on-site or were available virtually for the duration of scoring to work in partnership with the scoring supervisors and to ensure the implementation of the rubrics, procedures, and criteria for scoring as intended.

Scorers independently scored all three NCSC writing-scoring rubric traits (organization, idea development, and conventions) at individual workstations. Each scoring supervisor oversaw a team of approximately five scorers. Images of each student's response were automatically routed to a scorer based on his or her assignment. All student responses at each grade were selected for back-reading, which involved viewing responses recently scored by a particular scorer and, without knowing the scorer's score, assigning a score to that same response to review a specific scorer's work or a specific score point. Back-reading was useful in tracking specific areas of misunderstanding by a given scorer or group of scorers and assisted the scoring supervisor in knowing how to direct retraining activities. NCSC approved the dismissal of scorers whose performance remained below established standards for the particular item being scored. (For further information, see Appendix 2-H: Pilot Phase 2 Test Writing Constructed-Response Hand Scoring Plan.)

Pilot Phase 2 Mathematics Tests

The NCSC mathematics Pilot Phase 2 tests contained SR and CR items. The Pilot Phase 2 mathematics forms consisted of two sessions with 20 items per session. Session 1 was a common (anchor) session across all forms within a grade. Items across both session 1 and session 2 consisted of a mix of levels of complexity. Developers designed the mathematics forms to be similar in difficulty and to have a specific distribution for complexity levels—1, being the lowest, through 4, being the highest. Specifically, there was 20%, 35%, 35%, and 10% representation, respectively. NCSC partners constructed this distribution based on content and measurement expert recommendations, with input from the NCSC

TAC. (For more information, see Appendix 2-G: Pilot Phase 2 Blueprint and Forms.) Student responses to NCSC mathematics CR items were scored by the TA, and all other student responses to NCSC SR test items were machine-scored, as previously described in the discussion of the Pilot Phase 1.

Summary of Pilot Phase 1 and Phase 2 Results

Summary of Pilot Phase 1 Results

The NCSC pilot development partner completed analyses to address critical questions about the developed test content and inform item data reviews. Analyses conducted on Pilot Phase 1 test data focused on the following:

- Evaluating the appropriateness and accessibility of developed items when administered to a sample of the student population;
- Identifying (flagging) items for additional evaluation based on classical item statistics (difficulty and discrimination) or time demands—this information informed both item data review and ongoing development of the operational blueprint;
- Establishing evidence to understand item performance with regard to intended levels of complexity;
- Laying the groundwork for developing the NCSC test scales following Pilot Phase 2 testing.

Results indicated that content development processes generally succeeded in presenting grade- and age-appropriate content to students of the target population. No valid student case resulted in a score of 0; mean raw score, percent correct ranged from 47% to 67% of the total points per form in ELA and from 38% to 56% of the total points per form in mathematics.

Student counts per item were lower than anticipated, preventing some desired analytics (e.g., Item Response Theory (IRT) scaling, differential item functioning). However, classical item analysis and common-item comparison in mathematics indicated that items were generally appropriate and accessible to students. Additionally, item omit rates were very low (< 2%) and median times were generally reasonable, indicating feasible item and form length. It is important to note that because of a relatively small overall pilot sample (valid N=5161 across all grades and all assessed content areas), item statistics were based on sample sizes of 30 to 126 in mathematics and 40 to 135 in reading. (For additional information, see Appendix 2-D: Pilot Phase 1 Prioritized Sample Characteristics and Student Demographics and Appendix 2-I: Pilot Phase 1 Test Results and Item Statistics.) Consequently, item statistics must be interpreted in the context of these small samples. In addition, all of the SR items were either two-option or three-option items.

The pilot development partner completed analyses using Pilot Phase 1 results, addressing critical questions about the developed test content and using the data to inform item data reviews. Item flagging criteria were based on both item statistics (e.g., p-value, point-biserial correlations), as well as qualitatively observable problems (e.g., long scroll downs).

Summary of Pilot Phase 2 Results

Following the scoring of Pilot Phase 2 tests, the pilot development partner completed analyses of the Pilot Phase 2 results to inform item data reviews. Pilot Phase 2 analyses evaluated the appropriateness and accessibility of the developed items when administered to a sample of the student population and

identified (flagged) items based on classical difficulty and discrimination indices. In addition, the pilot development partner conducted analyses to determine the statistical properties of all items that were present on any of the forms. (See sections on Item Data Review, below in this chapter.) This information supports the creation of test forms that meet content and blueprint specifications as well as expectations for statistical characteristics, and it may be evaluated by test characteristic curves (TCCs) to help ensure statistical comparability across forms. (For further information, see Appendix 2-J: Pilot Phase 2 Test Results and Item Statistics.)

Anomalies in Writing Constructed-Response Collection

Approximately one week prior to the closing of the Pilot Phase 2 test window during an inspection of interim results data, analyses completed by the pilot development partner revealed some anomalies in the information entering their scoring system. For some CR items, text entry fields showed duplicate values—this affected the CR items at complexity level 2 for which the student was to compose text using multiple text fields within a single item template. Some of these items had as many as nine text fields. In all cases, the student response entered as the first text entry appeared across all of the text entry fields for the submitted item. This anomaly affected 14 of the 28 items, across grades, at complexity level 2 and approximately 38% of student responses. In addition, approximately 22% of students' responses to items at complexity level 3 were left blank. In the end, the final number of scorable student responses was far below the recommended minimum of 250 student responses per item. Based on these results, and a NCSC TAC recommendation to reduce testing burden for the first operational tests, NCSC decided to include writing CR items at complexity level 2 only as field test items within the spring 2015 test.

Overview of the Item Data Reviews

Following the administrations of Pilot Phase 1 and Phase 2, the pilot development partner computed classical statistics for each item piloted. The pilot development partner then compiled these statistics into data review books including individual item statistics and images of the items for use in item data review meetings. An item that had any statistics with values outside pre-established limits had an appropriate annotation (flag). In addition to judgments of content relevance, panelists evaluated the technical quality of items, checking each piloted item (including those with "appropriate" statistics) for flaws including the following:

- Inappropriate readability level;
- Ambiguities in the questions or answer options;
- Cluing within the body of the item;
- Keyed answers that were partially or wholly incorrect;
- Distractors that were partially or wholly correct;
- Unclear instructions;
- Factual inaccuracy;
- Any other concrete and material flaws.

NCSC sent out applications to every partner state during each of the recruitment phases. NCSC state partners identified experts from their states as potential panelists for item data reviews. Each group of potential panelists included subject matter experts with special education and/or content-area experience, educators with expertise applying college and career ready standards, and state/district-level

policy makers. Representatives from 16 states participated in the Pilot Phase 1 item data review and representatives from 10 states participated in the Pilot Phase 2 item data review.

Criteria for Pilot Phase 1 and Phase 2 Data Reviews

Prior to the item data review meetings, the assessment development partner applied item flagging criteria to identify items that may not have functioned as intended due to errors in content or rendering. The following list describes the item flagging criteria developed based on input from the NCSC TAC and NCSC partners.

- Difficult item: Low p-value < 0.50 Tier 1 (two answer choice options)
  - For items at the lowest complexity level in both reading and mathematics, there are only two answer choices. If the p-value was less than 0.50 for this type of item, the item was flagged and reviewed at the item data review.
- Difficult item: Low p-value < 0.40 Tiers 2–4 (three answer choice options)
  - For items at complexity levels 2–4 in both reading and mathematics, there are three answer choices. If the p-value was less than 0.40 for this type of item, the item was flagged and reviewed at item data review.
- Easy item: High p-value > 0.90
  - If the p-value was greater than 0.90 for an item in reading or mathematics, the item was flagged and reviewed at item data review.
- Low biserial correlation < 0.00
  - When student performance on an item did not correlate with student performance on the rest of the test, the item was flagged for a low biserial correlation and reviewed at item data review.
- Complexity reversal: items harder at the lowest level of complexity than at the highest level of complexity
  - The test development partner conducted an evaluation of item complexity for the item data reviews, which included ordering items by difficulty and evaluating the order in the light of items' complexity design. A reversal flag indicated when p-values were not in descending order from low to high corresponding to intended complexity. In ELA, there are four variations of each passage, one at each level of complexity. Each passage has an associated set of items developed at the same complexity level. The number of items associated with a passage varies by complexity level as well as by passage. Because there is not a direct one-to-one correspondence of items across the levels of complexity, the pilot development partner calculated mean p-values in order to identify instances of reversal in reading. The expectation was that the highest p-values would be associated with the lowest complexity items and lowest p-values with the highest complexity items. The test development partner flagged items for a reversal, and they were reviewed at item data review if the p-value for the level 4 item was greater than the p-value for the level 1 item.
- Distractor analysis: Proportion selecting distractor greater than proportion selecting key
  - Items were flagged for a distractor and reviewed at item data review when statistics for the answer choices revealed that students were being drawn to a distractor more often than to the correct response. Items with two possible correct responses were flagged when the p-value of a distractor was similar or higher than the p-value of the correct

response. This could indicate a mis-key (correct response not correctly noted), a second possible correct response, or a distractor with elements of a correct response.

During item data review, items were reviewed based on their statistical properties and qualitatively observable problems (e.g., long scroll downs).

Item Data Review Trainings

During the trainings with item data review panelists, the pilot development partner's psychometricians stressed the fact that a range of item difficulty helps ensure breadth of coverage from the standpoint of potentially being able to effectively sample the entire range of student ability. The pilot development partner's psychometricians examined item performance, including the percentage of students who answered each item correctly and the correlation between each item score and the total test score. Mean item scores for SR items that were less than 0.40 indicated that students selected the correct response less often than would be expected by chance. If an item also had a high point-biserial (high discrimination), the most successful students were more likely to answer it correctly. If the point-biserial was low (low discrimination), the test development partner's psychometricians indicated that success on the item was only weakly related to success on the test. However, the pilot development partner suggested rejecting items only when the item exhibited concrete and identifiable flaws. Items without such flaws remained in the bank for the NCSC spring 2015 operational assessment.

The pilot development partner introduced item discrimination to the data review panelists as being perhaps the most important statistic for item evaluation from a functional quality perspective. The pilot development partner emphasized that on an item-by-item basis, the ability of an item to draw distinctions between students of differing overall ability level was most desirable. The pilot development partner requested that data review panelists carefully scrutinize items with values at or below 0.00 before accepting the item as viable. In addition, the pilot development partner instructed panelists that a low level of discrimination without an identifiable cause was not by itself grounds for rejecting an item.

The pilot development partner's psychometricians also asked data review panelists to further evaluate data in light of the information presented as to how students responded to each item choice (for SR items). The response distributions served as a means of verifying the discrimination indices as well as offering insight into how well the SR distractors were functioning. The pilot development partner's psychometricians guided panelists to scrutinize distractors selected by a large number of high-scoring examinees to ensure that they were not confusing or partially or wholly correct.

While the pilot development partner provided these basic instructions to panelists for evaluating item quality, these specifications were guidelines rather than strict rules. Knowing that the statistics tell only part of the story, NCSC considered the professional judgments of the panelists were central in the consideration of whether they deemed an item as eligible for use in the NCSC AA-AAS. The test development partner emphasized to panelists that items should not be rejected solely based on their statistics, but should be rejected only if a concrete and identifiable flaw were found. In addition, the test development partner instructed panelists that items with "good" statistics could also be flawed, and if an item was found to have an identifiable problem, it should be rejected or accepted after revision even if it had statistics that appeared to be acceptable.

Each panelist reviewed each item and chose one of the following responses:

1. Accept as is;
2. Accept with revisions (provide suggested revisions or say where the change should occur);
3. Reject (provide justification).

The facilitators solicited feedback from the panelists for each item, encouraged discussion, and then recorded consensus results on an electronic PDF of each item.

Pilot Phase 1 Item Data Review Results

Pilot Phase 1 Item data review panelists reviewed the Pilot Phase 1 items and provided ratings for each item (i.e., accept, accept with revisions, reject, or do not use). (See Table 2-12. Summary of Items Reviewed During Pilot Phase 1 Item Data Review for a summary of ratings for the Pilot Phase 1 items). Each grade had a number of flagged items across the entire grade. If an item was "accepted," there was no change to the item and it was available for use in future forms. If an item was "accepted with revisions," the panelists may have suggested edits to the item. If the item was "rejected," the panelists determined that the item was flawed. If panelists determined that "Do Not Use" was the appropriate rating, the item was flawed and should not be used on any test form.

Table 2-12. Summary of Items Reviewed during Pilot Phase 1 Item Data Review

Content	Grade	Flagged items	Accept	Accept with revisions	Rejected	DNU
Reading	3	61	36	25	0	0
	4	74	41	33	0	0
	5	43	11	32	0	0
	6	37	11	26	0	0
	7	54	12	41	1	0
	8	42	19	23	0	0
	11	48	20	27	1	0
Math	3	63	57	6	0	0
	4	97	35	62	0	0
	5	96	57	39	0	0
	6	50	39	11	0	0
	7	68	52	16	0	0
	8	64	45	17	1	1
	11	71	56	15	0	0

Following the item data review meeting, each of the pilot development partners' facilitators, by grade and content area, met with the respective NCSC content experts and at least one participating SEA representative to review and approve/reject the panelists' recommendations. NCSC content experts and the participating SEA representative reviewed the suggested modifications or edits, and subsequently approved the change, approved the change with consideration, or rejected the change.

Actions taken after the item data review included edits to item response options, art, audio, item text, and passages. The pilot development partner's psychometricians reviewed recommended edits to items to determine if these would result in a need for additional field-testing of the items. If any change resulted in additional field-testing, the test development partner assigned a new identification number to the revised item. As a result, the original item retained its original statistics in the item bank, and the new

item was ready to be tested as a new item during Pilot Phase 2. Items accepted at data review following Pilot Phase 1 were eligible for use in Pilot Phase 2. These items had two sets of associated statistics in the item bank after Pilot Phase 2.

Pilot Phase 2 Item Data Review Results

Panelists reviewed the Pilot Phase 2 items and provided ratings for each item (i.e., accept, accept with revisions, reject, or do not use; see Table 2-13. Summary of Items Reviewed During Pilot Phase 2 item data review for a summary of ratings following participant reviews of statistically flagged items). Each grade had a number of flagged items. If the item was "accepted," there was no change to the item. The intact item could be used again on future test forms. If the item was "accepted with revisions," the panelists suggested edits to the item. During reconciliation, if panelists determined a need to revise the item, the item was revised and could be used on future test forms as a field-test item. If the item was "rejected," the panelists determined that the item was flawed. The item would need to be rewritten and field-tested again on future test forms.

Table 2-13. Summary of Items Reviewed during Pilot Phase 2 Item Data Review

Content	Grade	Flagged items	Accept	Accept with revisions	Rejected
Reading	3	4	3	1	0
	4	2	2	0	0
	5	5	4	1	0
	6	7	7	0	0
	7	2	1	1	0
	8	5	5	0	0
	11	3	3	0	0
Mathematics	3	23	23	0	0
	4	49	49	0	0
	5	45	41	4	0
	6	24	22	2	0
	7	26	18	8	0
	8	30	26	4	0
	11	26	21	4	1
Writing	3	0	0	0	0
	4	0	0	0	0
	5	1	1	0	0
	6	0	0	0	0
	7	7	6	1	0
	8	6	6	0	0
	11	6	5	1	0

Post-Item Data Review Process for Pilot Phase 2

The process used for revising the mathematics, reading, and writing items after Pilot Phase 2 item data review was as follows:

1. The NCSC mathematics and ELA content experts met with the respective test development partner's facilitators and one participating SEA representative after the group completed all reviews to discuss global edits and approve those global edits across grades.
2. All suggested edits to items were approved, approved with revisions, or were rejected.
3. All edits to items reviewed by item data review panelists and approved with revisions were updated in the NCSC assessment system.
4. Approved edits to the remaining items from Pilot Phase 1 that were not used on forms in Pilot Phase 2, were applied after the completion of Pilot Phase 2 edits.

Items accepted at data review from Pilot Phase 2 were eligible for use as operational items beginning with the spring 2015 administration.

Pilot Phase 1 and Phase 2 Implications

The *Standards for Educational and Psychological Testing* documents various sources of validity evidence (e.g., test content, response processes, internal structure, relationships to other variables, testing consequences) and highlights the importance of ensuring that the item and test form tryouts are as representative as possible of the population for which the test is intended (AERA, APA, and NCME, 2014).

The results of NCSC Pilot Phase 1 and Phase 2 item data reviews and research supported NCSC decisions related to the development of the NCSC item pool, test design and delivery, and development of score scales. Range of item difficulty results indicated higher success rates than anticipated across the full range of items, providing support for NCSC's item development model and demonstrating that students with the most significant cognitive disabilities can show what they know and can do on rigorous assessments of grade and age-appropriate academic content, whether they are just beginning instruction on the content or have already made a lot of progress.

Report of Additional Studies and Surveys

NCSC's long-term goal, as stated in NCSC's theory of action, is to ensure that students with the most significant cognitive disabilities achieve increasingly higher academic outcomes and leave high school ready for post-secondary options. Achievement of this goal required that NCSC have a strong research agenda to investigate and address the complex challenges inherent in developing an AA-AAS. A research-to-practice approach was relevant and necessary to improve and validate the assessment conceptual framework that was based on an increased understanding of how students who participate in the AA-AAS access, communicate, and develop competence in the general curriculum through well-designed studies.

To collect validity evidence and information to inform the design of the NCSC AA-AAS and related assessment materials, NCSC researchers developed a research plan that included multiple studies and surveys scheduled to occur throughout assessment development, Pilot Phase 1 and Pilot Phase 2, and the operational assessment. NCSC action research studies created opportunities to "check" the partnership's application of the model of learning as represented in the items for the target population.

The action research studies also created opportunities to evaluate item content quality, format, and clarity; analyze the range of student performance in the exemplar items; and garner survey feedback from teachers in the field to inform next steps in the assessment development process. Included below are brief summaries of the studies and surveys; the summaries include in brief the purpose, methodology, results, recommendations for future work, and potential impact on assessment design and development.

Exploring Students' and Teachers' Interactions with NCSC Task Templates

Following the completion of the design pattern and task template reviews, and before completing development of the NCSC item pool, NCSC researchers created three studies evaluating the task templates for mathematics, reading, and writing. These studies evaluated how students and teachers interacted with items and gathered evidence related to item complexity and usability for grades 3–8 and grade 11 across the 2012–2014 school years. Researchers used results from each of the three task template tryout studies to advise NCSC content experts and the item development partner on the accessibility of the items. They also used results to improve the items and passages by changing the question format, adjusting differences in the level of complexity across items in a family, or adjusting the stimulus materials accompanying the items. Researchers addressed three sets of research questions based on the topical areas of (a) task feedback, (b) student interaction, and (c) teacher administration.

According to the *Standards for Educational and Psychological Testing*, validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (AERA, APA, & NCME 2014, p. 15). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, analysis of item response time data, and evaluation of the reasoning processes that examinees use when solving test items (Embretson [Whitley], 1983; Messick, 1989; Mislevy, 2009). NCSC sought this evidence through these studies to confirm that the NCSC assessments would include accessible items that measured a range of student learning of academic knowledge and skills.

The findings from the studies provided NCSC content experts and the item development partner with feedback to improve items; using this feedback, they revised the use of some language, reviewed the complexity levels of the items within families, and improved the item directives. Recommendations for NCSC content experts and the item development partner included, but were not limited to, the following:

- Consider minimizing the number of materials received by TAs.
- Review directives to remove any superfluous words that increase item length.
- Consider the number of stimulus materials associated with items.
- Consider reducing the length of the passage or providing more visual supports.
- Consider most efficient ways to communicate the variable features through the TAM and training modules.
- Add accommodations to address accessibility for students with visual impairments.

Student Interaction Study: Exploring Student Interactions With and Teacher Perceptions of Mathematics and Reading Items

After completing item development and item reviews, NCSC researchers conducted the Student Interaction Study to acquire information related to the responses of students with the most significant

cognitive disabilities to NCSC mathematics and reading items. NCSC and the pilot development partner used the evidence collected from this study to inform the implementation of the Pilot Phase 2 test that occurred in the fall of 2014.

Researchers recruited teachers and their students with significant cognitive disabilities in two different areas of the country and collected validity evidence from two samples of students. The two student samples focused on students with specific levels of communicative competence in terms of expressive and receptive communication to allow for comparisons across these communication levels for mathematics and reading.

During the administration of items for this study, researchers directed teachers to use a cognitive laboratory approach, when appropriate, to ask students about the process they used to answer items. Researchers observed the teachers' administration of the items to their students and then interviewed teachers post-administration to gain detailed feedback about the content and difficulty of the items, the cognitive processes students appeared to apply, the administration process, student engagement, the effects of the technology platform on the administration process, and teachers' suggestions for improvement.

Researchers reported the results from both locations and compared the two samples across mathematics and reading. Based on the findings from this study, NCSC researchers provided recommendations for NCSC state partners and the item development partner.

Recommendations for NCSC state partners included the following:

- Work with teachers to understand student engagement.
- Consider conducting classroom observations to better understand student cognitive processes prior to implementing another assessment.
- For test administration, encourage TAs to use technology with which a student is most comfortable.
- Recommendations for the item development partners included the following:
- Collaborate with the technology team to address the issues related to item presentation.
- Consider whether rereading sections of a passage is necessary for all students.
- Clarify for teachers (who will communicate the information to students) how students can be prompted to use the item supports.
- Clarify the appropriateness of using paraprofessionals and provide training for those purposes.

Pilot Phase 1 End of Test Survey: Mathematics and Reading

As part of the Pilot Phase 1 administration, NCSC researchers developed an EOTS specific to each content area and requested the TA complete and submit a survey for each student administered a test in either mathematics or reading. TAs received directions in the Pilot Phase 1 TAM for completing and submitting the EOTS. The EOTS consisted of three different versions: the Reading EOTS, the Math EOTS, and the Accessibility EOTS (related to the administration of items to students who are blind, deaf, or deaf/blind). Researchers designed the EOTS to understand (1) the characteristics of students who completed the mathematics or reading Pilot Phase 1 test, (2) the opportunities these students had to learn grade-level academic content, and (3) TA perceptions of the administration procedures for the test.

Because students who participated in Pilot Phase 1 completed either the mathematics or reading tests, each student had only one survey completed by the TA. If a TA used the addendum to the TAM, which provided strategies and support to administer the Phase 1 test to a student who was blind, deaf, or deaf/blind, the TA was directed to complete the Accessibility EOTS for that student.

Following the administration of the spring 2014 Pilot Phase 1, NCSC researchers analyzed TAs' responses to the EOTS separately for the mathematics, reading, and accessibility surveys. Researchers merged the record containing the survey data provided by the student's TA with the student's test performance data for each student for which the survey and test data could be matched. They analyzed the results and provided findings specific to reading and mathematics for teacher characteristics, student characteristics, opportunity to learn, and test administration.

NCSC and the test development partner used the findings from the analysis of EOTS data to inform future decisions regarding the operational tests and the online assessment system. Based on the results for the Pilot Phase 1 mathematics and reading EOTS, NCSC researchers provided the following recommendations for NCSC state partners and the pilot development partner:

- Provide students with more instructional focus in the content areas assessed as well as in the use of computers and other electronic devices to build familiarity with online testing.
- Provide students with additional practice with the item format and the assessment platform.
- Consider making adjustments to the presentation of items in the online system and to the test directions.

Pilot Phase 1 End of Test Accessibility Survey

After administering the spring 2014 Pilot Phase 1, NCSC researchers analyzed TAs' responses to the accessibility survey section of the EOTS. The purpose of this analysis was to gain information about the accessibility of the assessment items for students who are blind, deaf, or deaf/blind. Researchers matched the accessibility section of the EOTS with mathematics and/or reading test data, when possible, and analyzed the combined data. NCSC and the pilot development partner used the findings from the analysis to improve the Pilot Phase 2 and the operational tests and to provide evidence for NCSC's validity argument.

TAs who reported using the Accessibility Addendum in the TAM, and who reported that the student who took the Pilot Phase 1 test was blind, deaf or deaf/blind, represented the target respondents for this study. The Accessibility Addendum provided administration strategies and supports to use with students who are blind, deaf, or deaf/blind. However, researchers determined that only 66 of the 283 TAs who completed accessibility sections were associated with students identified as blind, deaf, or deaf/blind. For purposes of the study, researchers focused on these 66 response sets. Researchers separately analyzed the remaining 217 accessibility survey response sets associated with students not identified as being blind, deaf, or deaf/blind.

The Accessibility EOTS questions focused on TA certification, teaching experiences, and experience administering AA-AAS; characteristics of students participating in the Pilot Phase 1; employed accessibility strategies during the administration of the items and which worked best; and what additional administration supports would have been helpful. Results indicated that the majority of TAs

held certification in special education and had five or more years of experience in administering an alternate assessment, but results did not include information as to TAs' experience in working with students who are blind, deaf, or deaf/blind. The survey results indicated that the most commonly used accessibility strategies were the same as those used for students identified as being blind, deaf, or deaf/blind and those who were not. The TAs who had students identified as being blind, deaf, or deaf/blind recommended that other teachers spend more time on instruction and testing preparation. Results of the survey indicate that many of the teachers with only a few years of teaching experience who administered the Pilot Phase 1 test to students identified as being blind, deaf or deaf/blind had more experience teaching students without disabilities than with disabilities.

Researchers recommended conducting further studies involving teachers administering the assessment to students who are blind, deaf, or deaf/blind. In addition, researchers recommended carefully clarifying the intended audience for the Accessibility Addendum in the TAM, training modules, and registration to ensure that TAs use the recommended accessibility strategies and supports for the correct students.

NCSC used results of these analyses, in conjunction with other data gathered from the Pilot Phase 1 administration and the expert/practitioner/state partner task force to help guide assessment revisions, training needs, and revisions to the Accessibility Addendum for Pilot Phase 1 test administration.

Spring 2014 Writing Evaluation Study

In spring 2014, as part of the overall Pilot Phase implementation, NCSC researchers conducted a small-scale writing evaluation study (WES) to pilot CR and SR writing items and to assess the administration process for writing. This study evaluated the writing items by comparing student performance across complexity levels of writing items, analyzing the interaction between writing item characteristics and student characteristics, and considering how the scoring process could best recognize students' writing skills. NCSC used the results from the WES to inform further development of writing items prior to the Pilot Phase 2 test and the operational NCSC AA-AAS.

Researchers designed the WES so that the student sample included students whose teachers represented the full range of current classroom writing instruction practices. Researchers assigned students to one of three groups, and each group contained students with expressive and receptive characteristics. Dependent on the group assignment, researchers directed TAs to administer CR and SR items of different difficulty in a specified order. After administering the items, NCSC researchers conducted focus groups with the TAs to obtain data that would help them determine ways to enhance the operational administration of the writing items. The final WES-related activity was a range-finding and scoring event to evaluate both student performance on the items and the quality and appropriateness of the writing rubrics and scoring procedures.

The results from the WES helped developers improve and refine the administration procedures (including TA directives and item supports), the scoring system for writing items, and the items themselves. Recommendations provided by focus groups and the analysis of results included suggestions for improving item elements: length, scaffolds and supports, scribing, and transcription. Recommendations for areas of further study included the use of accommodations, the use of AAC for

instruction and assessment, student performance as it relates to the extent of focused writing instruction, and the auditing of test administration.

Pilot Phase 2 End of Test Survey: Mathematics, Reading, and ELA

In conjunction with test administration in Pilot Phase 2, NCSC researchers asked TAs to complete an EOTS that accompanied the mathematics, reading, and ELA tests. Researchers designed the EOTS to understand (1) the characteristics of students who completed the particular Pilot Phase 2 test(s); (2) the opportunities these students had to learn grade-level academic content; and (3) TA perceptions of the administration procedures for the test. NCSC researchers used data from Pilot Phase 2 EOTS to inform future decisions regarding the AA-AAS and the online assessment system.

Following the administration of Pilot Phase 2, NCSC researchers analyzed TAs' responses to the EOTS. Researchers first merged the EOTS and test data to create one record for each student with each record containing the survey data provided by the student's TA and the student's test performance data. Researchers analyzed results and reported findings specific to mathematics, reading, and ELA for student characteristics, opportunity to learn, and test administration.

Based on results from the mathematics, reading, and ELA test records representing those students who completed Pilot Phase 2 tests and the corresponding EOTS, NCSC researchers recommended the following:

- Continue dissemination of NCSC instructional materials and provide additional teacher training, as the data indicated that students need more instructional focus on assessed content.
- Continue improving materials used to prepare for the administration of the NCSC AA-AAS.
- Encourage teachers to use the computer and electronic devices for instructional activities.
- Make the layout of the computer-based assessment more flexible (e.g., horizontal layout for item presentation).

TEST SPECIFICATIONS AND BLUEPRINTS

Test Specifications

The NCSC test specifications reflect NCSC's theory of action and overall assessment design. Consistent with the CCSS and the NCSC development processes and theory of action, the blueprints address specific characteristics of each content area. The operational design mirrored the Pilot Phase 2 forms and takes into account the NCSC-created items with varying levels of complexity. This test design is outlined in detail within the pilot section of this chapter.

Complexity

The NCSC content development processes addressed levels of cognitive and language complexity, specifically the CCSS and the heterogeneous characteristics of the target student population. The assessment items vary systematically in complexity, yet remain aligned to the KSAs behind the prioritized CCCs. NCSC designed its AA-AAS to capture student performance through two specific item design features: (1) levels of content complexity, and (2) degrees and types of scaffolds and supports. Items were built as item families, and each tier within the family addressed both the content complexity and the degree of scaffolding and support provided with the items. The items were written to measure a range of academic abilities within the target population. The array of item characteristics to facilitate

varying student needs includes reminders, examples, and models. These are provided to focus the student on the task and elicit a response without guiding the student's response. Please see NCSC Brief 6: *Age- and Grade-Appropriate Assessment of Student Learning* in Appendix 1-A for further detail on the item family structure. A range of items representing both differing levels of complexity and scaffolding and support were provided on each test form.

The NCSC Blueprint

The NCSC test blueprint was designed to be consistent with the NCSC theory of action, the principled design approach used to develop the summative assessment, and best practices in educational measurement. Tables 2-14a and 2-15a show the broad targets developed to guide the item development process and to inform test construction. They provide general guidance for identifying areas of emphasis in the development of the mathematics and ELA test forms. The blueprint tables in Appendices 2-K and 2-L incorporate the overall content distributions used for the development of the operational tests. Each grade level/content area is represented by a table that first describes the domain (e.g., operations and algebraic thinking) or text type (e.g., reading informational text), weights by domain and ELA strands and text types, grade-level content priorities (CCCs), item types, and numbers of items. The items for each form of the test in each grade and content were revisited following the operational assessment window in order to balance both the content requirements of the blueprints and the psychometric characteristics of the items. The core set of items on each form was established from this balance. Tables 2-14b and 2-15b show the actual distribution of content across the 2015 operational forms. The writing CR items (prompts) were field tested as part of the 2015 administration, so are not represented as part of the actual ELA distribution of content shown in Table 2-15b.

Table 2-14a. NCSC 2015: Guidelines for Distribution of Mathematics Content by Grade Level

Math Content Category	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 11
Operations and Algebraic Thinking	30%	30%	10%				
Number and Operations Base Ten	20%	10%	40%				
Number and Operations Fractions	20%	30%	20%				
Measurement and Data	20%	20%	20%				
Geometry	10%	10%	10%	10%	20%	30%	10%
Ratio and Proportions				30%	40%		
Expressions and Equations				20%	10%	20%	
The Number System				30%	20%	10%	
Statistics and Probability				10%	10%	20%	20%
Functions						20%	
Algebra and Functions							50%
Number and Quantity							20%

Table 2-14b. Actual distribution of Mathematics content on the NCSC 2015 Operational Assessments

Mathematics Domain	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 11
Operations and Algebraic Thinking	28-32%	28-32%	9-11%	-	-	-	-
Number and Operations Base Ten	17-23%	9-11%	34-40%	-	-	-	-
Number and Operations Fractions	17-23%	28-32%	17-23%	-	-	-	-
Measurement and Data	17-23%	17-23%	17-23%	-	-	-	-
Geometry	9-11%	9-11%	9-11%	9-11%	17-23%	28-32%	9-11%
Ratio and Proportion	-	-	-	28-32%	34-40%		-
Expressions and Equations	-	-	-	17-23%	9-11%	17-23%	-
The Number System	-	-	-	28-32%	17-23%	9-11%	-
Statistics and Probability	-	-	-	9-11%	9-11%	17-23%	17-23%
Functions	-	-	-	-	-	17-23%	
Algebra And Functions	-	-	-	-	-	-	47-52%
Number and Quantity	-	-	-	-	-	-	17-23%

Table 2-15a. NCSC 2015: Guidelines for Distribution of ELA Content by Grade Level

ELA Content Category	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 11
Reading Literary	27%	27%	30%	20%	20%	20%	15%
Reading Informational	27%	27%	30%	40%	40%	40%	45%
Reading Vocabulary	10%	10%	10%	10%	10%	10%	10%
Reading Foundational	6%	6%					
Writing	30%	30%	30%	30%	30%	30%	30%

Table 2-15b. Actual distribution of ELA content on the NCSC 2015 Operational Assessments

ELA Content Category	Gr 3	Gr 4	Gr 5	Gr 6	Gr 7	Gr 8	Gr 11
Reading Literary	30-33%	35-43%	34-44%	33-35%	30-32%	26-32%	24-27%
Reading Informational	30-33%	23-30%	28-34%	31-33%	35-38%	35-42%	40-44%
Reading Vocabulary	7-10%	10-12%	9-11%	13-14%	13-14%	13-13%	13-17%
Reading Foundational	6-7%	6-8%					
Writing *	19-20%	19-20%	19-21%	20-21%	19-20%	19-20%	19-21%

* The writing CR items (prompts) were field tested as part of the 2015 administration, so are not represented as part of the actual ELA distribution of content shown in Table 2-15b.

Mathematics

Mathematics items are aligned to prioritized grade-level content targets (CCCs), as described in the measurement construct development section. Mathematical knowledge is assessed across the CCCs through SR items and CR items. CR items were present at grades 3, 4, 5, 8, and 11 only. The need for CR items was determined by the focal KSA associated with a given CCC. Students might construct a graph, solve a problem, or complete a table in a CR item. CR items were scored dichotomously, with "correct" or "incorrect" options only.

In some cases, the selected focal KSA was best addressed by separating the skill into two parts. Therefore, two unique items are necessary to fully address a single content standard. For example, the CCC 8.DPS.1h1 asks students to both graph bivariate data using scatter plots and identify possible associations between the variables. Items were developed to address both parts of the standard, as found in the blueprint documents.

In addition, in mathematics there were items that were identified as not allowing the use of calculators to respond to the item. These items tended to be items related to computation in which the construct being assessed would be masked by the use of a calculator.

English Language Arts

ELA items in reading and writing are aligned to prioritized grade-level content (CCCs), as described above in the measurement construct development section. The distribution of ELA items related to various text types (e.g., literary, informational, and argument) aligns to the text type emphasis in reading and writing outlined in the CCSS. The project determined that all reading comprehension assessment items be presented in an SR format. Thus, to measure more complex reading skills, some SR items were built as a set of two or three sequenced items ("multipart") which, when combined, serve to measure the breadth of one prioritized content standard. In other words, in some instances the focal KSA aligned to a specific prioritized CCC is designed to have two or three SR items associated with it. In grades 5–8 and 11, some prioritized content standards require evaluation of content across more than one passage. These skills are measured using "paired passage sets." All paired passages are written in the informational text type.

The three CCCs prioritized for writing at each grade level consist of one CCC assessed by a CR item and two CCC assessed by SR items. The CR writing items are designed to measure a student's ability to generate a permanent product to represent organized ideas specific to a writing mode, supported with details or facts to clarify meaning and the use of Standard English conventions. The CR writing items were considered field-test items and did not count toward the student's score. Each state was given the choice of reporting information about the field test CR writing item to the student's teacher/school.

FORM ASSEMBLY

Process/Mechanism

Operational Design

The operational NCSC assessment program was designed to produce valid and reliable mathematics and ELA (reading and writing) scores. The mathematics and reading portions of the test comprised primarily SR items. Some grade levels in mathematics included CR items. Writing comprised

both SR and CR items. The CR items did not count in the student's operational score, as they were being field-tested. Within a particular grade and content area, the operationally scored items were administered on four overlapping forms—that is, the four forms included some items that were common to all four forms.

All items selected for the operational test were evaluated through inclusion in two phases of pilot testing administered to different student samples. Pilot Phase 1 testing occurred April–May, 2014, and Pilot Phase 2 occurred in October–November, 2014. Due to time constraints, the forms from Pilot Phase 2 were administered for the 2014–2015 operational assessment. These forms were judged sufficient to represent the blueprint and the assessment targets. Based on the results of the Pilot Phase 2 item data reviews, the scope of the edits to the items for the operational tests was not substantial. As noted, the one major change that was made to the Pilot Phase 2 forms for the operational assessment was the addition of writing sessions to create a full ELA test consisting of both reading and writing items. This occurred because the Pilot Phase 2 forms were constructed as either a reading test made up of only two sessions of reading items or an ELA test made up of one reading session and one writing session. In order to create a full ELA form, the two reading sessions from the reading test were used with the SR writing items (session 3). This formed the operational ELA test. In addition, a fourth session was added as an embedded field test of the CR writing items. Each grade and content area consisted of four fixed forms, each made up of two to four sessions.

Operational Core Items and Embedded Field Test Items

Using the operational blueprints, NCSC partners identified an initial list of core items for each of the grade-level and content-area test forms. Because of the inherent limitations of the statistical evidence from the pilots, all core item selections were made based entirely on ensuring the best possible content characteristics for each form. Based on the initial core set, each mathematics test form included five field-test SR items, and each ELA test form had between 5 to 7 field-test SR items. In addition, for ELA, the fourth test session comprised one of two CR writing items per grade level that were being field-tested. Appendices 2-M and 2-N outline the initial core items per content area, grade, and form.

Once the initial core items were identified, analyses using Pilot Phase 2 data were performed on the test forms and items, including classical statistics and item response theory statistics (e.g., TCCs and conditional standard error of measurement). Theses analyses provided a preliminary set of information on the test forms and items. No item substitutions were recommended through this process. NCSC partners reviewed the NCSC Initial Core Item List Report and expressed some concerns about the statistical equivalence of the four forms corresponding to the reading tests in grades 7, 8, and 11. Consequently, the partners recommended that the test development partner construct revised forms for these reading tests where statistical comparability would be emphasized without regard to content considerations. The results indicated that, as desired, the revised reading forms within grades 7, 8, and 11 exhibited a notably improved degree of statistical similarity with each other. Appendices 2-O and 2-P contain the two documents that were developed for and during the initial core item identification process: Initial Core Item List Report and Response to NCSC Steering Committee Concerns. These documents outline the process in more detail.

The initial core items were identified prior to the operational administration. Once operational item data were available, the core item sets for each grade and form were evaluated and adjusted as needed for any items with egregious data (e.g., negative item-total correlations or low item-total correlation with unsatisfactory model-data fit). An iterative process that involved the test development partner's psychometricians, content developers, special education staff, and NCSC organizational and state partners was used to determine the final core items. Items were evaluated for statistical characteristics and content. The post-operational evaluation of the initial core items resulted in the replacement or removal of items from the core item set in both mathematics and ELA. Across all grades and forms in mathematics, 10 core mathematics items were replaced by other items in the operational pool, and 19 core mathematics items were removed without replacement—making the core item set smaller by one or two items for a particular grade and form. Across all grades and forms in ELA, there were no core item replacements. However, six ELA core items were removed without replacement—again making the core item set somewhat smaller for a particular grade and form.

Specifications

Mathematics

Mathematics forms consisted of two sessions with 20 items per session. Of the 40 items, 35 were intended to be used as core and 5 were intended to be used as embedded field test slots. Session 1 was a common (anchor) session across all forms. Items were presented as a series of items tapping progressively higher levels of a construct based on increasing tier and difficulty information from Pilot Phase 1. As a result of Pilot Phase 1 quantitative and qualitative data and observations, tier representation was closely attended to for the Pilot Phase 2 pool. Mathematics tier distributions were as follows: Tier 1, 20% representation; Tier 2, 35% representation; Tier 3, 35% representation, Tier 4, 10% representation. Four forms of the mathematics test were created at each grade level.

Table 2-16. NCSC 2015: Mathematics Forms

Form	Session 1	Session 2
1	20 items	20 items
2		20 items
3		20 items
4		20 items

English Language Arts

For the operational assessment, all students needed to take an ELA test, not just a reading test. Therefore the forms and sessions for the operational assessment comprised the Pilot Phase 2 reading forms for sessions 1 and 2 and new sessions 3 and 4, which included writing items only. The new session 3 consisted of SR writing items and the new session 4 consisted of a single CR writing item that was being field-tested.

For ELA item complexity, the number of passages and the number of items were carefully attended to in the test design. Each ELA form contained five passages, with the first three passages used as common passage sets in session 1 across all forms. Two unique passage sets (with varying tier designations) were included in the second session: 2A, 2B, 2C, and 2D. The forms followed this passage tier representation: Tier 1: 1–2 passage and item sets; Tier 2: 1–2 passages and item sets; Tier 3: 1–2

passage and item sets; Tier 4: 1 passage and item set. This distribution addressed concerns regarding student engagement during testing and, alternately, helped to avoid student frustration and testing overload. Session 1 and 2 each comprised about 20 reading items. Within each form, four of the passage/item sets were intended to be used as core, and one passage/item set was intended to be used as embedded field-test slots. Session 3 included four SR writing items, and the Tier 1 writing prompt item was made up of four to six SR items. The number of SR items varied depending on the grade level. Session 3 was the same across all forms of the test at each grade. Session 4 was an embedded field test and comprised a Tier 2 passage and a CR writing prompt. There were two distinct Tier 2 writing prompts field-tested at each grade level.

Table 2-17. NCSC 2015: ELA Forms

Form	Session 1	Session 2	Session 3	Session 4
1	3 reading passages and items	2A: 2 reading passages and items (a/b)*	Writing 4 SRs and Tier 1 Prompt	Writing Tier 2 Prompt A
2		2B: 2 reading passages and items (c/d)*		
3		2C: 2 reading passages and items (e/f)*		Writing Tier 2 Prompt B
4		2D: 2 reading passages and items (g/h)*		

**Indicates unique passage/item sets*

Form Accessibility

The NCSC AA-AAS was developed to ensure that all students are able to participate in an assessment that is a measure of what they know and can do in relation to the grade-level CCSS. The NCSC Accessibility Committee developed accessibility guidelines and made recommendations for the selection of reading passage sets and mathematics items for students who have significant cognitive disabilities and who are blind, deaf, or deaf/blind. These items were included in Form 1 for each grade and content area, creating the most accessible form possible while still meeting the blueprint requirements.

At grades 3 and 4 ELA only, each test form included a set of foundational items that required students to decode words. Based on the population of students tested within the alternate assessment, foundational items were developed in two formats, a verbal and a nonverbal form. The forms included directions to present items in specific ways to adapt to a student's current mode of communication. Therefore, for ELA at grades 3 and 4, a total of eight forms rather than four were constructed, and each of the forms was provided in a verbal and a nonverbal format. In addition, braille versions of the word cards for the foundational items were created and made available to order for students who needed this accommodation.

As described in the previous section, core items were carefully selected to ensure inclusion of high quality items and comparability among the four operational forms. Details are given in the "Operational Core Items and Embedded Field Test Items" section of this chapter.

CHAPTER 3: SYSTEM COHERENCE AND ALIGNMENT TO GRADE-LEVEL CONTENT STANDARDS

Evidence that test content reflects the concepts that were meant to be measured is one of the critical sources of information necessary to support valid interpretations of test scores (AERA, APA, and NCME, 2014). Alignment is about coherent connections across various aspects within and across a system (Forte, 2013a, 2013b). Traditional alignment procedures describe the degree of intersection, overlap, or relationship among content embedded in state content standards, assessment, and instruction (Webb, 1997).

NCSC designed an AA-AAS assessment that reflected the belief that academic content from the standards at each grade level provides assessment targets for all students, including those with the most significant cognitive disabilities. NCSC did not develop extended academic content standards, but incorporated aspects of item design to vary the levels of content complexity. This approach ensures that all students, including those students with the most significant cognitive disabilities, have access to challenging mathematics and English language arts (ELA) content. Detailed information about the test design can be found in Chapter 2 (Test Development).

An illustration of the coherence among AA-AAS components is shown in Figure 3-1. The foundation for academic content for most current state assessments is defined by college and career ready standards (CCR), and additionally for the NCSC assessment, community readiness (CCCR) standards. NCSC's assessment design was based on the Common Core State Standards (CCSS), although several NCSC states conducted additional alignment studies based on their unique CCR standards.

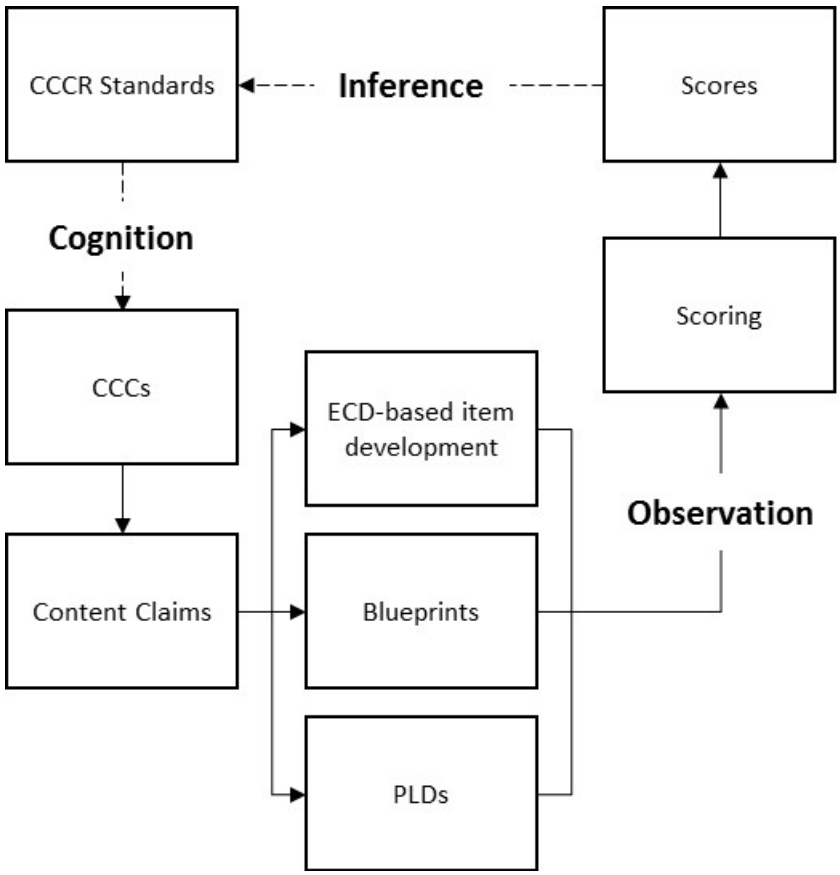
Using research and theories about how students with significant cognitive disabilities acquire academic skills and knowledge, core content connectors (CCCs) were generated that identified the most salient core academic content in the mathematics and ELA grade-level standards. The CCCs are not extended standards; they include the same content expectations found in the CCSS. NCSC further refined and reduced the number of CCCs for assessment purposes so that they reflected the necessary knowledge and skills students with significant cognitive disabilities need to reach critical learning targets or big ideas within the standards at each grade level. See NCSC Brief 7: *NCSC's Content Model for Grade-Aligned Instruction and Assessment: The Same Curriculum for All Students*, Appendix 1-A, for additional orientation to the NCSC content model.

NCSC generated claims about the inferences from the NCSC assessment scores based on the CCCs and associated research and learning theory. Claims served as the evidence-based rationale for the reduced breadth within each grade's CCSS. These claims can be thought of as a higher-order level of organization that is more attuned to both instruction and assessment. The claims were then used to generate:

(a) models for the tasks and test items that students would encounter;
(b) blueprints for how tasks and test items would be combined into test forms;
(c) performance level descriptors to convey the meaning of test scores in relation to the claims, and ultimately, to the academic standards on which those claims were based.

Taken together, the carefully-developed tasks and items, combined as indicated in the blueprints, offer students the opportunity to demonstrate their knowledge and skills in relation to all of the claims and allow for a range of performance as described in the performance level descriptors.

Figure 3-1. Coherence among AA-AAS Components



To maintain the integrity of the grade-level CCSS, NCSC carried out a series of studies to address these and several other key questions to test the logic chain and ensure that the assessments were aligned such that students' scores can be interpreted as reflecting the knowledge and skills defined in the standards and claims.

This chapter addresses the following alignment questions:

1. What is the degree of alignment between the CCCs and the grade-level CCSS?
2. What is the degree of alignment between instructional student learning expectations and measurement targets?
3. To what degree do the assessment tasks and items align to the grade-level CCSS?
4. To what degree do the assessment tasks and items align to the performance level descriptors (PLDs)?

5. How well do the claims align with grade-level content and provide useful information for tracking student progress toward achieving the knowledge and skills in the grade-level standards?

In designing a summative assessment aligned to grade-level CCSS, an iterative procedure was used and evidence was gathered to ensure that grade-level CCSS were the foundation of the NCSC AA-AAS. The first two alignment questions were examined early in the assessment development process to provide formative evaluation information for making adjustments before designing the summative assessment. The last three alignment questions were examined after the summative tests were created. A summary of the studies used to collect evidence supporting the alignment among the components of the NCSC AA-AAS system is presented in the following sections. Table 3-1 provides a list of the studies. Appendix 3-A describes the characteristics of the panelists for the studies; Appendix 3-B contains detailed study reports. Because the NCSC content model uses terms specific to the model, a glossary of terms used in the study summaries is at the end of this chapter as a tool for reviewers.

Table 3-1. Studies Related to Evidence of System Coherence

Study	Conducted	Claim for which evidence is provided
Relationship Studies	Mathematics – Summer 2012; Reading – Winter 2013; Writing; Summer 2013	The content and skills in the CCCs represent an adequate and appropriate sample of the grade-level CCSS. Evidence for alignment question #1.
UMASS Study of Coherence	Fall 2013	The targets for measurement provide information useful for tracking student progress in the CCSS and to teachers for providing instruction focused on academic expectations. Evidence for alignment question #2.
Task/Item Alignment Study	Summer 2015	The content and skills assessed by the NCSC AA-AAS represent an adequate and appropriate sample of the grade level CCSS. Evidence for alignment question #3.
Item Mapping Study	Summer 2015	The score reports are accurate and support appropriate inferences about student knowledge and skills. Evidence for alignment question #4.
Vertical Coherence Study	Summer 2015	The targets for measurement provide information useful for tracking student progress in the CCSS and to teachers for providing instruction focused on academic expectations. Evidence for alignment question #5.

Relationship of the CCCs to Grade-level CCSS Academic Content Standards: Alignment Question 1

Introduction

Evidence examining the alignment between the CCSS and the CCCs was collected early in the development phase to ensure the CCCs maintained the relationship to the CCSS. NCSC conducted the

Mathematics Relationship Study (conducted in 2012), Reading Relationship Study (conducted in 2013), and Writing Relationship Study (conducted in 2013). All studies used the Links for Academic Learning (LAL) model developed by Flowers, Wakeman, Browder, and Karvonen (2009) to evaluate alignment. The basic premises of the LAL method include the following expectations for AA-AAS:

- The assessments must be linked to grade-level content standards.
- The target for achievement must be academic content that corresponds to the student's assigned grade based on chronological age.
- Functional activities and materials may be used to promote understanding, but the target skills for student achievement are academically focused.
- Prioritization of the content, if it occurs, should reflect the major domains of the curricular area (e.g., strands of ELA) and have fidelity with this content and how it is typically taught in general education.
- The alternate expectations for achievement may include a focus on prerequisite skills or some partial attainment of the grade level, but students should still have the opportunity to meet high expectations, to demonstrate a range of depth of knowledge, to achieve within their symbolic level, and to show growth across grade levels or grade bands.

All panels consisted of general educators who were knowledgeable of the content standards and special educators who were knowledgeable about the student population. See Appendix 3-A for profiles of panelists. Details about rating scales, panelists, training, reliability of ratings, and procedures for implementing for these studies are available in full reports in Appendix 3-B.)

Given that alternate achievement is defined as reduced depth, breadth, and complexity related to the general achievement expectations, the results should reflect strong content centrality (i.e., degree of fidelity between the content in the CCCs and the content in the CCSS) and performance centrality (i.e., degree to which the CCCs and the CCSS have the same performance expectations) but possibly lower DOK of the CCCs (have a different cognitive complexity) than the general assessment DOK. The findings from the three studies are summarized below, and in more detail in Appendix 3B.

Mathematics Results

Results indicated that nearly all prioritized academic grade-level content targets (98.7%) were aligned to their intended CCSS. Raters' evaluation of the content centrality of the prioritized academic grade-level content targets matched to their intended CCSS resulted in all (100%) having a rating of all or part of the content found in the CCSS. Similarly, raters' evaluation of the performance centrality of prioritized academic grade-level content targets matched to the intended CCSS resulted in all (100%) having a rating of all or some of the performance found in the CCSS.

Raters' DOK ratings showed at least 50% of the prioritized academic grade-level content targets in grades 3–5 as having DOK levels at or above the corresponding CCSS. However, for grades 6-11, panelist rated the DOK level of less than 50% (18%-44%) of the prioritized academic grade-level content targets as at or above the DOK level of the corresponding CCSS. This difference is expected given that NCSC did not design the academic grade-level content targets to replicate the intended CCSS match in terms of level of complexity. Overall, the findings provide evidence to support alignment of the

prioritized academic grade-level content targets to CCSS with regard to content centrality, performance centrality, and DOK.

Reading Results

Panelists rated the intended CCSS as the best match for 90.2% prioritized academic grade-level content targets. Panelists gave ratings of some or all for content centrality to 97.8% prioritized academic grade-level content targets matched to the intended CCSS. They assigned a rating of none for content centrality to one prioritized academic grade-level content target because it covered only a small part of the CCSS. The panelists rated 93.5% prioritized academic grade-level content targets as having some or all performance centrality with their associated CCSS.

The distribution of DOK levels among the prioritized academic grade-level content targets broadly mirrors that of their associated CCSS, although, as expected from an AA-AAS, the academic grade-level content targets reflect slightly lower DOK levels. Across all grades, panelists ratings indicated the majority of prioritized academic grade-level content targets (56.5%) had the same DOK level as the target CCSS. Given the more narrow focus of the content targets, finding the same DOK for over half of these targets provides evidence of a strong overall match to the intended CCSS in terms of complexity. Overall, the evidence provided supports the alignment of the prioritized academic grade-level content targets to the CCSS for content centrality, performance centrality, and DOK.

Writing Results

Panelists aligned 95.2% academic grade-level content targets to their intended CCSS. Panelists chose the intended CCSS as the best match for all of them (100%). Panelists gave ratings of some or all for content centrality of all prioritized academic grade-level content targets and rubric-trait academic grade-level content targets. Similarly, panelists rated all of the prioritized academic grade-level content targets matched to their intended CCSS and all rubric-trait academic grade-level content targets as having "all" or "some" of the performance centrality found in the CCSS. When panelists chose a rating of "some" for the centrality indices, they often evaluated the academic grade-level content target as being one part of the overall CCSS, which is permissible given that the academic grade-level content targets are based on both the CCSS and the Learning Progressions Frameworks.

In general, panelists rated the DOK for the prioritized academic grade-level content targets slightly lower than that of the CCSS. A comparison of DOK ratings of the prioritized academic grade-level content targets and their intended CCSS illustrate fidelity to DOK levels of their CCSS. Panelists rated the majority of prioritized academic grade-level content targets (70%) at the same DOK level as their target CCSS. A comparison of DOK ratings of the rubric-trait academic grade-level content targets and their target CCSS revealed that the majority of rubric-trait academic grade-level content targets (63.9%) were given a DOK rating lower than their target CCSS, while only 30.6% and 5.6% of their academic grade-level content targets were rated at the same level or higher DOK level, respectively.

While the degree of DOK match was lower for the writing rubric-trait content targets, this finding is not surprising given the complexity of the intended CCSS and NCSC's focus on creating content targets teachers would find instructionally useful. This was facilitated, in part, by breaking these complex standards into more manageable, teachable components to build students' skills.

Given that NCSC did not intend for every prioritized academic grade-level content target to encompass all aspects of its intended CCSS, researchers expected ratings of "some" for both content and performance centrality as well as lower ratings for DOK. Consequently, the evidence from this study supported a reasonably strong relationship between the prioritized academic grade-level content targets and their intended CCSS with regard to content centrality, performance centrality, and DOK.

Summary

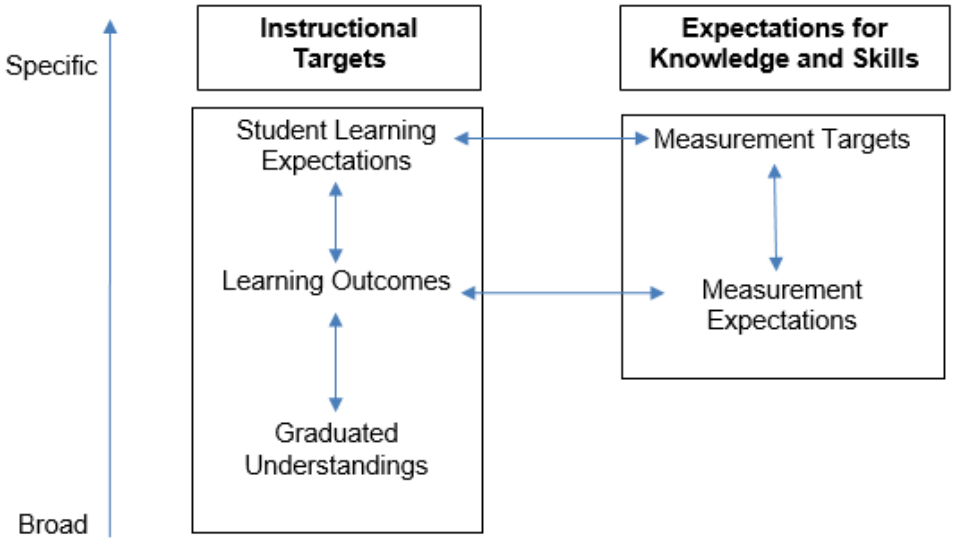
All the evidence suggested that the mathematics, reading, and writing CCCs had a strong relationship to the CCSS standards. In all content areas, the content centrality (i.e., degree of fidelity between the content in the CCCs and the content in the CCSS) and the performance centrality (i.e., degree to which the CCCs and the CCSS have the same performance expectations) suggested that the CCCs maintained the integrity of the grade-level CCSS. The DOK in the grade-level CCSS tended to be higher than the CCCs DOK, but this is acceptable given AA-AAS have a different cognitive complexity than the general education assessments. At all grades, there were CCCs rated at the high DOK levels suggesting students had access to challenging academic standards.

Alignment between Knowledge, Skills, and Abilities in Assessment to Student Learning Expectation for Instruction: Alignment Question #2

Introduction

Since instructional materials and professional development were an important component of the NCSC system, evidence was gathered to evaluate the quality of alignment among the instructional and assessment contexts of the system. A detailed description of the methods and findings can be found in the University of Massachusetts research report, Study of Coherence (conducted in 2013). The purposes of the study were to: (a) evaluate the degree of coherence across the learning outcomes and the measurement targets, (b) identify gaps or areas for improvement, and (c) inform the development of the grade-level PLDs. An illustration of the connections being investigated is shown in Figure 3-2.

Figure 3-2. Evaluation of Coherence among System Components



Note: Arrows indicate links among system components that UMASS investigated.

Findings

In all three content areas, findings suggested that the Graduated Understandings, learning outcomes, student learning expectations, measurement targets, and measurement expectations demonstrated a coherent relationship in terms of the skills, knowledge, abilities, and student expectations delineated across grades. Findings provided recommendations related to the vocabulary and precision of the language across components. In the context of gaps and redundancies, results indicated that a content expert might explain the rationale behind certain gaps or redundancies (e.g., that they were based on the progression of learning expectations) and be able to match some of the domain-specific terms and language that varied across documents. The findings indicated that consistency in the use of domain-specific terms and language would benefit examination of coherence. It was also suggested that NCSC address such consistency within other aspects of the NCSC system, such as an evaluation of the terms and descriptors used within the grade-level PLDs.

Following each recommendation was an action step taken by the NCSC partnership.

- Ensure the consistent use of technical terms across grades, outcomes, and expectations.
    - NCSC Response: NCSC content and measurement experts identified where the alignment and connections did and did not exist. Subsequent revisions of NCSC documents addressed the report's feedback to ensure consistency, clarity of language, and accuracy of the content.
- Ensure learning outcomes appropriately represent key aspects from the Graduated Understandings.
    - NCSC Response: Given that the learning outcomes intentionally represent a subset of skills considered critical aspects of learning connected to future content, NCSC partners collaborated with content and special education experts to review and revise the learning outcomes prior to the operational test.

- Ensure that specific aspects of writing techniques and mechanics are consistent across all grades, outcomes, and expectations.
    - NCSC Response: NCSC content and measurement experts examined identified inconsistencies. Subsequent revisions of NCSC documents addressed the feedback from the UMASS researchers to ensure consistency across grades, outcomes, and expectations.

Alignment of the Tasks and Items to Grade-level CCSS Standards: Alignment Question #3

Introduction

This section describes the traditional alignment study that examined the alignment of items/tests to the standards and were conducted after tests were developed. As with the previous alignment between the CCSS to CCCs, Links for Academic Learning (LAL) was used to determine the alignment of the items/tests to the grade-level standards. Panelists who were not part of NCSC were used to evaluate alignment in this study (See Appendix 3-A).

Four forms of the mathematics and four forms of ELA 2015 operational tests for grades 3-8 and high school were used in this study. At all grade levels, the mathematics forms contained 40 items. The number of items on each ELA test varied, depending upon the number of items that accompanied each passage. Each mathematics and ELA item corresponded to an academic grade-level standard. NCSC item developers created items at four levels of complexity to provide students multiple points of entry to show what they know and can do relative to each content target. Both the mathematics and ELA tests contained items that corresponded to all four levels of graduated complexity.

Details about rating scales, panelist, training, reliability of ratings, and procedures for implementing for the alignment study can be found in the full report (Item Alignment Study, Appendix 3-B). All panels consisted of general educators who were knowledgeable of the content standards and special educators who were knowledgeable about the student population (Appendix 3-A). The findings from the study are summarized below and organized by LAL criteria for alignment.

Findings

Criterion 1: The content is academic.

Researchers asked panelists to identify the CCSS to which the items aligned. If panelists indicated an item did not align to a CCSS, then panelists indicated if the content of the item was foundational. Both ELA and mathematics panelists indicated that 100% of the items aligned to a CCSS.

Criterion 2: The focus of achievement maintains fidelity with the content of the original grade-level standards (content centrality) and when possible, the specified performance (performance centrality).

Researchers compared panelists' ratings of the content centrality of the operational assessment items to the academic grade-level content targets (see Table 3-2). Panelists rated over 94% of the items in both content areas as a far or near link to the academic grade-level content target (a far link indicates a partial match, which is deemed acceptable within the LAL methodology (Flowers et al., 2009). Of those

items that panelists rated as having no connection to the academic grade-level content targets, most were at complexity level 1 and were aligned to Essential Understandings.

Table 3-2. Content Centrality of Items in Relation to Academic Grade-Level Content Targets

	Items	No Link		Far Link		Near Link		Linked to Target
Content	*N*	*n*	%	*n*	*%*	*n*	*%*	**%**
ELA	597	10	1.7	301	50.4	286	47.9	98.3
Math	699	40	5.7	263	37.6	396	56.7	94.3

Nearly 90% or more of all items demonstrated some or all of the performance expectations found in the academic grade-level content targets (see Table 3-3). Because NCSC designed items at complexity level 1 using the Essential Understandings, NCSC intended the performance centrality of those items to be lower than that of items designed using the focal KSAs. Thus, NCSC expected that for all items, overall performance centrality would be lower than overall content centrality.

Table 3-3. Performance Centrality of Items in Relation to Academic Grade-Level Content Targets

	Items	None		Some		All		Linked to Target
Content	*N*	n	%	n	%	n	%	**%**
ELA	597	42	7.0	293	49.1	262	43.9	93.0
Math	699	75	10.7	178	25.5	446	63.8	89.3

Overall, the ELA and mathematics operational items were aligned to the academic grade-level content targets. As expected with alternate assessment items, the performance centrality was lower than content centrality; however, most of the items maintained the performance expectations in the academic grade-level content targets.

Criterion 3: The depth of knowledge (DOK) differs from grade-level standards, but maintains high expectations for students with the most significant cognitive disabilities.

In examining DOK, researchers first calculated the frequency of item DOK ratings for the items by content area. DOK results for the ELA and math items indicated that panelists primarily rated items at DOK levels 2 and 3 (the memorization and performance levels). Panelists did not rate any items at the lowest or highest DOK levels 1 and 5 (attention and application, respectively), and rated very few at DOK level 4 (comprehension).

Researchers also compared the item DOK ratings to the academic grade-level content target DOK ratings to investigate if panelists rated items at, above, or below the DOK level of the corresponding target. Researchers found that the DOK level for more than half of the items was rated at or above the DOK level of the content target (see Table 3-4).

Table 3-4. Comparison of the DOK Ratings of Items and the DOK Ratings of Academic Grade-Level Content Targets

	Items	Above		At		Below		At/Above
Content	*N*	n	%	n	%	n	%	**%**
ELA	597	20	3.4	291	48.7	286	47.9	52.1
Math	699	73	10.4	333	47.6	293	41.9	58.0

In addition to examining the alignment of the operational items with the academic grade-level targets and either the focal KSAs or Essential Understandings, panelists rated the alignment among the focal KSAs, Essential Understandings, and the academic grade-level content targets. Panelists rated all of the focal KSAs as having some link (either far or near) to the academic grade-level content targets, and panelists rated all of the focal KSAs as possessing some or all of the performance expectations found in the academic grade-level content targets. The results also showed that the DOK levels for the focal KSAs and the academic grade-level content targets were nearly identical.

Panelists rated nearly all of the Essential Understandings as having either a far or near link to the academic grade-level content targets (95%) and as well as having at least some of the performance expectations found in the academic grade-level content targets (95%). As expected based on their purpose of being associated with introductory points of access to the grade-level content, the DOKs of most of the Essential Understandings (71%) were below the DOKs of the academic grade-level content targets. This is acceptable given that NCSC designed the Essential Understandings to provide students with entry-level access to the knowledge and skills of the academic grade-level content targets.

Criterion 4: There is some differentiation in achievement across grade levels or grade bands.

Across ELA grades 3–11, panelists indicated that the items changed to some degree in terms of all five definitions, and provided examples of change. Panelists at all grades who provided ratings reported at least limited or partial evidence of change in the broader, deeper, prerequisite, and identical relationships.

Across mathematics grades 3–11, panelists indicated that the items provided limited or partial evidence of change by all five definitions, and provided examples of change. Panelists in the grades 3–4 panel indicated that the items did not measure broader skills. Panelists in grades 8 and 11 panels indicated that no skills at grade 11 were identical to skills at grade 8—the skill discrepancy is reasonable given the gap between the two grades.

Criterion 5: The potential barriers to demonstrating what students know and can do are minimized in the assessment.

Special education experts indicated whether students who were blind, deaf, deaf-blind, or who used a nonverbal mode of communication could demonstrate their knowledge on the assessment (a) as

designed, with flexibility built into the tasks/items; (b) with accommodations; (c) with modifications; or (d) not at all. All education experts indicated students could do the alternate assessment as designed with flexibility built into tasks, and with accommodations. No special education experts indicated that students would need modifications.

Overall Analysis of Alignment

NCSC designed the operational items to assess the knowledge and skills of a wide variety of students with the most significant cognitive disabilities. All items referenced grade-level content, and panelists rated over 90% of the items as having a far (partial) or near (full or complete) link to the content targets. A majority of the items (93% in ELA and 89% in mathematics) maintained the performance expectations found in the academic grade-level content targets. Most of the items' DOK ratings were in the middle of the DOK distribution. When researchers compared the DOKs of the items to the DOKs of the academic grade-level content targets, over half of the items' DOK levels were at or above the DOK of the academic grade-level content targets. The focal KSAs and Essential Understandings had a strong link, in both content and performance, to the academic grade-level content targets. The DOK levels for the focal KSAs were almost identical to those of the academic grade-level content targets, and, as designed, the DOK levels for the Essential Understandings were below those found in the academic grade-level content targets. Overall, there was strong coherence among the operational items and the content targets for both ELA and mathematics, and there was strong vertical coherence in skills assessed by the items across the grade levels.

The study provided evidence that the assessment's operational items allowed students using various communication modes and with specific characteristics to access the items. Panelists indicated that the items were suitable for students who used various communication modes, and panelists indicated that no modifications were necessary to enable student access to the test items.

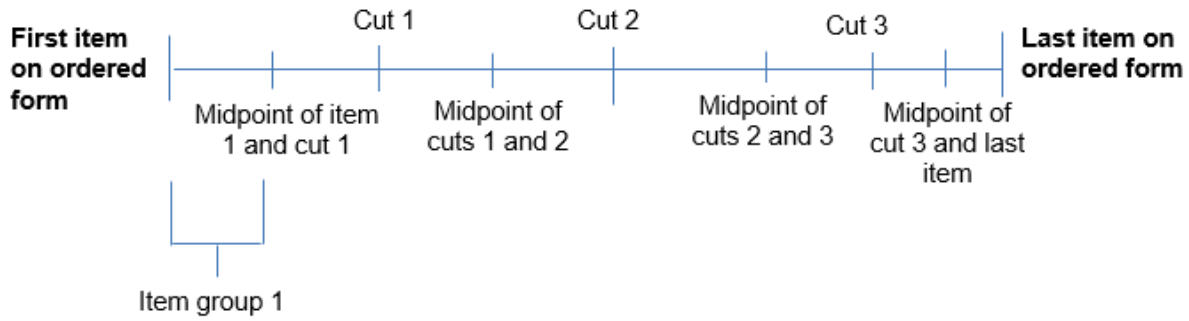
Alignment of NCSC AA-AAS Items to the Performance Level Descriptors (PLDs): Alignment Question #4

Introduction

Evidence was collected to examine the relationship between the knowledge, skills, and abilities (KSAs) accessed by items and the KSAs referenced in each level of the PLDs. A detailed description of the method, panelists, procedures, and results can be found in the full report of the Item Mapping Study in Appendix 3-B. Panelist profiles are provided in Appendix 3-A. Results from this study were used to provide evidence that the score reports are accurate and support appropriate inferences about student knowledge and skills.

Groups of items that corresponded to each of the four PLD levels were created based on cut scores established during standard setting, item difficulty estimates, and the ordered item booklets. The four item groups were mapped onto the scoring scale using the midpoint within the performance level. An illustration of the item grouping method is shown in Figure 3-3.

Figure 3-3: Placement of Item Groups on Scale



Findings

Panelists rated item groups based on their judgment of whether the KSAs in the item groups represent the KSAs in the PLDs. All items in the item groups were rated as having the (a) same PLD KSAs, (b) higher PLD KSAs, or (c) lower PLD KSAs Results are summarized in Table 3-5. For all content areas and grade levels, the majority of the item groups were rated as having the same KSAs as the PLDs, ranging from 57% to 78%. Some panelists indicated that some KSAs were missing in the item groups or the PLDs, but overall the overlap of KSAs found in the item groups and PLDs was acceptable. Results of this study will be used to inform future item development.

Table 3-5. Percent by Content Area/Grade of Item PLD Levels as Identified by Panelists

		Percent		
Content area	**Grade**	**Same**	**Higher**	**Lower**
ELA	3	75.8	6.8	17.4
	4	57.0	28.5	14.5
	5	75.0	10.7	14.3
	6	71.3	18.4	10.3
	7	73.6	20.0	6.4
	8	73.5	19.1	7.4
	11	77.6	11.0	11.4
	Total	71.9	16.4	11.7
Mathematics	3	74.8	24.5	0.6
	4	65.8	27.5	6.7
	5	78.8	16.3	5.0
	6	75.0	11.3	13.8
	7	70.6	17.5	11.9
	8	67.5	11.9	20.6

	11	77.2	16.1	6.7
	Total	72.8	17.9	9.3

VERTICAL COHERENCE STUDY: ALIGNMENT QUESTION #5

Introduction

As part of the process of developing a system that includes curriculum, instruction, and professional development resources aligned to college and career ready standards, the alignment between the measurement model, the instructional model, and the content claims was examined in the Vertical Coherence Study (conducted in 2015). The following section summarizes the findings of the study.

Findings

Results suggested that the focal KSAs/Essential Understandings across content areas and grades provided evidence in support of the content claims. Mathematics panelists reported that the focal KSAs/Essential Understandings provided full support for some, but not for all four mathematics claims. The reading panelists reported that the focal KSAs/Essential Understandings provided full support of the reading claim, and writing panelists indicated that the focal KSAs/Essential Understandings provided some evidence of the writing claim but fell short of providing full evidence in support of it.

In general, panelists also agreed that the student learning expectations provided full evidence in support of the content claims. Mathematics panelists agreed that the student learning expectations at all but two grades provided full evidence in support of the content claims. At grades 3 and 4, panelists reported that the student learning expectations provided full evidence for most but not all of the claims. Both reading and writing panelists indicated that the student learning expectations across grades provided full evidence in support of the claims.

Panelists indicated that the focal KSAs/Essential Understandings in all three content areas represented vertical coherence across the five vertical relationships (broader, deeper, prerequisite, new, and identical). In the higher grades, panelists reported some gaps in the focal KSAs/Essential Understandings, which they found reasonable given the gap between the NCSC assessment in grades 8 and 11. Panelists also reported few instances of new content. (Researchers had expected this finding, given the broad nature of focal KSAs/Essential Understandings in comparison with narrower targets such as items.) Only writing panelists indicated that many of the focal KSAs/Essential understandings at higher grades represented skills identical to those at the lower grades.

Panelists also indicated that the student learning expectations in all three content areas represented vertical coherence across the five vertical relationships. Unlike panelists' focal KSAs/Essential Understandings ratings, panelists' student learning expectations ratings indicated that there were no gaps across grade spans in the student learning expectations. As expected, panelists reported fewer new or identical student learning expectations across grades; the evidence panelists provided suggested that the student learning expectations grew primarily in terms of breadth and depth rather than in terms of introducing new expectations.

In their holistic judgment ratings, panelists across content areas agreed or strongly agreed that the focal KSAs/Essential Understandings within and across grades provided evidence of strong coherence

between the measurement model and NCSC's long-term outcome of college, career, and community readiness as expressed in the content claims. Panelists also agreed or strongly agreed that the student learning expectations within and across grades provided evidence of strong coherence between the instructional model and NCSC's long-term outcome of college, career, and community readiness as expressed in the content claims.

SUMMARY OF EVIDENCE

Throughout development and implementation of the NCSC system, key points to test the assumption that the AA-AAS were identified to provide evidence that support intended interpretations and uses. A summary of evidence is organized by the alignment questions.

1. What is the degree of alignment between the CCCs and the grade-level CCSS?

NCSC first investigated the relationship between the CCCs and the CCSS as articulated by the Learning Progressions Frameworks (Alignment Question #1). The results from the Mathematics, Reading, and Writing Relationship Studies indicated that the prioritized academic grade-level content targets and their alignment to intended college and career ready standards was strong with regard to content centrality, performance centrality, and DOK.

2. What is the degree of alignment between instructional student learning expectations and measurement targets?

To provide evidence for the evaluation of the Alignment Question #2, NCSC investigated the degree of coherence among system indicators and between system indicators and NCSC's overarching content claims. Study results indicated that a few gaps existed between the measurement and instructional targets, but overall the results suggested a strong connection between the focus of instruction and assessment.

3. To what degree do the assessment tasks and items align to the grade-level CCSS?

As evidence for the evaluation of Alignment Question #3, all tasks and items referenced grade-level content, and panelists rated over 90% of the items as having a far (partial) or near (full or complete) link to the content targets. A majority of the items (93% in ELA and 89% in mathematics) maintained the performance expectations found in the academic grade-level content targets. Most of the items' DOK ratings were in the middle of the DOK distribution. The focal KSAs and Essential Understandings had a strong link, in both content and performance, to the academic grade-level content targets. Overall, there was strong coherence among the operational tasks/items and the content targets for both ELA and mathematics, and there was strong vertical coherence in skills assessed by the items across the grade levels. Evidence supported that the assessment's operational items allowed students using various communication modes and with specific characteristics to access the items. Panelists indicated that the items were suitable for students who used various communication modes, and panelists indicated that no modifications were necessary to enable student access to the test items.

4. To what degree do the assessment tasks and items align to the performance level descriptors (PLDs)?

To provide evidence related to Alignment Question #4 related to the assumption that the score reports are accurate and support appropriate inferences about student knowledge and skills, NCSC conducted the Item Mapping Study to examine the extent to which the PLDs reflected what students had the opportunity to show evidence of, at varying levels, through their performance on the assessment. The focus of this study was on collecting evidence regarding the connections between the knowledge and skills the NCSC AA-AAS items measure and the description of student performance within and across categories of the PLDs. In general, results from the Item Mapping Study indicated that the knowledge, skills, and abilities captured by the items corresponded to and represented the content NCSC intended to measure, with minimal gaps in the information the assessments provided relative to the PLDs.

5. How well do the claims align with grade-level content connectors and provide useful information for tracking student progress toward achieving the knowledge and skill in the grade-level standards?

NCSC designed the Vertical Coherence Study to investigate the links between the measurement model, the instructional model, and the content claims, which represent the overarching focus for learning and assessment across the NCSC system. Specifically, the study assessed the links between the focal KSAs/Essential Understandings (measurement) and the content claims, and the links between the student learning expectations (instruction) and the content claims. The results indicated that the mathematics and ELA focal KSAs/Essential Understandings provided evidence in support of the claims, and that the mathematics and ELA student learning expectations provided evidence in support of the claims. In addition, study panelists agreed that the focal KSAs/Essential Understandings and student learning expectations provided evidence of strong coherence between the measurement model and NCSC's long-term outcome of college, career, and community readiness as expressed in the content claims. Results from the study confirmed that the learning expectations provided to teachers to guide instruction were connected to the expectations used to guide development of the NCSC AA-AAS.

Chapter 3 Glossary of Terms Used in NCSC Content Model

Core Content Connectors (CCCs): The CCCs clarify concepts in the Common Core State Standards (CCSS) by deconstructing the progress indicators in the Learning Progressions Frame work into teachable and assessable segments of content." The CCCs were designed to frame instruction for assessing students with the most significant cognitive disabilities while retaining the grade-level content focus of the CCSS and ensuring students are prepared for college, career, and community options.

Graduated Understandings: The Graduated Understandings articulate the big ideas, learning targets, and related instructional content within and across grades and help teachers conceptualize the progression of knowledge and skills students need to meet the rigor of the academic expectations expressed through the CCSS.

Learning Outcomes: Learning outcomes represent the "big ideas" and essential skills that students are expected to demonstrate by the end of a particular academic grade. NCSC partners wrote the learning outcomes to group together related indicators in a logical manner, help guide the development of lessons (which include key learning components), and tie the learning expectations to classroom instruction.

Student Learning Expectations: The student learning expectations integrate content knowledge and skills to describe student learning expectations by the end of each grade. Developers structured the student learning expectations to demonstrate the increasingly more sophisticated learning outcomes as students progress through the grades and provide a clear picture of end-of-year expectations for student learning.

Measurement Expectations: Measurement expectations are the prioritized content for assessment. The grade specific skills are intended to provide evidence to support the NCSC content claims with consideration of appropriate and practical content for a summative assessment, and congruency with the range of complexity reflected in the CCSS.

Measurement Targets: Measurement targets are narrative descriptions that integrate the selected assessment focal knowledge, skills, and abilities (focal KSA) of the measurement expectations. The organization of the focal KSAs in the measurement targets enabled NCSC to systematically organize the kinds of observations that would provide evidence of skill acquisition, relevant content, context, and features of task situations that allow students to provide the evidence associated with the prioritized content for each grade.

CHAPTER 4: TEST ADMINISTRATION

The NCSC AA-AAS was developed to ensure that all students with significant cognitive disabilities are able to participate in an assessment that is a measure of what they know and can do in relation to the grade-level Common Core State Standards (CCSS). NCSC's AA-AAS is a component of a system of curriculum, instruction, and professional development that allows students with the most significant cognitive disabilities to access grade-level content aligned to the CCSS. The test provides eligible students in grades 3–8 and 11 the opportunity to demonstrate what they know in mathematics and English language arts (ELA) — reading and writing.

ADMINISTRATION PROCEDURES AND GUIDELINES

Responsibility for Administration

The National Center and State Collaborative (NCSC) Alternate Assessment Based on Alternate Achievement Standards (AA-AAS) Test Administration Manual (TAM) provided the guidelines for planning and managing the NCSC administration for district and school personnel. The TAM provided the roles and responsibilities of the Test Administrators (TAs) and Test Coordinators (TCs) involved in the administration of the NCSC Test. The NCSC Directions for Test Administration (DTA) in both ELA and mathematics by grade and test provided specific directions for test administrators from scheduling sessions to preparing classrooms and testing students. Uniformity of administration procedures from school to school was ensured by using DTAs that contained explicit directions and scripts read aloud to students by TAs.

Administration Procedures

To ensure that the test was administered in a standardized manner, NCSC state partners and the test development vendor instructed TCs and TAs for the NCSC AA-AAS to read the TAM before testing and to be familiar with the instructions provided in the grade-level DTA. These documents were provided electronically on the NCSC Assessment Platform dashboard screen; some states also provided printed versions to the field. The NCSC Assessment System User Guide was also provided on the dashboard screen.

The TAM included a section highlighting aspects of test administration for the first year of operational assessment and checklists to help prepare for testing. The checklists outlined tasks for school staff to perform before, during, and after test administration. In addition to these checklists, the TAM described the testing material needed for administration, how to maintain security during administration, and secure shredding of printed materials after testing was complete. The TAM checklists for the TCs assisted them in providing oversight of the test, typically at the district level, to ensure that the test was administered as intended and that the TAs and students had the support needed for a successful administration. The TAM also included checklists for the TAs to use to prepare themselves, their classrooms, and the students for the administration of the tests. The TAM contained sections that detailed the procedures for TAs to follow for each test session as well as instructions for completing End of Test Surveys following the administration of each NCSC AA-AAS.

The DTAs provided TAs directions and a script to administer each item of the test, and were specific to the test form assigned to the student. To standardize the one-to-one administration of the test, TAs were instructed to follow these directions and script exactly as written. Item-specific reference materials, scoring

rubrics for constructed-response mathematics items in certain grades, and the open-response foundational reading items in grades 3 and 4 were also provided in the DTAs.

The NCSC Assessment System User Guide provided TAs with the information needed to successfully navigate the online system, outlining how to access the system and what needed to be completed within the system prior to administration and after administration. A large portion of the guide focused on how to navigate the system during the administration of the test itself in order to ensure a positive administration experience for the TA and student. The guide also instructed TAs on how to complete the End of Test Survey. The user guide included many screenshots in order to provide TAs with an easy to understand user friendly resource.

For TAs of students who had hearing/vision/sensory-motor disabilities that made access to some NCSC AA-AAS items challenging, additional information was available to guide administration (e.g., suggestions related to tactile symbols, object replacement, and interpreting guidelines) in the *Procedures for Assessing Students Who Are Blind, Deaf or Deaf-Blind: Additional Directions for Test Administration* (secure materials).

Allowed Student Manipulatives

The DTAs provided the TA with a list of the allowable manipulative materials (e.g., calculator, counting tiles, scratch paper) needed to administer the items in mathematics. NCSC allowed the use of a calculator, without restriction, for mathematics in grades 6-11. However, they did not allow the use of a calculator for the number operations portion of the test in grades 3-5.

Additional Administration Policies

During administration it occasionally became necessary to address questions and scenarios from the field with policy and procedural decisions. Policy decisions were made by the NCSC Steering Committee with direct impact upon relevant procedures. The additional administration policies, addressing such issues as accidental administration of all or part of a test to the wrong student, were distributed to states by NCSC leadership during the administration window. These contained the additional policy and procedural decisions made by the steering committee during this, the first, administration. Subsequently, these decisions were made known to the NCSC Service Center to keep field support aligned to NCSC policy decisions. Please see Appendix 4-A for a list of the additional administration policies.

Participation Requirements and Documentation

NCSC created a participation guidance document titled *Guidance for IEP Teams on Participation Decisions for the NCSC Alternate Assessment* that provided a student's IEP team with detailed information on the NCSC AA-AAS participation criteria. NCSC developed the participation guidelines to help ensure that the appropriate students were identified for inclusion in testing (see Standard 1.1, p. 25, of the AERA, APA, & NCME, 2014 *Standards for Educational and Psychological Testing*). A student designated eligible by the student's IEP team for participation based on these criteria was eligible to participate in the NCSC AA-AAS (see www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC_Participation_Guidance-Nov-2013.pdf).

The criteria for student participation in the NCSC AA-AAS reflect the pervasive nature of a significant cognitive disability. A student who participates in NCSC AA-AAS participates in both

mathematics and ELA. Thus, an IEP team is encouraged to consider all content areas when determining who should participate in this assessment. The guidance provides specific evidence to inform participation decisions indicating evidence that does not meet criteria for participation (e.g., category of disability or instructional setting) and three participation criteria that must be met: (1) the student has a significant cognitive disability, (2) the student is learning content linked to the Common Core State Standards, (3) the student requires extensive direct individualized instruction and substantial supports to achieve measurable gains in the grade- and age-appropriate curriculum (see: www.ncscpartners.org/Media/Default/PDFs/Resources/NCSC_Participation_Guidance-Nov-2013.pdf).

Parents/guardians are partners in IEP team meetings and engage in assessment participation decisions. Thus, they need accurate information about the NCSC AA-AAS. NCSC has a resource library for parents at www.ncscpartners.org/resources. This resource library was available to parents/guardians prior to the 2014-2015 school year to provide resources that informed participation decisions for individual students. NCSC AA-AAS State Coordinators were responsible for dissemination of the participation guidelines and additional resources to parents/guardians.

Learner Characteristics Inventory

For each student assessed, the TA completed the Learner Characteristics Inventory (LCI) prior to testing. The LCI consisted of 16 learner characteristics. See NCSC Brief 8, *Characteristics of Students with Significant Cognitive Disabilities: Data from NCSC's 2015 Assessment*, Appendix 1-A for a description of the characteristics of the students who participated in the assessment.

Documentation of Accommodations

The TAM also provided directions about allowable accommodations. Consistent with the NCSC accommodations policies, TAs were to use only those accommodations included in a student's IEP. The TAM included a table of NCSC accommodations divided into accommodations for assistive technology, paper presentation of items, use of a scribe, and sign language. For each accommodation, the TAM provided details on how to access further information related to its use.

NCSC developed the *Procedures for Assessing Students Who Are Blind, Deaf, or Deaf-Blind: Additional Directions for Test Administration* to provide additional details on planning for and implementing accommodations for testing a student who is blind, deaf, or deaf-blind. TAs accessed this guide through their state's Department of Education Coordinator after submitting the state's Special Forms Request document. To insure the correct and appropriate form assignment, State Coordinators submitted Special Forms Requests to the test administration vendor, triggering assignment of the most accessible form and shipment of braille to the appropriate students.

The accommodations allowed were "designed to remove construct-irrelevant barriers related to individual characteristics that otherwise would interfere with the measurement of the target construct and therefore unfairly disadvantage individuals with these characteristics" (AERA, APA, & NCME, 2014, p. 67) while maintaining the assessment of the construct of interest. Thus, students using accommodations received scores on the NCSC AA-AAS considered valid and appropriate to aggregate with those of other students. NCSC and its vendor paid careful attention to the potential effects of testing conditions on test score interpretations and adhered to the Standards for Educational and Psychological Testing (2014).

The TAM specifically stated that the use of any physical prompting by the TA, including hand over hand, would invalidate the results of the test for the student. The use of physical prompting was considered to be a modification or change to the DTA. Physical prompting was not allowed and explicitly deemed an inappropriate test practice and a test irregularity. The TAM provided TAs a complete description of approved accommodations. Table 4-1 (from the TAM) contains the approved NCSC accommodations for the 2015 test.

Table 4-1. 2014-15 NCSC: NCSC Accommodations (Appendix references in Table 4-1 refer to TAM)

Accommodations	Access Information
Assistive Technology (AT) *Student may use assistive technology devices for viewing, responding to, or interacting with the test items. The student and TA should use the AT device with the sample items to ensure that it functions properly with the NCSC Assessment System. The NCSC Assessment System supports various AT devices, such as alternate keyboard, switches and hub, head mouse, etc.*	Refer to NCSC Assessment System User Guide for Test Administrators for information about: ▪ Compatibility of NCSC Assessment System with Assessment Features; and ▪ Compatibility of NCSC Assessment System with Assistive Technology Devices.
Paper Version of Item/s *The use of a paper-based presentation of test item/s is a state-specific policy. Refer to Appendix A. State Specific Information. [Appendix A is in the TAM]*	PDF version of test item/s is available in the NCSC Assessment System. All printed assessment materials must be given to the TC for secure shredding upon completion of the test. (Please refer to Appendix A. State Specific Information. [in the TAM])
Scribe *This accommodation may be used for Selected-Response or Writing Constructed-Response Items.*	Refer to: ▪ Appendix B ▪ NCSC Assessment System User Guide for TAs ▪ ELA (Reading/Writing) DTA
Sign Language (e.g., ASL, PSE, SEE) *TA may communicate passages, items and response options using sign language to student.*	

System Assessment Features

The online NCSC assessment system afforded students a variety of assessment features to support student access to the test. Assessment features were either built into the system or were typically available on a computer. Assessment features could be enabled by the student or TA at the time of testing. The test is designed to have all passages, items and response options read to the student, either by the NCSC Assessment System's Audio Player or the TA.

The NCSC Assessment System User Guide for Test Administrators listed and described the assessment features available to students at the time of administration:

- Answer Masking
- Audio Player
- Alternate Color Themes
- Increase/Decrease Size of Text and Graphics
- Increase Volume
- Line Reader Tool
- Magnification
- Read Aloud and Reread Item Directions, Response Options, Passage

The TAM instructed TAs to refer to the NCSC Assessment System User Guide for Test Administrators for descriptions of the assessment features and the directions to enable the assessment features. TAs were also instructed in the TAM to become familiar with the assessment features and were informed that they may practice using the assessment features with the system sample items prior to administration.

Student Response Check

To ensure that the TA could clearly identify which answer a student indicated, TAs were instructed to administer the Student Response Check (SRC) prior to administration if there was any doubt that the student had clearly and consistently observable responses to test items. The SRC was a 3-question content-neutral task during which a student was asked to respond using his or her intended communication methods to determine whether the student demonstrated a clear and consistently observable mode of communication. For students who communicate using gestures, eye-gaze, or other modes of communication that may make using the computer difficult, TAs were instructed to conduct the SRC using the paper version, and were provided instructions on how to access that version. Step-by-step instructions for conducting the SRC were provided in Table 14 of the TAM.

The TA used results of the SRC as an indication of whether a student was currently capable of providing an observable response to items on the test. This was considered important because, if a student's response to a test item was not observable by the TA, the TA could not enter the student's response in the NCSC Assessment System. Implications of the SRC were described in the TAM, instructing the TA how to proceed based on the results of the SRC. TAs were directed to administer all test items to students displaying observable responses during the SRC. For students who did not display observable responses during SRC, TAs were instructed to administer the first four items in Reading Session 1 or Mathematics Session 1. If the TA observed a student response to at least one of the first four items, TAs were to administer all test items in both content areas. If the TA did not observe a student response to any of the first four items, TAs were to close the test session using the procedures described in the NCSC Assessment System User Guide for Test Administrators, which included selecting the closure reason "No observable mode of communication rule applied."

End of Test Surveys

After each content test was completed, TAs were asked to complete the End of Test Survey (EOTS) for that specific student and content area. The option to complete the EOTS became available to the TA after

the content-area test was submitted or closed. These surveys were developed to glean insights from the experience of each TA administering the NCSC AA-AAS, and were intended to provide additional data about how the test functions for students with unique and varying needs, student engagement with the test, and the opportunity to learn the content represented by the Common Core State Standards. The surveys also provided an option for teachers to give feedback on the testing experience.

In order for TAs to provide complete information on the EOTS, they were invited to keep a log noting their experience, and that of each student, with the assessment. However, they were directed to submit notes like this to the TC, as specified in state law, for secure shredding following completion of the End of Test Surveys.

Test Security

Maintaining test security is critical to the success of the NCSC AA-AAS and to provide accurate, fair, and comparable measurement for all students. The required administration trainings and the TAM emphasized test security. The TAM provided policies for implementation by the District Test Coordinator (DTC), School Test Coordinator (STC), and Test Administrator (TA) related to testing security and integrity as well as appropriate and inappropriate test practices. These policies also included compliance with state's test security protocols and procedures, and signing and submitting state-specific required test security agreements as outlined in state law and policy. The TAM gave a detailed explanation of all test security measures and test administration procedures. The TAM directed school personnel to immediately report any concerns about breaches in test security to the STC and/or principal. The STC, principal, or both were responsible for immediately reporting the concern to the DTC and their State NCSC Coordinator. State professional codes of ethics and state law provided the guidelines for determining the consequences for any irregularity. The reporting of violations, as described, ensured that test score invalidation was possible, when necessary.

Handling and Use of Secure Test Materials

All tests and DTAs were secure materials. NCSC recognized that secure handling of assessment materials was key to protecting the integrity, validity, and confidentiality of test items and student results and implemented procedures to ensure that unauthorized persons could not access or view NCSC AA-AAS content on the platform. The TCs were required to document the receipt of secure materials, check the lists of students, and shred all test materials at the completion of test administration. TAs were expected to:

- Keep all test materials in locked storage.
- Not disclose any actual test items to students prior to testing.
- Not provide answers to any test items to any students.
- Not change or otherwise alter a student's answer.
- Follow the Test Administration Manual explicitly.

Although all materials were delivered online, TAs were provided the option to print materials for ease of use. For such use, NCSC developed guidelines for printed test materials including, but not limited to, DTAs, test-specific printouts (e.g., mathematics and reading reference sheets), rubrics, and test forms. TAs were required to:

- Maintain such printed materials in a secure, locked location.
- Protect secure materials from view by other students, teachers, parents, school staff members, or individuals who may enter or work in the school building.

- Ensure secure transport of testing material from building to building.
- Not duplicate, reproduce, or share items.
- Give any printed test forms or other printed material to the STC for secure shredding.

Preparing a Secure Testing Environment

NCSC anticipated that students would be administered the NCSC AA-AAS in a one-on-one setting, most likely in their classrooms or a similar environment familiar to the students. The TAM provided the following guidance for ensuring an appropriate, secure testing environment:

- Administer the test only through the password-protected testing environment.
- Restrict student access to resources explicitly identified in the DTA.
- Ensure test items are viewable only by the student taking the test and the certified, licensed, and trained TA administering the test.
- Remove electronic devices and photography technology that could jeopardize test content in the test-taking environment.
- Ensure a quiet test-taking condition, void of talking or other distractions.

Access to secure test materials in the NCSC Assessment System was limited to registered, permissioned users. To attain 'registered' status, the user was created in the system by another authorized user. For example, the district-level TC (whose user status was created by the state-level user) was allowed to create a school-level TA's user status. To access the DTAs, which contained secure test items and, in some instances, other secure test materials, TAs were required to complete the training and pass the final quiz, as described in the section Administration Training. Additional information about permissioned access to secure materials may be found in the section Monitoring and Quality Control.

Test Administration Window

The test administration window was March 30–May 15, 2015. Most states administered the test during the entire NCSC window, while others started the testing window later due to state holidays. The test vendor delivered the online administration using the NCSC online delivery platform, following NCSC's test design requiring test administration in six separate sessions (four for ELA and two for mathematics).

The NCSC AA-AAS was not a timed test. Testing time varied for each student with testing paused and resumed, based on student needs. If a student exhibited frustration, lack of engagement, refusal to participate, or became sick during the administration of the NCSC AA-AAS, TAs were directed to pause the testing, and take a break—which may have been a few minutes to a few days, depending on the student's needs. NCSC protocols allowed the TA to pause and resume the administration of the NCSC test as often as necessary during the test window, based on a student's needs.

Session Structure

Test Administrators could begin with either the mathematics test or ELA test. Once a content-area test was started, TAs were required to complete that test before beginning the test in the other content area. Each content area test consisted of a set of testing sessions. Students were administered the test sessions in order for a given content area. ELA consisted of four test sessions (see Table 4-2) and mathematics consisted of two test sessions (see Table 4-3) at each grade level.

Table 4-2. 2014-15 NCSC: ELA Test Sessions

NCSC ELA Test			
Session 1: Reading	*Session 2: Reading*	*Session 3: Writing*	*Session 4: Writing*
Literary and informational reading passages and associated Selected-Response Reading items	Literary and informational reading passages and associated Selected-Response Reading items	Selected-Response Writing items	One Constructed-Response Writing item
Open-Response Foundational Reading items (Grades 3 and 4 only)	Open-Response Foundational Reading items (Grades 3 and 4 only)		

Table 4-3. 2014-15 NCSC: Mathematics Test Sessions

NCSC Mathematics Test	
Mathematics Session 1	*Mathematics Session 2*
Selected-Response Mathematics items	Selected-Response Mathematics items
Constructed-Response Mathematics items in selected grades	Constructed-Response Mathematics items in selected grades

Testing Calendar

To provide sufficient time to prepare for successful administration, key administration activities were scheduled prior to the administration window. A more detailed view of the testing calendar, including enrollment, training, and administration, is provided in Table 4-4.

Table 4-4. 2014-15 NCSC: Key Administration Activities

Dates	*Activity*	*Responsibility*
March 9, 2015	Practice Items Available in Sandbox	Test Administration Vendor / Test Administrators
March 9, 2015	TAM and User Guides Available Training Modules Available	Test Administration Vendor / Test Coordinators / Test Administrators
March 20, 2015	Enrollment File Upload Deadline	State NCSC Coordinators / Test Coordinators
March 10, 2015	DTAs Available	Test Administration Vendor / Test Administrators
March 18, 2015	Procedures for Assessing Students Who Are Blind, Deaf, or Deaf-Blind:	Test Administration Vendor / State NCSC Coordinators

	Additional Directions for Test Administration provided to states for dissemination to the field	
March 30-May 15, 2015	Administration Window Open Extended Service Desk Hours	Measured Progress / State NCSC Coordinators / Test Coordinators / Test Administrators

Administration Support

The TAM directed Test Administrators and School/District Test Coordinators to contact State NCSC Coordinators for assistance with general questions about the NCSC Assessment System or to obtain state-specific information. *Appendix A. State Specific Information* in the TAM listed contact information for each state's NCSC Coordinator and the link to state specific policies related to the NCSC AA-AAS.

To provide additional support to schools before, during, and after testing, the test administration vendor operated the NCSC Service Center. The NCSC Service Center provided a centralized location that those involved in test administration could call, using a toll-free number, to ask specific questions or report problems they may be experiencing. NCSC Service Center operators were responsible for receiving, responding to, and tracking calls, then routing issues to the appropriate person(s) for resolution. The NCSC Service Center was available for extended hours throughout the test window (from 8:00 a.m. to 8:00 p.m. ET, Monday through Friday) because the test was administered in multiple time zones.

NCSC Service Center

The test administration vendor provided technical support through the NCSC Service Center during administration of the NCSC assessment. The TAM directed TAs and TCs to contact the NCSC Service Center with questions pertaining to the NCSC Assessment System and test administration procedures. The NCSC Service Center responded to questions and requests from the field via phone and email throughout the administration window, with extended hours during key administrative activities such as registration. The NCSC Service Center's toll-free support number and email address were promoted to the field through the NCSC Assessment System and related communications.

Support was provided in a tiered manner, where Tier 1 support represented direct support to the caller by NCSC Service Center representatives, Tier 2 support represented promoted support by the program management team for items such as policy questions, and Tier 3 support represented technical requests that were escalated to the technology vendor for attention. Wherever possible, callers were directed to the appropriate section of the TAM or NCSC Assessment System User Guides, available to users within the NCSC Assessment System.

All activity was tracked in the NCSC Service Center log, and included in weekly status reports that were provided to NCSC leadership and operational states. These reports summarized activity, ticket requests, call analysis data (call duration, hold time, etc.), and per-grade/content and per-state test status summaries throughout the administration window. A sample report is provided in Appendix 4-B.

Additional Supports

In addition to the NCSC Service Center, the test administration vendor program management team periodically provided direct phone and email support to the state leads. In cases where logistical or procedural support was needed, program management worked with State Coordinators to resolve questions or issues. In cases with policy or consortium-wide implications, however, program management referred the state lead to the NCSC Steering Committee and/or NCSC leadership.

The test administration vendor also provided support and direction in the form of tip sheets that were created, as warranted, before or during an administration activity. All tip sheets were reviewed by NCSC prior to release, and provided to state leads for distribution throughout their states. Tip sheets were created to provide more specific directions than included in the TAM for specific scenarios or needs that arose during administration, as follows:

- Who to call for help
- How to capture evidence if you don't have a webcam
- How to move a student to a new school or class
- How to pause testing (specifically, when uploading evidence)
- How to request reopening a closed test
- How to request special forms during the assessment window
- How to unlock your user account (after 5 failed login attempts)
- How to update student grade assignment
- Preparing for the end of the administration window
- Submitting test results
- System timeout (after 10 hours of inactivity)

A messaging system in the NCSC Assessment System was implemented to notify users of system-wide or assessment-wide messages. Upon logging in to the system, a message appeared at the top of the screen that notified users of system information and upcoming system activities, such as known issues and scheduled system maintenance, as well as courtesy messages and errata notices.

Errata Notices

Due to errors discovered during administration, errata notices were issued for three items:

- Grade 3 ELA – an extraneous item was discovered on Form 4 version for verbal students: Session 1, Item 6
- Grade 7 ELA – an audio issue was discovered for the passage part related to Form 4, Session 2, Item 5
- Grade 4 Mathematics – an issue was discovered with the DTA for Form 4, Session 2, Item 8.

Once an issue was discovered, the test administration vendor worked with NCSC to determine the appropriate course of action, after which time an errata notice was created. Each errata notice identified the impacted item, a description of the issue, instructions to TAs using the affected form(s), and instructions for TAs to determine which form was assigned to a student. In each case, based on discussion with NCSC, the TAs were instructed to skip the impacted item and informed that omission of the item would not affect how students are scored on the assessment. None of the impacted items were "core" items, used in determining student scores, and thus had no impact on student scores. The errata notices were issued in a manner similar

to tip sheets and were also stored and accessed from the NCSC Assessment System dashboard, through a dedicated messaging system that appeared at the top of the screen for users who were logged in. This message, which persisted throughout the administration window, notified users that errata notices had been issued, and provided a link to where all the notices could be accessed in full.

ADMINISTRATION TRAINING

Training

NCSC adhered to the premise from the Testing Standards (AERA, APA, & NCME, 2014) that a key consideration in developing test administration procedures and manuals is that the test administration should be fair to all examinees. Because the NCSC AA-AAS was a computer-administered test, the administration procedures were consistent with the hardware and software requirements of the test specifications. NCSC required completion of training by all TAs and TCs to support standardized test processes and procedures. NCSC provided ancillary testing materials outlining specific practices and policies including (a) Test Administration Manual (TAM); (b) NCSC Online Test Administration Training; (c) NCSC Assessment System User Guide for Test Administrators; (d) NCSC Assessment System User Guide for Test Coordinators; and (d) grade and content specific Directions for Test Administration (DTA). TAs and TCs received both the online training and the supporting documents to ensure fidelity of implementation and the validity of the assessment results as well as to help NCSC prevent, detect, and respond to irregularities in academic testing, and testing integrity practices for technology-based assessments.

NCSC's test administration vendor made the modules available prior to the beginning of the test window and throughout the testing window. Training modules were customized to address the specific responsibilities of the TA and to provide important information from three documents TAs were required to use: (1) Test Administration Manual, (2) Directions for Test Administration (DTA), and (3) NCSC Assessment System User Guide for Test Administrators. NCSC developed 13 training modules for TAs and four for TCs (see Tables 4-5 and 4-6).

Table 4-5. 2014-15 NCSC: Modules for Test Administrators

Module 1: Training Requirements and Responsibilities of Test Administrators
Module 2: Overview of NCSC AA-AAS (Test) and Testing Integrity
Module 3: Optimal Testing Conditions and Assessment Features
Module 4: Test Accommodations and Procedures for Assessing Students Who Are Blind, Deaf, or Deaf-Blind: Additional Directions for Test Administration
Module 5: Navigate the Assessment System
Module 6: Before Test: Complete Demographics, LCI, and Accommodations
Module 7: Student Response Check
Module 8: Student Experience in the NCSC Assessment System
Module 9: Mathematics DTA – Administer the Test
Module 10: ELA DTA: Reading – Administer the Test
Module 11: ELA DTA: Writing – Administer the Test

Module 12: Upload Evidence for ELA Constructed Response Writing Item
Module 13: Submitting or Closing a Test, Accommodations- After Test, and End of Test Survey

Table 4-6. 2014-15 NCSC: Modules for Test Coordinators

Module 1: Responsibilities of Test Coordinators
Module 2: Overview of NCSC AA-AAS (Test) and Testing Integrity
Module 3: Navigate the NCSC Assessment System
Module 4: Create Users and Organizations

All online training recordings were accessed by TAs and TCs through the NCSC Assessment System. It was a requirement that the online training modules be viewed in sequence and one module had to be viewed before the link to the subsequent module would become accessible. Once a module was accessed, that module would be marked as complete in the NCSC Assessment System and the link to the next module in the sequence was available. Once all 13 modules were marked as complete, an end-of-training final quiz would become available to TAs within the NCSC Assessment System. TCs were required to complete the online training for TCs but were not required to take the end-of-training final quiz.

Additionally, TAs were instructed to become familiar with the online system by accessing the system sample items. NCSC content and measurement experts developed a set of sample assessment items for teachers, administrators, and policymakers. The sample items did not address all assessed content at each grade level, and were not representative of every item type. Rather, NCSC developed the sample items to provide a preview of the array of items and to illustrate multiple item features supporting the ways in which students with a wide range of learner characteristics interact with the assessment process. In addition, NCSC developed an electronic presentation to allow viewing of sample test items, and shared the presentation with state partners for inclusion on each state's Department of Education website.

Certification

At the end of each online training module for TAs and TCs were quiz questions pertaining to information from the module. The quiz questions were included as a review of the content and to prepare TAs for the end-of-training final quiz. The end-of training final quiz was accessed via the NCSC Assessment System following completion of all online training modules. TAs were required to take the end-of-training final quiz which covered content across all modules; they had to obtain a score of 80% or higher in order to be provided access to secure test administration materials. If TAs did not fulfill this certification requirement, they were not allowed access to the secure test materials. The TAs were notified within the NCSC Assessment System if they had passed or not passed the end-of-training final quiz. TAs were allowed multiple attempts to obtain a score of 80% or higher on the end-of-training final quiz. Some NCSC State Partners did not require Test Administrators who had previously completed training, and participated in Pilot Phase 2, to complete the online training modules; however, all NCSC State Partners required Test Administrators to complete the end-of-training final quiz and obtain a score of 80% or higher in order to access the secure test materials.

MONITORING AND QUALITY CONTROL

To ensure that proper testing procedures and appropriate test practices were maintained throughout administration, numerous measures were taken both to communicate participants' responsibilities and to monitor the appropriateness, accuracy, and completion of key procedures and tasks. The TAM explained to TAs and TCs that each person participating in the assessment program was directly responsible for immediately reporting any violation or suspected violation of test security or confidentiality by notifying the school or district TC. TCs were then instructed to follow state procedures regarding reporting the issue or suspected issue; however, district TCs were informed that they must report to the State NCSC Coordinator any incidents involving alleged or suspected violations that would be considered a serious irregularity. The TAM further explained that the consequences for inappropriate test practices would be determined by their state's professional codes of ethics and state law.

The online NCSC Assessment System contains built-in measures to ensure proper testing procedures, as seen in the session-based test design. As described in the *Session Structure section of this chapter*, tests were administered in item groupings referred to as test sessions. Immediately prior to completion of a test session, the online system displayed an End of Session screen that notified the TA of the number of questions left unanswered in that session, allowing the TA to ensure that no student responses were inadvertently omitted prior to leaving the session. The TA could navigate back to an item in the session and continue working or pause the test at that item in order to return to that session at a later time. However, the TA was notified that the system would effectively "lock" the session once it was completed by clicking Save & Exit on the End of Session screen. This prevented them from returning to the earlier session, which ensured that items in a subsequent session could not be used to inform responses to items in the previous session.

Throughout the administration window the test administration vendor monitored and provided weekly updates to NCSC on the technical performance of the NCSC Assessment System, test statuses across operational NCSC states, and trends identified in support calls. This provided a mechanism for concerns to be identified early and the appropriate measures to be taken, such as creation of tip sheets and the creation of additional NCSC administration policies, as described earlier in this chapter. This high level of communication and responsiveness throughout the assessment process contributed to a proper and valid administration of the NCSC AA-AAS.

RESULTS

During test administration it became clear that certain decisions and actions would be desirable to ensure the validity of the test and its results. Some insights applied to aspects of the administration window directly, while others informed the decision-making process for creating the decision rules.

As trends emerged in both NCSC Service Center requests and field communications with the State NCSC Coordinators, NCSC was able to identify unforeseen process questions and administration scenarios that required policy decisions, as described in the section Administration Procedures – Additional Administration Policies, or simply process clarifications. In cases where common process questions were evident, NCSC and the test administration vendor collaborated to craft clear instructions that were then distributed either to the field through the State Coordinators or to the NCSC Service Center representatives, whichever was deemed a more appropriate and helpful response to the issue at hand.

Other trends informed the data clean-up processes and the development of the decision rules. For example, qualitative data from the field, such as the previously unforeseen administration scenarios, reassigned tests due to grade corrections or special forms requests, and the potential for tests still in progress at the end of the administration window, emphasized the importance of considering all potential administration scenarios when determining the decision rules. A full description of the final set of decision rules is included in Chapter 9 – Reporting Interpretation and Use.

Beginning with a thorough set of online training modules for TAs and TCs, and ending with appropriate and informed data analysis and reporting, the processes for test administration were integrated and consistent, and reflected responsiveness to needs observed in the field.

CHAPTER 5: SCORING

ITEM SCORING PROCESS

Overview of Scoring Processes

The NCSC operational tests were completed through an online administration. Students responded to a variety of item types, including selected-response (SR), multiple choice, and constructed-response (CR) items. In addition, all students responded to a constructed-response writing item. The student responses were entered directly into the NCSC Assessment System and/or uploaded into the system through scanning or webcam. The SR items were scored according to the answer keys provided in each test package.

Student responses to the writing CR items and any uploaded material were exported from the platform and imported to the test administration vendor's *iScore* system. Through *iScore*, qualified scorers read and evaluated student responses and submitted scores electronically. The processes by which images were logged in, scanned, and uploaded into *iScore* provided anonymity to individual students and ensured random distribution of all responses during scoring. Details on the procedures to hand-score student responses are provided below.

Scoring Processes and Rules for SR and CR Items in Mathematics and SR Items in Reading and Writing

Overview of scoring process by item type

SR: Reading, Mathematics, and Writing

Selected-response (SR) items (multiple choice) were presented to students in a standard format. All directions and materials needed for administering SR items were provided in the secure Directions for Test Administration (DTA) that accompanied each test form. Test Administrators (TAs) received training in the administration and scoring of selected-response reading and mathematics items in online training module 9 (Mathematics DTA – Administer the Test) and module 10 (ELA DTA: Reading – Administer the Test). In module 11(ELA DTA: Writing – Administer the Test) TAs were given instruction to refer back to module 10 and follow the same protocols outlined for the SR reading items with the SR writing items. The DTA provided the full items, including the teacher scripts, to be read aloud to the student and any direction to the teacher related to the item and item set up, such as what to point to in the item as the script was read to the student. Every item was presented in the following order:

- Item stimulus (which could include a passage, passage part, picture, graphic, or other illustration)
- Item question
- Answer options presented in stacked or vertical formation

Students selected a response from the options in a variety of ways (e.g., using the computer mouse, verbalizing, gesturing, using eye gaze or communication devices, assistive technology, etc.). Many students entered responses directly into the NCSC Assessment System. If the student had the scribe accommodation, the scribe entered the student-selected response on behalf of the student.

The SR items were scored according to the answer keys provided in each the test package. All item responses were exported from the system and provided to the test administration vendor's Data and Reporting Services (DRS). DRS then applied the scoring rules. Items were scored as correct or incorrect, with the

majority of them contributing a score of 1 or 0 to the content area raw score. The SR items for the Tier 1 Writing multi-part selected response suite were treated as a set for scoring purposes. Each item set consisted of 4, 5, or 6 items. In all cases, a student's score on the item set ranged from 0 to 2 points. If the student got none of the items correct he or she earned 0 points, if the student got one or two items correct the student earned one point, and if the student got three or more of the items correct, two points were earned.

CR: Mathematics Completion

The Constructed-Response (CR) items, in selected grades for mathematics, required students to develop an answer instead of selecting an answer from response options. CR items were presented as novel tasks using materials and content presented in an on-demand test format. Each item was presented to the student in a standardized, scripted sequence of steps culminating in the TA scoring the student performance using the Mathematics Scoring Rubrics provided for the item. The Mathematics Scoring Rubrics provided scoring standards that were used to evaluate student responses. Test Administrators received training in the administration and scoring of CR mathematics items in online training module 9 (Mathematics DTA – Administer the Test). Directions and materials needed for administering mathematics CR items were included in the secure DTA that accompanied each mathematics test form. The TA entered the student CR score into the NCSC Assessment System as either correct or incorrect. All item responses were exported from the system and provided to the test administration vendor's DRS. DRS then applied the scoring rules. Each CR item contributed a score of 1 or 0 to the mathematics raw score.

Open-Response: Foundational Reading

Open-Response (OR) Foundational Reading items were included in the Reading Test in grades 3 and 4 only. The items were word identification tasks. Students identified either three or five words depending on the tier level of the items presented. Test Administrators received training in the administration and scoring of OR foundational reading items in online training module 10 (ELA DTA: Reading – Administer the Test). Directions and materials needed for administering reading OR items were included in the secure DTA that accompanied each ELA grades 3 and 4 test forms. The TA entered the student's scores into the NCSC Assessment System.

Students with clear and consistent oral speech were administered the OR Foundational Reading items. Students using communication other than oral speech, such as Augmentative and Alternative Communication (AAC) devices, American Sign Language, braille or eye gaze were administered the SR Foundational Reading items included in the Reading Test.

All item responses were exported from the system and provided to the test administration vendor's DRS. DRS then applied the scoring rules. Individual items were scored as correct or incorrect. For scoring purposes the Tier 1 foundational items required the student to identify all three words correctly in order to earn one score point for the foundational item set. The Tiers 2, 3 and 4 foundational items required the student to identify 4 or 5 of the 5 words correct in order to earn one score point for the foundational item set.

Overview of scoring process within the assessment system

The NCSC Assessment System provided automated machine scoring for all item types, aside from extended/constructed response which required human scoring. The system also allowed for teacher entry of student responses to be used for paper-based test delivery. The NCSC Assessment System automatically

scored question types that were machine-scorable and where scoring data had been provided. At the completion of the operational test, all test data were extracted from the system and were then compiled to generate full result sets for each student's tests.

During the initial pilot and operational test year, it was discovered that several scenarios within the online system could lead to a potential loss of student responses. Through the collaboration of the technology vendor, with the approval of the NCSC states, these situations were eliminated. Additionally, the mechanism that allowed the test delivery platform to scale to multiple web servers was extended to support a more fault-tolerant environment. This approach allowed test delivery results to persist even after the test results had been exported from the system.

In cases where test results were questioned, it is this repository of results that allowed system administration teams to reconstruct test results regardless of downstream publishing and scoring activities. This facility was used during the operational test period to recover results that would have otherwise been lost. The technology and test administration vendors' response teams worked together closely to identify any discrepancies in the students' results and ensured that any data discrepancies were quickly identified and remediated.

Administrator/scorer training

All TAs were required to participate in administration training modules and pass a final quiz in order to be certified to administer the NCSC assessment. The training included modules for each of the content area (Mathematics, Reading and Writing) DTAs. The modules reviewed the parameters for the administration and scoring of each item type, as well as how to enter the student responses into the NCSC Assessment System (Mathematics and Reading only).

During test administration, TAs used the content area DTAs to administer each item. The DTAs included the teacher scripting and directions related to any item set up, providing directions for the teacher to follow during administration. For the mathematics constructed response items, the DTA included any templates required by the items, the directions related to how to present the items to the student, and the rubrics used to score the items.

Further direction was provided to TAs on the entering of item responses in NCSC Assessment System through the NCSC Assessment System User Guide for Test Administrators. The guide outlined the use of the system, including how to enter student responses and submit each content area test.

During the administration window TAs were able to call or email the NCSC Service Center with any questions that they might have related to the administration of test items and submission of the student responses within the NCSC Assessment System.

Scoring Processes and Rules for Field Test Writing Constructed Response Items

Overview of scoring process by item type

CR - Writing

The writing CR items were field tested and required students to produce a permanent product in response to a writing prompt. The student, or a qualified scribe, recorded the response to the writing prompt on either the response template that was in the NCSC Assessment System or on the paper response template that was included in the writing DTA. The NCSC grades 3-8 and grade 11 writing assessments included CR items to which a student generated a response using a response template, which included pre-populated sentence starters.

The writing CR item was presented to the student by the TA in a standardized, scripted sequence of steps and included directions to present grade- and prompt-specific writing stimulus materials that needed to be printed and prepared. All writing stimulus materials, including the response template, were identified by a card number and were included in the Writing DTA. If the student used a paper version of the template to write a response, the following parameters were followed:

- TAs were instructed to annotate or interpret the student's writing directly on the student's written product if the TA believed that a novel reader (i.e., a scorer who might not be able to interpret a component of the student's written product, such as inventive spelling, penmanship, or use of symbolic expressions.
- TAs were instructed to transcribe or type exactly the student's written response, including any annotations, into the NCSC Assessment System.

Field Tested Writing CR – Administrator training; monitoring

All Test Administrators were required to participate in administration training modules and pass a final quiz in order to be certified to administer the NCSC assessment. The training included module 11 (ELA DTA: Writing – Administer the Test) which reviewed the parameters for the administration of the CR item, as well as how to enter student responses into the NCSC Assessment System.

During the test administration, TAs used the DTAs to administer each writing CR item. The DTAs included the teacher scripting and directions related to any item set up, providing directions for the teacher to follow during administration.

Further direction was provided to TAs about entering item responses in the NCSC Assessment System through the NCSC Assessment System User Guide for Test Administrators. The guide outlined the use of the system, including how to enter student responses and submit the writing session of the test. TAs were able to upload the student's final writing response template directly in the system, retype the student response within the item response field of the item, or upload the template and retype it within the item response field of the item. The item responses (no matter how they were entered) were then extracted from the online system and provided to the test administration vendor for human scoring.

Field Tested Writing CR: Scoring

Range finding and exemplar identification

In preparation for writing constructed response scoring, the test administration vendor scoring team carefully reviewed the range finding sets created during Pilot 2 to ensure clear understanding of all training materials. Several conference calls were held between the test administration vendor and NCSC representatives prior to the start of scoring. These discussions resolved training papers that contained unclear annotations or needed additional clarification of the scoring rationale applied during Pilot 2 range finding. Adjustments were made to scores and annotations where appropriate and all changes were approved by NCSC representatives. Additionally, there were occasions where it was determined that a selected training paper from a Pilot 2 range finding set was not an exemplar and should not be utilized for scorer training. In these instances, the test administration vendor and NCSC representatives identified a replacement chosen from additional student responses in Pilot 2 range finding sets. All replacements made to the training papers were approved by NCSC representatives.

Recruiting and training scorers

Scorer Recruitment and Qualifications

The NCSC scorers were a diverse group of individuals with a broad range of backgrounds, including teachers, business professionals, graduate students, and retired educators. They were primarily obtained through Kelly Services, a temporary employment agency. All scorers assigned to work on the NCSC assessment held a 4-year college degree, which included ELA or writing coursework. Approximately 80% of the leadership and scorer group assigned to NCSC had previous experience in scoring alternate assessments. All scorers signed a nondisclosure/confidentiality agreement.

Table 5-1 summarizes the qualifications of the 2015 NCSC scoring leadership and readers.

Table 5-1. NCSC 2015: Qualifications of Scoring Leadership and Scorers

Scoring Responsibility	Educational Credentials			
	Doctorate	Master's	Bachelor's	Total
Scoring Leadership[1]		3 (30%)	7 (70%)	10
Scorers	1 (3%)	14 (42%)	18 (55%)	33

[1] Scoring Leadership=Scoring Supervisors and Scoring Team Leaders

Test Administration Vendor Staff and Scoring Leadership

The NCSC operational items for writing constructed-response items were scored in Dover, New Hampshire between June 24 and July 9, 2015. The following staff members were involved in scoring:

- Assistant Director (AD), Scoring Operations: Primarily responsible for coordinating scheduling, budgeting and logistics of all Scoring Centers. In addition, the AD for Scoring Operations had overall responsibility for NCSC scoring related activities.
- ELA Group Manager for Scoring: Responsible for managing scoring related activities, monitoring reports, and leadership and scorer training to ensure overall consistency of scoring.

- Scoring Content Specialist: Responsible for overseeing scoring activities across grades, and monitoring accuracy and productivity across groups.
- Special Education Specialist: Responsible for overseeing scoring activities and acting as special education lead in coordination with the test administration vendor scoring staff.
- *iScore* Operations Manager: Set up and maintained *iScore* system for scoring and coordinated technical communication.
- Scoring Supervisor: Responsible for selecting calibration responses, training Scoring Team Leaders and scorers, resolving arbitrations and monitoring the consistency of scoring for items in assigned grades.
- Scoring Team Leader (STL): Responsible for performing quality control measures, resolving arbitrations and monitoring the accuracy of a small group, usually consisting of not more than 6 scorers.

Training

Scoring Supervisors worked closely with NCSC representatives prior to the beginning of scoring to ensure clear understanding of all training materials. Scoring Supervisors trained the STLs who were required to meet the minimum accuracy standard of 80% exact and 90% exact plus adjacent agreement on all items. This process was applied to each trait individually across Qualification Sets 1 and 2. The STLs were also present during scorer training which further reinforced their understanding of the rubric and training materials. NCSC representatives were present and available during all trainings.

The test administration vendor conducted training on each CR item before scorers were allowed access to student responses. Scorers were divided into two groups. One group focused on CR items in grades 3-5 and the other group focused on CR items in grades 6-8 and 11. Training sessions for scorers were facilitated by a Scoring Supervisor and conducted in the following manner:

- Scorer training for NCSC commenced with an introduction to Scoring and an overview to explain the purpose and goal of the testing program and any unique features of the test and/or testing population.
- A general discussion addressed the security, confidentiality, and proprietary nature of testing, scoring materials, and procedures.
- Training for each item focused on the three traits of the NCSC analytic rubrics for writing and how the scoring for each trait would be applied to student work.
- The training for each item included pertinent information on the testing instructions and item stimuli.
- Scorers reviewed actual responses with an item-specific Anchor Set, averaging 10 responses representing a range of scores across traits.
- Scorers were instructed to refer back to the Anchor Set frequently during scoring.
- Anchor exemplars were presented in a pre-determined order and consisted of responses that were typical, rather than unusual or uncommon, solid, rather than controversial or borderline, and true.
- The Scoring Supervisor announced the anchor response score and explained the scoring rationale, allowing scorers to internalize typical characteristics of each score point.
- Supplementary training materials, averaging 7 responses, contained practice responses representing all score points across traits, when possible, and often contained responses that were more unusual and/or less solid (e.g., are shorter than normal, employ atypical approaches, contain

both very low and very high attributes, etc.). None of the practice papers contained non-scorable codes.

- During the review of practice responses, the Scoring Supervisor often focused review efforts on the lines between adjacent score points or clarification of other scoring issues that are traditionally difficult for scorers to internalize.
- Scorers independently read and scored each practice response and the Scoring Supervisor discussed the actual score, and explained the scoring rationale for each response.
- A Q&A segment addressed any remaining questions from scorers and provided clarification.

Qualification

Following the training for each item, scorers were required to complete a Qualification Set to determine eligibility to score student work. There were two Qualification Sets consisting of 10 responses each. These responses were pre-selected from Pilot Test 2 materials and approved by NCSC representatives. The responses, which represented a range of score points, were randomly distributed to scorers through *iScore.*

Scorers had two opportunities to qualify for scoring each item. If scorers attained a score match of at least 80%/exact/90% adjacent agreement on the first Qualification Set, they were admitted to score live student responses. If scorers were unable to attain a score match of at least 80% exact/90% adjacent agreement (80%/90%) on the first Qualifying Set, the Scoring Supervisor conducted a retraining by discussing the qualifying responses in terms of the scorepoint descriptions and the original Anchor Set. Following this training, scorers were assigned Qualification Set 2. If a scorer passed the first two traits on Qualification Set 1, but missed the third trait, they would be required to pass the third trait on Qualification Set 2. When a scorer achieved an accuracy rate on any traits they failed in Qualification Set 1, they were allowed to score student responses. When the disaggregated data indicated that a qualified scorer had demonstrated a weakness in a particular trait, that qualified scorer received additional training prior to start of scoring. NCSC representatives on site were satisfied that all scorers were well qualified to score student work.

Scorers who failed to achieve the minimum levels of agreement were not allowed to score live student responses. When scorers demonstrated a level of understanding and the ability to apply feedback, Scoring leadership and NCSC representatives chose to include the scorer in future trainings. During the scoring of all NCSC grades, there were 9 scorers who failed to qualify to score one of the 14 CR items.

Once the first CR item was completely scored, the training process was repeated for the next CR item. This continued until all CR items were scored. Table 5-2 summarizes the qualification rates for NCSC.

Table 5-2. Qualification Summary

Note: *iScore* designated each item per grade as WRCC001 and WRCC002 for identification purposes

Grade 3	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**11**	**2**	**13**	**13**	**0**	**13**

Total Failed	**4**	**2**		**0**	**0**	
Grade 4	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**12**	**1**	**13**	**2**	**11**	**13**
Total Failed	**1**	**0**		**11**	**0**	
Grade 5	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**5**	**13**	**18**	**10**	**11**	**21**
Total Failed	**20**	**7**		**11**	**0**	
GRADE 6	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**17**	**0**	**17**	**15**	**0**	**15**
Total Failed	**0**	**0**		**2**	**0**	
Grade 7	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**12**	**4**	**16**	**18**	**0**	**18**
Total Failed	**4**	**0**		**0**	**0**	
Grade 8	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**11**	**7**	**18**	**14**	**7**	**21**
Total Failed	**7**	**0**		**7**	**0**	
Grade 11	W R C C 0 0 1 Qual 1	W R C C 0 0 1 Qual 2	Scorers Qualified	W R C C 0 0 2 Qual 1	W R C C 0 0 2 Qual 2	Scorers Qualified
Total Passed	**17**	**10**	**27**	**21**	**2**	**23**
Total Failed	**10**	**0**		**3**	**0**	

Methodology for Scoring Field Tested Writing CR Items

All student responses were scored from either uploaded evidence or computer generated text, defined as student work directly entered into the NCSC Assessment System. TAs were directed to capture an image of the Student Response Template pages when the student did not enter a response directly into the system. The template contained the student's original product and the TA uploaded the captured image into the NCSC

Assessment System. The system allowed a TA to either use a webcam to take a snapshot of the student's paper, or scan it with the school's scanner, and upload to the system.

When both uploaded and computer generated text were available, the uploaded evidence was scored first and the computer generated text was used for clarification and confirmation of the uploaded student writing evidence. When there was only uploaded writing evidence but no computer generated text to provide clarification and confirmation, then the uploaded writing evidence was scored. When there was only computer generated text but no uploaded writing evidence, the computer generated text was scored.

The following processes were in place during the scoring of the NCSC writing CR items:

- The iScore system forced scorers to review all available pages before allowing a score to be submitted.
- All scoring was "blind," i.e., no student names were visible to scorers, only booklet numbers within iScore linked student responses.
- The test administration vendor maintained security during scoring by using a highly secure; server-to-server interface to ensure that access to all student response images was limited to only those who were scoring or working for the test administration vendor in a scoring management capacity.
- During scoring, iScore, enabled a constant measuring and monitoring of scorers for scoring accuracy and consistency. Each scorer's reading rate and total number of scored responses was also monitored.
- Scorers were required to maintain an acceptable scoring accuracy rate (80% exact/90% adjacent agreement) on a daily basis as measured across read-behinds, double-blinds, and daily calibration sets.
- Scorers who repeatedly fell below standard were retrained and/or dismissed from scoring that item after consultation with NCSC representatives (3 scorers were dismissed due to low performance over the course of scoring).
- Scoring rules were in place to determine the final score of record, or when a final score was to be provided by scoring leadership.

Table 5-3 represents the total number of student responses scored by item in each grade.

Table 5-3. Student Responses Per Grade

Grade	Item	Number of Student Responses
3	WRCC001	2058
	WRCC002	1924
Total		3982
4	WRCC001	2119
	WRCC002	2071
Total		4190
5	WRCC001	2165
	WRCC002	2111
Total		4276

6	WRCC001 WRCC002	2124 2197
Total		4321
7	WRCC001 WRCC002	2105 2211
Total		4316
8	WRCC001 WRCC002	2353 2156
Total		4509
11	WRCC001 WRCC002	2123 1916
Total		4039

Scoring Rules

All writing CR items were scored against a three-trait analytic rubric (see Appendix 5-A). The scoring scale options of 0, 1, 2, and 3 were applied to each trait. When a response did not conform to score point parameters, scorers could designate the response as one of the following:

- Blank: There is no attempt to respond to the item; no uploaded material is provided and no response has been typed.
- Unreadable: The text on the scorer's computer screen is indecipherable or too faint to read accurately.
- Non-English: The response is written in a language other than English.
- Repeats the Prompt: The response is a direct copy of the prompt without any original text.
- Escalate: The response requires clarification or adjudication by Scoring Leadership.
- No Score: The response requires designation by Scoring Leadership.

Table 5-4 shows the scoring resolution process for each of these designations.

Table 5-4. NCSC 2015: Scoring resolution process

Designation	Resolution Process
Blank	Responses scored Blank were sent to another scorer for a second read. Responses scored Blank twice were converted to zeros ('0's) for reporting purposes. Any discrepancies were resolved by the Scoring Leadership.
Unreadable	Those responses judged unreadable were forwarded to special queue within *iScore* to be reviewed by a Scoring Supervisor who resolved the student score. (If the response remained unreadable after review, the Scoring Supervisor assigned a score of "0")
Non-English	Responses written in a language other than English were marked Non-English and converted to zeros ("0") for reporting purposes with the exception of Non-English responses from the state of New Mexico. A post-scoring edit review of those responses

	identified New Mexico students and if a translation of the foreign language was supplied, that translation was scored.
Escalation	Responses that required additional clarification or adjudication were escalated to Scoring Leadership. This included responses where it appeared that more than one students' work had been uploaded to the response. Responses where the uploaded evidence was a mismatch to the typed response were escalated. These responses were forwarded to Scoring Leadership for response appraisal and scoring. Responses that were a legitimate response to another item were escalated for review by Scoring Leadership. Any student response indicating potential cheating and/or security lapses before, during, or after the test administration were scored based on its merits and then forwarded for review. If further attention was warranted, the Client Services team notified the State Test Coordinator and NCSC partner states.
No Score	Responses that were determined to be non-scorable were resolved by the test administration vendor leadership team in partnership with NCSC representatives.

Scorers also had the option of flagging a response as an "Alert" paper requiring immediate review and possible immediate action by scoring leadership and NCSC representatives. Only one student response was deemed "Alert" and it was forwarded to the appropriate partner state.

"Alert" responses could include but were not limited to one or more of the following:

- Thoughts of suicide
- Criminal activity
- Alcohol or drug use
- Extreme depression
- Violence
- Rape, sexual or physical abuse
- Self-harm or intent to harm others
- Neglect

Monitoring of Scoring Quality Control

Test administration vendor scorers assigned to the NCSC project that met or exceeded the minimum standard for qualification were allowed access to score live student responses. Scorers were continuously monitored to ensure that scoring was accurate and consistent. Read-behind and double-blind statistics were reviewed daily. Calibration sets were administered and reviewed repeatedly during the course of the project. Scoring leadership and NCSC representatives on site paid close attention to the disaggregated read-behind, double- blind, and calibration statistics. Scorers in need of additional clarification on applying scores to specific traits were coached by scoring leadership. This continuous training allowed scorers an opportunity to resolve issues, ask questions, reiterate scoring guidelines, and establish parameters for atypical student responses. Scorers who demonstrated inaccurate or inconsistent scoring were retrained, and allowed to resume

scoring under increased supervision. Scoring leadership, after consultation with NCSC representatives, removed scorers who continued to fall below accuracy standards. Their work for the day was voided and rescored by other qualified scorers. During the scoring of NCSC, the work of 4 scorers was voided and rescored. There were two voids in grade 7 and one each in grades 3 and 4.

Throughout the scoring of the field tested CR writing items for NCSC, read-behind scoring, double-blind scoring and calibration sets were used as quality control measures. NCSC representatives along with the test administration vendor's Scoring team, monitored reports daily.

Calibration Sets

Calibration Sets commenced on the second day and any subsequent day of scoring each item. Scoring leadership selected the responses used for daily calibration and NCSC representatives on site approved each set. For calibration, scorers were assigned 5 pre-scored responses representing a variety of scores and selected from recently scored responses. Scorers who correctly assigned at least 12 out of 15 exact scores on the Calibration Set began scoring for the day. Scorers who failed to meet the standard were retrained by discussing the calibration responses in terms of the rubric and the Anchor Set. Scoring leadership determined if scorers who were retrained after the Calibration Set should be given access to live student responses to begin scoring. Retrained scorers who did return to scoring student work continued to be closely monitored. Over the course of NCSC scoring, there were an average of 22 scorers per grade. Only 11 scorers, across all 7 grades and 14 items, required retraining. In most cases, scorers who received retraining successfully returned to scoring and, as mentioned previously, only 4 scorers had work voided during the course of scoring.

Read-behind Scoring

Read-behinds provide a crucial tool in verifying scorer accuracy. The STL completed read-behinds on individual scorers throughout each shift. The STL's evaluation of each response was performed with no knowledge of the scores assigned across traits. The scores were only available to the STLs after they also scored the response. If there was a difference in scores, either adjacent or discrepant, the STL score became the score of record. If the scores were discrepant, or if there were a significant number of adjacent scores between the scorer and the STL, the STL discussed the rationale with the scorer.

The average number of read-behinds for each scorer was 5–10 reads a day, but this varied depending on the accuracy of each scorer. Read-behinds provided an immediate means of identifying scorers in need of further clarification on how to effectively apply the scoring rubrics to student responses. If scorers fell consistently below the 80% /90% exact/adjacent threshold, scoring leadership had the prerogative to void their scores for the day and/or stop them from scoring the item. Read-behinds were also performed on the STLs. Scoring leadership and NCSC representatives monitored scoring accuracy and consistency by reviewing the read-behinds performed by the STLs as well as reading behind them whenever possible.

Double-blind Scoring

While read-behinds measure scorer accuracy in relationship to STL scores, double-blind scoring provides statistics on scorer-to-scorer agreement. Double-blind scoring is the practice of having two scorers independently score a response, without knowing either the identity of the other scorer or the score that was assigned. In double-blind scoring neither scorer knows which response will be (or already has been) scored by another randomly selected scorer. All responses for NCSC were 100% double-blind scored.

In addition to monitoring inter-rater agreement rates, double-blind scoring also allowed scoring leadership to resolve arbitrations when two scorers' double-blind scores did not agree across any of the three traits. If there was not exact agreement, *iScore* automatically placed the response into an arbitration queue. Scoring leadership, with no prior knowledge of the scores assigned, evaluated the response and the score of the Scoring leadership became the score of record. The double-blind statistics provided an overview of agreement rate among the entire pool of scorers and assisted in identifying any need for group retraining.

Final Score Resolution

Scoring leadership provides resolution scores for responses that do not have exact agreement on all traits after read-behind or double-blind scoring. Less than 10% of the scores required this level of resolution. Tables 5-5, 5-6, and 5-7 provide examples of how the final score of record may be determined through resolutions.

Table 5-5. 2015 NCSC Operational: Examples of Scoring Resolutions: Read-Behind Scoring

Read-Behind Scoring [1] (Trait 1-Trait 2-Trait 3)		
Scorer Score	Leadership Score	Final
4-4-4	4-4-4	4-4-4
4-3-3	3-3-3	3-3-3
4-3-3	2-2-2	2-2-2
[1]In these cases, the Leadership score was the final score of record.		

Table 5-6. 2015 NCSC Operational: Examples of Scoring Resolutions: Double-Blind Scoring

Double-Blind Scoring[1] (Trait 1-Trait 2-Trait 3)			
Scorer #1	Scorer #2	Leadership Resolution	Final
4-4-4	4-4-3	4-4-4	4-4-4
4-3-3	2-2-2	3-3-2	3-3-2
2-1-1	1-1-1	2-2-1	2-2-1
2-2-2	4-4-4	3-3-3	3-3-3
[1]All adjacent or discrepant scores were resolved in arbitration and in these cases the leadership score became the final score of record.			

Table 5-7. 2015 NCSC Operational: Examples of Scoring Resolutions: Edit Scoring

Edit Scoring[1] (Trait 1-Trait 2-Trait 3)					
Scorer #1	Scorer #2	STL #1 RB	STL #2 RB	Scoring Supervisor Resolution	Final
3-2-2	3-2-2	-	-	-	3-2-2

2-2-2	3-2-2	2-2-2	2-2-2	-	2-2-2
0-1-1	1-2-1	1-2-1	1-2-1	-	1-2-1
3-2-2	2-1-1	3-2-2	3-1-2	3-2-2	3-2-2
1-0-1	1-1-2	1-1-1	1-1-2	1-1-2	1-1-2

[1] If a response received more than one read-behind and the scores supplied by the STLs did not agree, a resolution score was needed. In these cases the Scoring Supervisor provided a final score during the post scoring edit process.

Quality and Production Management Reports

Reports generated through *iScore* were essential during the NCSC scoring. Reports provided real time statistics to be reviewed by the test administration vendor's Scoring team and NCSC representatives in order to closely monitor scoring and ensure:

- Overall accuracy, consistency, and reliability of scoring (group level) was maintained.
- Scorer data (individual level) was monitored in real-time to allow early scorer intervention when necessary.
- Individual traits in need of further clarification were identified.
- Scoring schedules were upheld.

Several reports provided the comprehensive tools and statistical information needed to execute quality control and manage production. Thse reports are described in Table 5-8.

Table 5-8. NCSC 2015: Scoring quality control and production management

Report	Description
Read-behind Disaggregated Summary	The Read-behind Disaggregated Summary report showed the total number of Read-behind responses read by both the scorer and the STL and noted the number and percentage of exact, adjacent, and discrepant scores across each trait.
Double-blind Disaggregated Summary	The Double-blind Disaggregated Summary report showed the total number of Double-blind responses read by a scorer and noted the number and percentage of exact, adjacent, and discrepant scores across each trait.
Compilation Report	The Compilation Report showed for each scorer, the total number of responses scored, the number of calibration responses scored, and the percentage of exact, adjacent, and discrepant scores across each trait.
Summary Report	The Summary Report listed the total number of student responses loaded into *iScore*. This report included how many reads had been completed to date and how many reads remained

Inter-rater reliability

Kappa statistics (kappa coefficient) measure the agreement between two or more raters. The calculation is based on the difference between how much agreement is actually present compared to how much agreement would be expected to be present by chance alone. Kappa is a measure of this difference standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate disagreement. The Kappa information in Table 5-9 shows that agreement between raters fell into the substantial agreement or almost perfect agreement range for the field tested CR Writing items across grades.

Table 5-9. Kappa Agreement – Field Tested CR Writing Items NCSC 2015

Agreement Translation:
< 0 Less than chance agreement
0.01–0.20 Slight agreement
0.21–0.40 Fair agreement
0.41–0.60 Moderate agreement
0.61–0.80 Substantial agreement
0.81–0.99 Almost perfect agreement

Table 5.9.	Organization Trait 1	Idea Development Trait 2	Conventions Trait 3
Grade 3			
WRCC001	.62	.68	.82
WRCC002	.68	.66	.84
Grade 4			
WRCC001	.64	.72	.85
WRCC002	.62	.69	.84
Grade 5			
WRCC001	.70	.64	.83
WRCC002	.66	.64	.83
Grade 6			
WRCC001	.77	.69	.89
WRCC001	.70	.67	.87
Grade 7			
WRCC001	.78	.63	.85
WRCC002	.79	.65	.83
Grade 8			
WRCC001	.72	.72	.77
WRCC002	.74	.73	.78
Grade 11			
WRCC001	.76	.66	.72
WRCC002	.81	.64	.76

CHAPTER 6: PSYCHOMETRIC ANALYSES

This chapter provides an overview of the psychometric analyses of the operational test data. The first section presents classical statistical item analyses, providing results that are directly derived from the raw data. The second section presents analyses based on the application of item response theory (IRT) modeling techniques. Because this is the first year of the NCSC testing program, the IRT section provides an extensive review of the IRT modeling procedures, the process used to establish the NCSC scale, and the methods used for linking and equating the multiple forms used in a particular grade for a specific content area.[8]

CLASSICAL ITEM ANALYSES

Both qualitative and quantitative analyses were conducted on all the operational items for NCSC's alternate assessment based on alternate achievement standards (AA-AAS) to provide a thorough description of the quality of the student tasks that comprise the NCSC assessment instruments. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students, in particular racial, ethnic, or gender groups. This section focuses on quantitative evaluations from the perspective of classical item statistics. Further item-based statistical analyses can be found in Chapter 8 (devoted to statistical validity analyses) in the section describing differential item functioning (DIF) analyses. The individual item statistics can be found in Appendix 6-A.

The item-based classical statistical evaluations presented in this section are divided into two parts: (1) item difficulty, and (2) item-test correlations. The item analyses presented here are based on the operational administration of the NCSC AA-AAS in spring 2015. Note that the information presented in this chapter is based on the items on which students' reported scores are calculated (i.e., core items). The detailed process that was used to select core items is provided in Appendix 6-B, *NCSC Core Item List Report*.

Classical Difficulty & Discrimination Indices

All items were evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. Multiple-choice and one-point short-answer items are scored dichotomously (correct versus incorrect); for these items, the difficulty index is simply the proportion of students who correctly answered the item. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student abilities, but may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance of 0.25 (for

[8] Although not presented in the body of this chapter, analyses of student demographics and accommodation frequencies were also produced and are provided in Appendices 6-K and 6-L, respectively.

four-option multiple-choice items or essentially 0 for constructed-response items) to 0.90, with the majority of items generally falling between approximately 0.4 and 0.7 for the lower mathematics grades and between about 0.6 and 0.8 for ELA and the higher mathematics grades. However, on a standards-referenced assessment, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students do. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the item-test correlation is referred to as the item's discrimination because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For constructed-response items, the item discrimination index used was the Pearson product-moment correlation; for multiple-choice items, the corresponding statistic is commonly referred to as a point-biserial correlation. The theoretical range of these statistics is –1.0 to 1.0, with a typical observed range from 0.2 to 0.7.

A summary of the item difficulty and item discrimination statistics for each content area and grade is presented in Table 6-1. Note that the statistics are presented for all items as well as by item type (multiple-choice and constructed-response). The mean difficulty and discrimination values shown in the table are within typically observed ranges. One item each in grades 3 and 5 ELA tests displayed slightly negative discrimination statistics, but were not significantly different from 0. Discrimination values near 0 indicate that getting the item correct or incorrect is not indicative of high or low performance on the test as a whole. These items were included in the operational test forms to ensure content representativeness.

Table 6-1. 2014–15 NCSC: Summary of Item Difficulty and Discrimination Statistics by Content Area and Grade[1]

Content Area	*Grade*	*Number of Items*	p-*value*				*Discrimination*			
			Min	*Max*	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*
ELA	3	78	0.25	0.89	0.58	0.16	-0.02	0.61	0.40	0.14
	4	82	0.18	0.84	0.57	0.14	0.01	0.61	0.38	0.16
	5	61	0.31	0.85	0.58	0.14	-0.06	0.48	0.32	0.12
	6	74	0.33	0.91	0.63	0.16	0.03	0.53	0.34	0.12
	7	80	0.24	0.87	0.61	0.14	0.08	0.52	0.34	0.12
	8	85	0.32	0.88	0.61	0.14	0.00	0.52	0.33	0.11
	11	55	0.29	0.90	0.64	0.15	0.05	0.55	0.37	0.13
Mathematics	3	80	0.28	0.76	0.52	0.12	0.06	0.60	0.33	0.12
	4	78	0.19	0.81	0.48	0.15	0.05	0.50	0.27	0.12
	5	80	0.25	0.78	0.47	0.14	0.00	0.49	0.24	0.11
	6	79	0.29	0.85	0.55	0.13	0.07	0.50	0.31	0.11
	7	80	0.33	0.88	0.53	0.13	0.14	0.45	0.31	0.08
	8	79	0.28	0.87	0.51	0.13	0.09	0.46	0.32	0.09
	11	80	0.33	0.80	0.49	0.12	0.11	0.45	0.30	0.09

[1]Note: The number of items does not equal the number of raw score points.

The individual item statistics can be found in Appendix 6-A. The numbers of items corresponding to each tier and form are displayed in Tables 6-2 (ELA) and 6-3 (mathematics). Because the items were administered across multiple forms, the item statistics are summarized by form for ELA (see Table 6-4) and

mathematics (see Table 6-5). Note that the classical statistics should be interpreted with caution because the items are primarily two- or three-option selected-response items. The average classical item statistics indicate that the forms are highly consistent in terms of average item difficulty and discrimination. Because the items were developed to correspond to different tiers, the item statistics have also been summarized by tier in Table 6-6. Although the Tier 1 items tend to be easier than items from the other tiers, the relative difference is much greater when comparing the Tier 1 items to the other tiers than it is among Tiers 2, 3 and 4.

Table 6-2. 2014–15 NCSC: Number of Items by Grade, Form, Tier—ELA2

Grade	*Form*	*Tier 01*	*Tier 02*	*Tier 03*	*Tier 04*
3	01	11	12	18	1
	02	12	12	12	1
	03	12	12	12	6
	04	12	22	1	7
4	01	11	11	18	1
	02	14	11	13	1
	03	9	11	15	6
	04	11	21	1	8
5	01	10	7	14	1
	02	10	7	14	1
	03	9	7	17	1
	04	10	12	1	8
6	01	12	7	13	1
	02	11	13	7	1
	03	12	7	13	1
	04	11	13	1	7
7	01	13	7	12	1
	02	11	13	11	1
	03	15	7	12	1
	04	11	12	1	9
8	01	15	7	12	1
	02	11	17	7	1
	03	11	7	16	1
	04	11	13	1	11
11	01	19	6	6	1
	02	15	6	11	1
	03	15	6	11	1
	04	15	11	0	9

[2] The number of items does not equal the number of

Table 6-3. 2014–15 NCSC: Number of Items by Grade, Form, Tier—Mathematics

Grade	*Form*	*Tier 01*	*Tier 02*	*Tier 03*	*Tier 04*
3	01	7	12	12	4
	02	7	13	11	4
	03	7	13	11	4
	04	7	13	11	4
4	01	7	11	12	3
	02	7	12	12	2
	03	7	12	11	3
	04	8	10	13	2
5	01	7	11	13	4
	02	8	12	11	4
	03	7	12	12	3
	04	7	11	12	3
6	01	7	12	13	3
	02	7	12	12	4
	03	6	12	12	4
	04	7	13	12	3
7	01	7	13	12	2
	02	7	14	12	2
	03	6	13	13	2
	04	6	13	13	2
8	01	7	12	12	4
	02	7	12	13	2
	03	7	12	13	3
	04	7	12	13	3
11	01	7	12	13	2
	02	7	13	13	2
	03	6	13	12	3
	04	6	13	13	2

Table 6-4. 2014–15 NCSC: Item-Level Classical Test Theory Statistics—Summary by Grade, Content, and Form— ELA3

Grade	*Form*	*Number of Items*	p-*value*				*Discrimination*			
			Min	*Max*	*Mean*	*SD*	*Min*	*Max*	*Mean*	*SD*
3	01	42	0.34	0.88	0.60	0.14	-0.02	0.57	0.40	0.14
	02	37	0.34	0.89	0.62	0.15	-0.02	0.58	0.39	0.14
	03	42	0.25	0.89	0.59	0.18	-0.02	0.57	0.39	0.13
	04	42	0.29	0.89	0.61	0.15	0.14	0.61	0.41	0.11
4	01	41	0.42	0.84	0.60	0.12	0.01	0.61	0.39	0.16
	02	39	0.43	0.84	0.62	0.13	0.02	0.58	0.35	0.15
	03	41	0.18	0.82	0.58	0.15	0.02	0.59	0.40	0.14
	04	41	0.40	0.84	0.59	0.12	0.08	0.58	0.41	0.14
5	01	32	0.31	0.85	0.60	0.16	-0.06	0.48	0.31	0.13
	02	32	0.35	0.85	0.60	0.15	-0.06	0.48	0.31	0.13
	03	34	0.35	0.84	0.60	0.14	-0.06	0.48	0.34	0.11
	04	31	0.35	0.85	0.59	0.15	-0.06	0.48	0.32	0.12

Grade	Form	Number of Items	p-value				Discrimination			
			Min	Max	Mean	SD	Min	Max	Mean	SD
6	01	33	0.34	0.91	0.63	0.12	0.03	0.53	0.35	0.11
	02	32	0.33	0.88	0.62	0.16	0.07	0.47	0.33	0.11
	03	33	0.34	0.86	0.63	0.13	0.03	0.51	0.35	0.11
	04	32	0.34	0.88	0.65	0.15	0.09	0.50	0.35	0.12
7	01	33	0.35	0.87	0.64	0.14	0.13	0.52	0.35	0.11
	02	36	0.24	0.86	0.61	0.15	0.08	0.52	0.34	0.12
	03	35	0.35	0.86	0.62	0.13	0.13	0.51	0.34	0.11
	04	33	0.35	0.86	0.61	0.14	0.12	0.48	0.32	0.11
8	01	35	0.36	0.88	0.65	0.14	0.13	0.48	0.33	0.09
	02	36	0.45	0.88	0.63	0.12	0.00	0.49	0.34	0.12
	03	35	0.44	0.88	0.62	0.14	0.13	0.44	0.32	0.09
	04	36	0.32	0.88	0.62	0.15	0.00	0.52	0.33	0.12
11	01	32	0.29	0.90	0.70	0.15	0.05	0.55	0.37	0.13
	02	33	0.29	0.90	0.67	0.15	0.05	0.55	0.35	0.14
	03	33	0.29	0.90	0.67	0.16	0.05	0.55	0.36	0.13
	04	35	0.31	0.90	0.67	0.14	0.17	0.55	0.38	0.08

[3]Note: The number of items does not equal the number of raw score points.
MC = multiple-choice; OR = open-response

Table 6-5. 2014–15 NCSC: Item-Level Classical Test Theory Statistics—Summary by Grade, Content, and Form—Mathematics

Grade	Form	Number of Items	p-value				Discrimination			
			Min	Max	Mean	SD	Min	Max	Mean	SD
3	01	35	0.31	0.76	0.53	0.13	0.06	0.57	0.32	0.11
	02	35	0.28	0.76	0.53	0.14	0.09	0.58	0.35	0.11
	03	35	0.30	0.76	0.53	0.12	0.15	0.57	0.33	0.11
	04	35	0.28	0.76	0.52	0.12	0.15	0.60	0.35	0.12
4	01	33	0.19	0.81	0.48	0.14	0.08	0.49	0.28	0.11
	02	33	0.23	0.81	0.48	0.16	0.06	0.46	0.25	0.12
	03	33	0.24	0.81	0.49	0.15	0.07	0.50	0.28	0.12
	04	33	0.24	0.81	0.49	0.16	0.05	0.45	0.24	0.11
5	01	35	0.25	0.78	0.48	0.14	0.00	0.49	0.23	0.11
	02	35	0.25	0.78	0.48	0.15	0.01	0.41	0.23	0.10
	03	34	0.25	0.78	0.46	0.14	0.01	0.45	0.24	0.11
	04	33	0.25	0.78	0.48	0.15	0.01	0.49	0.25	0.11
6	01	35	0.30	0.80	0.54	0.14	0.08	0.48	0.29	0.11
	02	35	0.30	0.85	0.54	0.13	0.07	0.47	0.29	0.09
	03	34	0.30	0.75	0.53	0.12	0.07	0.50	0.30	0.11
	04	35	0.29	0.78	0.54	0.13	0.08	0.48	0.30	0.10
7	01	34	0.33	0.88	0.53	0.14	0.14	0.44	0.29	0.07
	02	35	0.37	0.88	0.54	0.14	0.17	0.44	0.32	0.07
	03	34	0.34	0.88	0.53	0.14	0.20	0.45	0.31	0.07
	04	34	0.37	0.88	0.53	0.14	0.17	0.42	0.30	0.07
8	01	35	0.34	0.81	0.51	0.11	0.09	0.44	0.30	0.10
	02	34	0.34	0.76	0.52	0.10	0.09	0.46	0.32	0.10
	03	35	0.34	0.87	0.50	0.12	0.09	0.44	0.31	0.09
	04	35	0.28	0.76	0.49	0.12	0.09	0.41	0.29	0.09

Grade	Form	Number of Items	p-value				Discrimination			
			Min	Max	Mean	SD	Min	Max	Mean	SD
11	01	34	0.33	0.80	0.49	0.13	0.11	0.45	0.30	0.10
	02	35	0.33	0.77	0.49	0.12	0.11	0.45	0.29	0.08
	03	34	0.33	0.77	0.48	0.12	0.11	0.44	0.27	0.09
	04	34	0.33	0.77	0.48	0.11	0.11	0.44	0.30	0.08

MC = multiple-choice; OR = open-response

Table 6-6. 2014–15 NCSC: Item-Level Classical Test Theory Statistics—Summary by Grade, Content, and Tier

Content Area	Grade	Tier	Number of Items	p-value				Discrimination			
				Min	Max	Mean	SD	Min	Max	Mean	SD
ELA	3	1	16	0.59	0.89	0.77	0.10	0.18	0.51	0.38	0.08
		2	22	0.29	0.78	0.54	0.14	0.14	0.61	0.44	0.13
		3	28	0.29	0.77	0.55	0.13	-0.02	0.58	0.36	0.17
		4	12	0.25	0.68	0.49	0.16	0.19	0.55	0.44	0.10
	4	1	18	0.66	0.84	0.77	0.05	0.10	0.59	0.40	0.15
		2	21	0.41	0.69	0.53	0.07	0.15	0.58	0.43	0.14
		3	30	0.40	0.71	0.55	0.09	0.01	0.61	0.34	0.17
		4	13	0.18	0.64	0.43	0.15	0.08	0.57	0.39	0.15
	5	1	14	0.65	0.85	0.77	0.07	0.25	0.45	0.37	0.06
		2	12	0.35	0.72	0.53	0.11	-0.06	0.48	0.30	0.15
		3	27	0.31	0.74	0.52	0.09	0.00	0.48	0.31	0.13
		4	8	0.35	0.69	0.49	0.10	0.07	0.43	0.29	0.13
	6	1	23	0.60	0.91	0.78	0.09	0.15	0.53	0.40	0.08
		2	19	0.37	0.79	0.63	0.13	0.16	0.50	0.39	0.09
		3	25	0.33	0.73	0.54	0.11	0.03	0.50	0.29	0.12
		4	7	0.34	0.58	0.44	0.09	0.09	0.47	0.21	0.14
	7	1	25	0.47	0.87	0.76	0.10	0.17	0.50	0.39	0.09
		2	18	0.42	0.75	0.62	0.09	0.15	0.52	0.37	0.11
		3	28	0.24	0.71	0.52	0.10	0.08	0.52	0.32	0.12
		4	9	0.40	0.59	0.50	0.06	0.12	0.39	0.24	0.10
	8	1	23	0.50	0.88	0.76	0.11	0.00	0.48	0.34	0.11
		2	23	0.47	0.77	0.62	0.09	0.06	0.52	0.39	0.11
		3	28	0.36	0.70	0.53	0.09	0.14	0.48	0.30	0.09
		4	11	0.32	0.54	0.45	0.06	0.11	0.41	0.26	0.10
	11	1	19	0.59	0.90	0.79	0.08	0.19	0.55	0.39	0.08
		2	11	0.53	0.77	0.66	0.09	0.31	0.55	0.44	0.08
		3	16	0.29	0.72	0.54	0.14	0.05	0.55	0.30	0.18
		4	9	0.31	0.64	0.51	0.10	0.17	0.50	0.34	0.09
Mathematics	3	1	16	0.54	0.76	0.68	0.06	0.16	0.37	0.30	0.05
		2	30	0.37	0.72	0.51	0.09	0.19	0.59	0.38	0.12
		3	24	0.28	0.62	0.47	0.10	0.15	0.60	0.33	0.12
		4	10	0.28	0.59	0.41	0.09	0.06	0.58	0.26	0.16
	4	1	17	0.47	0.81	0.68	0.09	0.09	0.38	0.23	0.08
		2	27	0.23	0.67	0.43	0.12	0.08	0.48	0.30	0.13
		3	27	0.24	0.59	0.42	0.09	0.05	0.50	0.27	0.13
		4	7	0.19	0.45	0.37	0.09	0.10	0.40	0.19	0.10

Cont.

Content Area	Grade	Tier	Number of Items	p-value				Discrimination			
				Min	Max	Mean	SD	Min	Max	Mean	SD
	5	1	17	0.51	0.78	0.69	0.08	0.11	0.38	0.26	0.07
		2	28	0.33	0.55	0.46	0.05	0.08	0.49	0.26	0.11
		3	27	0.25	0.58	0.38	0.08	0.00	0.43	0.21	0.12
		4	8	0.25	0.52	0.38	0.10	0.07	0.49	0.20	0.13
	6	1	15	0.59	0.85	0.73	0.07	0.21	0.40	0.29	0.05
		2	28	0.31	0.76	0.56	0.10	0.13	0.48	0.36	0.09
		3	28	0.29	0.64	0.48	0.08	0.07	0.48	0.27	0.12
		4	8	0.30	0.65	0.44	0.12	0.08	0.50	0.30	0.15
	7	1	14	0.62	0.88	0.73	0.08	0.17	0.38	0.25	0.06
		2	32	0.37	0.72	0.52	0.10	0.21	0.45	0.35	0.07
		3	29	0.33	0.59	0.45	0.06	0.14	0.42	0.31	0.09
		4	5	0.37	0.53	0.46	0.06	0.17	0.36	0.30	0.08
	8	1	16	0.46	0.87	0.68	0.11	0.10	0.41	0.27	0.09
		2	27	0.38	0.70	0.50	0.08	0.17	0.43	0.33	0.07
		3	30	0.29	0.69	0.45	0.09	0.09	0.46	0.33	0.10
		4	6	0.28	0.48	0.41	0.08	0.14	0.44	0.33	0.10
	11	1	14	0.44	0.80	0.68	0.12	0.12	0.34	0.25	0.06
		2	30	0.33	0.64	0.47	0.07	0.14	0.45	0.32	0.09
		3	30	0.34	0.54	0.43	0.05	0.15	0.44	0.31	0.07
		4	6	0.34	0.49	0.39	0.06	0.11	0.35	0.22	0.12

IRT CALIBRATING, SCALING, & EQUATING PROCESS

This section describes the procedures used in the application of item response theory (IRT) modeling techniques to the NCSC tests, including descriptions of: the psychometric model, the calibration procedures, the reporting scale, and the equating procedures. Because this is the first year of the NCSC assessment program, these descriptions include detailed accounts of the decisions made on the psychometric model and the process used to establish the NCSC scale.

During the course of these psychometric analyses, a number of quality-control procedures and checks on the processes were implemented. These procedures included evaluations of the calibration processes (e.g., checking the number of Newton cycles required for convergence for reasonableness, checking item parameters and their standard errors for reasonableness, examination of Test Characteristic Curves [TCCs] and Test Information Functions [TIFs] for reasonableness); evaluation of model fit; evaluation of core items (e.g., comparability among the four forms administered within each grade/content combination) and evaluation of the scaling results (e.g., parallel processing by the Psychometrics and Research department and Data and Reporting Services department). More complete documentation can be found in the Appendix 6-B, NCSC Core Item List Report, submitted to the NCSC steering committee for approval to inform operational scaling, the choice of the final set of operational items for each form, and standard setting.

Table 6-7 lists items that required intervention either during item calibration and equating. For each flagged item, the table shows the reason it was flagged (e.g., a-parameter) and what action was taken. The number of items identified for evaluation was typical across grades and content areas.

Table 6-7. 2014–15 NCSC: Items That Required Intervention During IRT Calibration and Equating

Content Area	*Grade*	*Item ID*	*Reason*	*Action*
ELA	3	121596A	*a*-parameter	*a* set to initial
	4	121635A	*a*-parameter	*a* set to initial
	5	119972B	*a*-parameter	*a* set to initial
		117110B	*a*-parameter	*a* set to initial
		121720B	*a*-parameter	*a* set to initial
	6	123356A	*a*-parameter	*a* set to initial
		120013B	*a*-parameter	*a* set to initial
		121349B	*a*-parameter	*a* set to initial
	7	121502A	*a*-parameter	*a* set to initial
	8	121160A	*a*-parameter	*a* set to initial
	11	121065A	*a*-parameter	*a* set to initial
		121953B	*a*-parameter	*a* set to initial

Overview of Measurement Model

All NCSC items were calibrated using item response theory (IRT). IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, all items are assumed to be independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton & Swaminathan, 1985; Hambleton & van der Linden, 1997). The process of determining the specific mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p. Once the item parameters are known, an estimate of θ for each student can be calculated based on the student's observed responses to the items. This estimate, $\hat{\theta}$, is considered to be an estimate of the student's true score or a general representation of student performance. It has characteristics that may be preferable to those of raw scores for equating purposes.

For the 2014–15 NCSC tests, the two-parameter logistic (2PL) model was used for dichotomous (multiple-choice) items. The 2PL model for dichotomous items can be defined as:

$$P_i(X_i = 1|\theta_j) = \frac{\exp[Da_i(\theta_j - b_i)]}{1+\exp[Da_i(\theta_j - b_i)]}$$

where
X_i indexes the scored response on an item,
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
D is a normalizing constant equal to 1.701.

Since the 2014-15 NCSC AA-AAS was the first operational administration, the specific IRT models to be used for calibration had not been previously determined. Per recommendations of the NCSC Technical Advisory Committee (TAC), the test administration vendor's psychometricians fitted both Rasch and two-parameter logistic (2PL) models to the dichotomously scored items in the operational NCSC data (i.e., 14 grade/content combinations). Evaluation and comparison of the model-data fit facilitated the determination of the appropriate IRT model. A detailed comparison of the fit of these two models to the dichotomously scored items is given in Appendices 6-C and 6-D.

Calibration

After the modeling decision studies were completed as described above, the final decisions were implemented in the operational calibration of the 2014-15 NCSC assessment. PARSCALE 4.1 was used for all analyses. All command files were set up in a way that all general settings were identical across all grade/content combinations. The calibration statement for all the analyses reads: CAL GRADED, LOGISTIC, CYCLE=(100,1,1,1,1), TPRIOR, SPRIOR;

The logistic version of the item response theory (IRT) model was used, and default priors were used for all parameter estimates. Each item occupied its own unique block in the command file. For all grade/content area combinations, the largest change in parameter values (from one iteration to the next) was decreasing and tended to flatten out toward the end of the calibration process. The number of Newton cycles to convergence for each grade/content for the initial calibrations when 2PL (the model used to generate student scores) was fitted to the data is listed in Table 6-8. Note that the number of cycles observed for each test was far less than the prescribed maximum 100 in the calibration statement, thus indicating good convergence. For more information about item calibration and determination, refer to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

Table 6-8. Number of Cycles to Convergence

Content Area	*Grade 3*	*Grade 4*	*Grade 5*	*Grade 6*	*Grade 7*	*Grade 8*	*Grade 11*
ELA	38	43	41	43	53	38	55
Mathematics	28	20	17	27	28	29	19

ITEM RESPONSE THEORY RESULTS

The tables in Appendix 6-E give the IRT item parameters for all the core items on the 2014–15 NCSC tests by grade and content area. The statistics for the core items are summarized in Tables 6-9 through 6-13. The mean item parameter estimates shown in the tables below are within generally acceptable and expected ranges. For easy reference, Table 6-9 displays the means and standard deviations averaged over all forms for each grade and content area.

Table 6-9. 2014–15 NCSC: IRT Summary Statistics Overall

Content Area	*Grade*	*Number of Items*	*a*	*SD (a)*	*b*	*SD (b)*
ELA	3	78	0.81	0.37	-0.27	1.04
	4	82	0.78	0.42	-0.33	0.75
	5	61	0.66	0.33	-0.06	1.45
	6	74	0.78	0.42	-0.35	0.89
	7	80	0.79	0.50	-0.34	0.73
	8	85	0.70	0.37	-0.39	0.67
	11	55	0.89	0.48	-0.38	0.92
Mathematics	3	80	0.67	0.36	0.00	0.71
	4	78	0.57	0.34	0.33	0.95
	5	80	0.51	0.28	0.38	0.99
	6	79	0.66	0.29	-0.07	0.83
	7	80	0.63	0.22	-0.05	0.68
	8	79	0.62	0.22	0.03	0.71
	11	80	0.61	0.21	0.19	0.68

Because the items were administered across multiple forms, the IRT statistics are summarized by form for ELA (see Table 6-10) and mathematics (see Table 6-11).

Table 6-10. 2014–15 NCSC: IRT Summary Statistics by Grade and Form- ELA4

Grade	*Form*	*Number of Items*	*a*	*SD (a)*	*b*	*SD (b)*
3	01	42	0.81	0.37	-0.29	1.19
	02	37	0.78	0.35	-0.35	1.28
	03	42	0.78	0.33	-0.23	1.29
	04	42	0.83	0.34	-0.48	0.69
4	01	41	0.81	0.44	-0.43	0.62
	02	39	0.71	0.41	-0.62	0.74
	03	41	0.80	0.38	-0.35	0.72
	04	41	0.83	0.39	-0.40	0.64
5	01	32	0.67	0.36	-0.03	1.87
	02	32	0.67	0.37	0.01	1.91
	03	34	0.70	0.33	-0.08	1.75
	04	31	0.68	0.35	-0.04	1.83
6	01	33	0.72	0.34	-0.44	0.80
	02	32	0.76	0.42	-0.37	0.83
	03	33	0.75	0.38	-0.44	0.84
	04	32	0.81	0.43	-0.44	0.81
7	01	33	0.80	0.47	-0.49	0.61
	02	36	0.81	0.56	-0.27	0.89

Grade	Form	Number of Items	a	SD (a)	b	SD (b)
	03	35	0.72	0.40	-0.43	0.58
	04	33	0.76	0.57	-0.36	0.65
8	01	35	0.71	0.35	-0.57	0.65
	02	36	0.70	0.33	-0.58	0.54
	03	35	0.68	0.35	-0.47	0.63
	04	36	0.70	0.38	-0.44	0.75
11	01	32	1.02	0.52	-0.58	1.08
	02	33	0.87	0.40	-0.45	1.07
	03	33	0.87	0.39	-0.45	1.07
	04	35	0.89	0.33	-0.61	0.63

[4]Note: The number of items does not equal the number of raw score points.

Table 6-11. 2014–15 NCSC: IRT Summary Statistics by Grade, and Form-Mathematics

Grade	Form	Number of Items	a	SD (a)	b	SD (b)
3	01	35	0.62	0.32	-0.03	0.75
	02	35	0.68	0.34	-0.05	0.72
	03	35	0.65	0.30	-0.08	0.71
	04	35	0.69	0.37	-0.05	0.67
4	01	33	0.54	0.26	0.22	0.89
	02	33	0.54	0.33	0.33	1.03
	03	33	0.59	0.35	0.26	0.89
	04	33	0.51	0.30	0.31	1.04
5	01	35	0.50	0.30	0.35	1.09
	02	35	0.47	0.22	0.33	1.06
	03	34	0.48	0.26	0.39	1.01
	04	33	0.52	0.27	0.28	0.99
6	01	35	0.61	0.28	-0.01	0.92
	02	35	0.58	0.23	-0.07	0.83
	03	34	0.61	0.28	0.00	0.80
	04	35	0.59	0.24	-0.02	0.90
7	01	34	0.58	0.19	-0.08	0.78
	02	35	0.65	0.22	-0.13	0.69
	03	34	0.63	0.20	-0.03	0.70
	04	34	0.59	0.21	-0.06	0.71
8	01	35	0.56	0.21	-0.01	0.67
	02	34	0.61	0.25	-0.06	0.58
	03	35	0.57	0.20	0.06	0.63
	04	35	0.53	0.17	0.15	0.75
11	01	34	0.61	0.21	0.11	0.75
	02	35	0.58	0.20	0.16	0.71
	03	34	0.56	0.22	0.25	0.77
	04	34	0.58	0.18	0.25	0.66

Although the IRT statistics appear slightly more variable than the classical, they remain consistent with the classical statistics; the forms are relatively uniform in terms of average difficulty and discrimination, particularly in math. The difference between the content areas is somewhat expected because the design of the

assessments calls for ELA items to be administered in sets while the math design does not. The comparability of forms is further explored by comparing the TCCs of the various forms (see Appendix 6-F). Because the items were developed to correspond to tiers, the item statistics are also summarized by tier for ELA (see Table 6-12) and mathematics (see Table 6-13).

Table 6-12. 2014–15 NCSC: IRT Summary Statistics by Grade and Tier- ELA

Grade	*Tier*	*Number of Items*	*a*	*SD (a)*	*b*	*SD (b)*
3	1	16	0.87	0.26	-1.16	0.44
	2	22	0.91	0.39	-0.13	0.66
	3	28	0.70	0.42	0.01	1.37
	4	12	0.78	0.26	0.00	0.65
4	1	18	0.98	0.47	-1.25	0.57
	2	21	0.84	0.39	-0.14	0.36
	3	30	0.65	0.42	-0.22	0.53
	4	13	0.73	0.27	0.38	0.76
5	1	14	1.02	0.33	-1.06	0.31
	2	12	0.54	0.28	0.64	2.72
	3	27	0.57	0.25	0.08	0.90
	4	8	0.50	0.23	0.19	0.52
6	1	23	1.18	0.41	-1.04	0.28
	2	19	0.79	0.27	-0.42	0.59
	3	25	0.53	0.24	0.05	0.98
	4	7	0.39	0.23	0.74	0.81
7	1	25	1.19	0.59	-0.92	0.34
	2	18	0.77	0.36	-0.44	0.44
	3	28	0.57	0.29	0.10	0.81
	4	9	0.37	0.13	0.12	0.56
8	1	23	0.93	0.47	-1.03	0.41
	2	23	0.81	0.32	-0.52	0.37
	3	28	0.54	0.20	-0.07	0.54
	4	11	0.42	0.15	0.42	0.58
11	1	19	1.17	0.52	-1.08	0.27
	2	11	1.00	0.34	-0.56	0.28
	3	16	0.65	0.43	0.31	1.19
	4	9	0.62	0.19	0.13	0.63

Table 6-13. 2014–15 NCSC: IRT Summary Statistics by Grade and Tier- Mathematics

Grade	*Tier*	*Number of Items*	*a*	*SD (a)*	*b*	*SD (b)*
3	1	16	0.64	0.18	-0.88	0.32
	2	30	0.78	0.38	-0.02	0.40
	3	24	0.60	0.36	0.26	0.60
	4	10	0.54	0.43	0.82	0.71
4	1	17	0.56	0.24	-0.92	0.42
	2	27	0.66	0.38	0.53	0.73
	3	27	0.55	0.37	0.67	0.71
	4	7	0.33	0.14	1.24	0.78

Cont.

Grade	Tier	Number of Items	a	SD (a)	b	SD (b)
5	1	17	0.66	0.28	-0.89	0.37
	2	28	0.50	0.28	0.34	0.44
	3	27	0.46	0.25	1.03	0.91
	4	8	0.40	0.34	1.01	0.92
6	1	15	0.72	0.25	-1.05	0.28
	2	28	0.76	0.28	-0.14	0.57
	3	28	0.52	0.23	0.34	0.75
	4	8	0.65	0.46	0.62	0.88
7	1	14	0.69	0.27	-1.10	0.31
	2	32	0.70	0.21	-0.03	0.47
	3	29	0.56	0.17	0.38	0.47
	4	5	0.54	0.15	0.30	0.49
8	1	16	0.64	0.31	-0.86	0.50
	2	27	0.62	0.17	0.06	0.43
	3	30	0.61	0.21	0.35	0.57
	4	6	0.59	0.21	0.61	0.79
11	1	14	0.66	0.23	-0.81	0.60
	2	30	0.65	0.22	0.27	0.50
	3	30	0.58	0.16	0.41	0.32
	4	6	0.41	0.19	0.99	0.61

Item difficulty tends to have a positive relationship with tier; as the tier increases the items tend to be more difficult. Consistent with the classical stats, the Tier 1 items appear to be less similar from the other tiers in terms of magnitude of difficulty, and the Tiers 2 and 3 items occasionally overlap. This reversal of difficulty (between Tiers 2 and 3) tends to happen more frequently in ELA than in mathematics. Further investigation may be warranted in Grade 5 ELA where the Tier 2 items appear to be more difficult than anticipated.

The TCCs provide a more complete picture of the equivalence of the various forms. TCCs display the expected (average) raw score associated with each θ_j value between –4.0 and 4.0. Mathematically, the TCC is computed by summing the expected score on all the items or item sets (for the foundational item sets and the Tier 1 writing prompt selected-response item sets) that contribute to the raw score. The expected raw score at a given value of θ_j is

$$E(X|\theta_j) = \sum_{i=1}^{n} E(X_i|\theta_j)$$

where
X indexes total raw test score,
X_i indexes the scored response on an item,
i indexes the items (and n is the number of items contributing to the raw score),
j indexes students (here, θ_j runs from –4 to 4), and
$E(X|\theta_j)$ is the expected raw score on the test for a student of ability θ_j.

The expected raw score monotonically increases with θ_j, consistent with the notion that students of high ability tend to earn higher raw scores than do students of low ability. Most TCCs are "S-shaped"—flatter at the ends of the distribution and steeper in the middle.

The TIF, $I(\theta)$(see Lord, 1980, for theoretical definitions and examples of equations), displays the amount of statistical information the test provides at each value of θ_j. Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its Standard Error of Measurement (SEM). The SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, & Rogers, 1991), as follows:

$$\mathrm{SEM}(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the θ distribution where most students generally are located and where most items are sensitive by design. Appendix 6-F shows graphs of the TCCs and TIFs for each content area by grade and form. Also, Appendix 6-G displays the mathematical derivation of TCC and TIF equations for the foundational items and the Tier 1 writing prompt selected-response item sets.

Establishing the NCSC Scale

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for NCSC. The reporting scales are simple linear transformations of the underlying θ scale. The reporting scales were developed such that they range from 1200 through 1290 for all grade/content combinations. The second cut is fixed at 1240 for each grade level. In other words, to be classified in Level 3 or above, a minimum scale score of 1240 was required at all grades.

By providing information that is more specific about the position of a student's results, scale scores supplement performance level scores. Students' raw scores (i.e., total number of points) on the 2014–15 NCSC tests were translated to scale scores using a data analysis process called scaling. Scaling simply converts from one scale to another scale. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the 2014–15 NCSC tests can be expressed in raw or scale scores.

It is important to note that converting from raw scores to scale scores does not change students' performance level classifications. Given the relative simplicity of raw scores, it is fair to question why scale scores for NCSC are reported instead of raw scores. Scale scores make for more consistent reporting of results. The psychometric advantage of scale scores over raw scores comes from their being linear transformations of θ. Raw scores are not comparable from year to year because they are affected by differences in group ability and/or difficulty of the items that appear on each test form. Equating is a statistical procedure that is used to adjust for differences in form difficulty so that scores on alternate forms can be used interchangeably (Kolen & Brennan, 2014). Since the θ scale is used for equating, scale scores are comparable from one year to the next.

The scale scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scale score metric.

Students' ability estimates are based on their raw scores and are found by mapping through the test characteristic curve (TCC). Scale scores are calculated using the linear equation:

$$SS = m\hat{\theta} + b$$

where
m is the slope, and
b is the intercept.

NCSC requested various scales within the 1200's range for various standard deviations with the same cut score used (either 1230 or 1240) for all tests. After discussing the pros and cons of the various scales, the NCSC Steering Committee reviewed the possible scales provided to them by the test administration vendor and preferred a proficient cut of 1240 and standard deviation of 15 with a highest obtainable scale score (HOSS) of 1290 over other possible methods. This is the method that was adopted for operational scaling. A separate linear transformation is used for each grade and content area combination. As previously stated, the transformation function is determined by fixing the Level 2/Level 3 cut score and the standard deviation of the scale—that is, the cut score is set at 1240 and the scale score standard deviation is fixed at 15. A raw score of 0 is assigned a scale score of 1200, and the maximum possible raw score is assigned a scale score of 1290. Because only one point within the θ scale score space and the standard deviation of the scale is fixed, the scale score cut points between Level 1 and Level 2 and between Level 3 and Level 4 are free to vary across the grade and content area combinations.

Table 6-14 shows the slope and intercepts terms used to calculate the scale scores for each content area and grade. The values in Table 6-14 will not change unless the standards are reset.

Table 6-14. 2014–15 NCSC: Scale Score Slope and Intercept by Content Area and Grade

Content Area	*Grade*	*Slope*	*Intercept*
Mathematics	3	13.06	1243.67
	4	13.10	1239.87
	5	13.08	1241.41
	6	12.82	1241.25
	7	12.91	1243.24
	8	13.02	1242.36
	11	12.99	1242.48
ELA	3	11.72	1242.05
	4	12.06	1240.09
	5	12.42	1241.61
	6	12.35	1237.81
	7	12.30	1242.43
	8	12.61	1239.46
	11	11.49	1244.22

Appendix 6-H contains raw score to scale score (RS-SS) look-up tables for the four main forms. However, there were multiple additional RS-SS tables created and used for special instances (e.g., those adapted for students with no verbal mode of communication). The maximum for any grade/subject

combination was 15 in Grade 7 Mathematics. These are the actual tables used to determine student scale scores, error bands, and performance levels.

Appendix 6-I presents both the impact data for each grade by content area and the cumulative scale score distribution graphs for each grade and content area. These distributions were calculated using the sparse data matrix files that were used in the IRT calibrations. The cumulative graphs show the proportion of students at or below each scaled score.

PERFORMANCE LEVEL & SCALE SCORE DISTRIBUTIONS

Cutpoints for NCSC in ELA and mathematics were set in August 2015. Details of the standard setting procedures can be found in the standard setting report (Measured Progress, 2015c). The cuts on the theta scale that were established at those meetings are presented in Table 6-15. The θ-metric cut scores that emerged from the standard setting meetings will remain fixed throughout the assessment program unless standards are reset. Also shown in the table are the cutpoints on the reporting score scale.

.

Table 6-15. 2014–15 NCSC: Cut Scores on the Theta Metric and Reporting Scale

Content Area	Grade	Theta			Scale Score				
		Cut 1	Cut 2	Cut 3	Minimum	Cut 1	Cut 2	Cut 3	Maximum
Mathematics	3	-0.65	-0.28	0.77	1200	1236	1240	1254	1290
	4	-0.55	0.01	0.82	1200	1233	1240	1251	1290
	5	-0.84	-0.11	0.99	1200	1231	1240	1255	1290
	6	-0.61	-0.10	0.53	1200	1234	1240	1249	1290
	7	-0.91	-0.25	0.77	1200	1232	1240	1254	1290
	8	-0.66	-0.18	0.44	1200	1234	1240	1249	1290
	11	-0.70	-0.19	0.44	1200	1234	1240	1249	1290
ELA	3	-0.70	-0.18	0.72	1200	1234	1240	1251	1290
	4	-0.53	-0.01	1.43	1200	1234	1240	1258	1290
	5	-0.84	-0.13	1.16	1200	1232	1240	1256	1290
	6	-0.63	0.18	1.19	1200	1231	1240	1253	1290
	7	-0.59	-0.20	0.95	1200	1236	1240	1255	1290
	8	-0.75	0.04	0.78	1200	1230	1240	1250	1290
	11	-0.77	-0.37	0.90	1200	1236	1240	1255	1290

Table 6-16 shows the percentage of students by performance level categories along with the average and standard deviation of the scale scores for each grade/content area combination. Only the students whose scores were reported on the NCSC scale were included in the calculation of these values. Graphs of the performance level distributions are presented in Appendix 6-I. Appendix 6-J is the *NCSC Guide to Score Report Interpretation,* also referenced in Chapter 8 - Reporting Interpretation and Use.

Table 6-16. 2014–15 NCSC: Percentage of Students by Performance Level Categories

Content Area	Grade	Number of Students	Levels				Average Scale Score	SD of Scale Score
			Level 1	Level 2	Level 3	Level 4		
Mathematics	3	3,724	34.00	16.54	30.16	19.31	1239.16	20.36
	4	3,850	35.40	23.58	23.61	17.40	1236.17	18.88

Content Area	Grade	Number of Students	Levels				Average Scale Score	SD of Scale Score
			Level 1	Level 2	Level 3	Level 4		
	5	3,883	23.74	29.72	32.94	13.60	1237.71	18.82
	6	3,890	36.35	23.11	18.87	21.67	1237.27	18.98
	7	3,736	20.64	32.36	29.34	17.67	1239.28	18.99
	8	3,760	30.69	23.40	22.85	23.06	1238.27	18.51
	11	2,891	30.16	26.81	23.83	19.20	1237.04	18.82
ELA	3	3,719	35.31	16.54	26.06	22.10	1238.19	20.32
	4	3,848	41.27	13.83	33.52	11.38	1236.98	19.38
	5	3,884	28.78	24.97	31.75	14.50	1237.89	19.45
	6	3,903	38.87	26.90	20.91	13.32	1234.43	18.68
	7	3,747	38.24	15.13	27.86	18.76	1238.72	19.35
	8	3,772	32.66	28.98	17.82	20.55	1235.74	18.72
	11	3,151	33.26	18.44	31.58	16.72	1239.88	19.70

.

Linking and Equating Methods

Year to year equating has not been conducted this year because the 2014-15 NCSC was the first operational administration. However, because of the test design used (i.e., four forms for each grade/content area), within year equating has been conducted by IRT concurrent calibration. A portion of items were common across all four forms within each grade/content area.

CHAPTER 7: STANDARD SETTING

EXECUTIVE SUMMARY[9]

The purpose of this report is to summarize the activities of the standard-setting meeting for the National Center and State Collaborative (NCSC) in English language arts (ELA) and mathematics (grades 3–8 and 11). The need for standard setting arises because NCSC is a new assessment that will be administered for the first time in 2015. For this new assessment, performance standards must be set. The primary goal of the standard setting was to determine the knowledge, skills, and abilities (KSAs) that are necessary for students to demonstrate in order to be classified into each of the performance levels. The NCSC standard setting meeting was held on August 10 through 13, 2015. In all, there were 8 panels with 81 panelists participating in the process. Within each panel, the panelists were organized into two tables of four to six panelists plus a table facilitator provided by Measured Progress.

The standard setting process used was the bookmark procedure (see, e.g., Cizek & Bunch, 2007; Lewis et al., 1996; Mitzel et al., 2000). The main reason for choosing this method was that the assessment consists primarily of multiple-choice items but also includes some constructed-response items, and the bookmark procedure is appropriate for use with assessments that contain primarily or exclusively multiple-choice items, scaled using item response theory (IRT) (Cizek & Bunch, 2007). The standard setting was conducted over three rating rounds followed by a cross-grade articulation committee.

At the completion of the standard setting activities, including review of the cut scores by a cross-grade articulation panel, representatives from the partner states evaluated the recommended cut scores and related data in light of the Performance Level Descriptors, Borderline Descriptors, and Ordered Item Books. Final acceptance of the cut scores was determined by a majority vote of participating Partner States. Theta cuts at each stage of the process are shown below.

Chapter 7 Executive Summary Table-1. 2015 NCSC Standard Setting: Theta Cuts—ELA

Grade	*Process*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
3	Round 3		-0.57	-0.05	0.72
	Vertical Articulation		-0.73	-0.18	0.72
	Policy Meeting		-0.73	-0.18	0.72
4	Round 3		-0.53	-0.01	1.43
	Vertical Articulation		-0.53	-0.01	1.43
	Policy Meeting		-0.53	-0.01	1.43
5	Round 3		-0.51	-0.29	1.16
	Vertical Articulation		-0.84	-0.29	1.16

[9] All of this chapter, except the Results Summary section pp. 158-161 and the final section External Evaluation of Standard Setting, is the test administration vendor's *Final Report of the NCSC 2015 Standard Setting*.

Grade	*Process*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
	Policy Meeting		-0.84	-0.13	1.16
6	Round 3		-0.63	0.18	1.19
	Vertical Articulation		-0.63	0.18	1.19
	Policy Meeting		-0.63	0.18	1.19
7	Round 3		-0.59	-0.18	0.95
	Vertical Articulation		-0.59	-0.18	0.95
	Policy Meeting		-0.59	-0.18	0.95
8	Round 3		-0.75	0.04	0.66
	Vertical Articulation		-0.75	0.04	0.78
	Policy Meeting		-0.75	0.04	0.78
11	Round 3		-0.77	-0.37	0.52
	Vertical Articulation		-0.77	-0.37	0.90
	Policy Meeting		-0.77	-0.37	0.90

Chapter 7 Executive Summary Table 2. 2015 NCSC Standard Setting: Theta Cuts—Mathematics

Grade	*Process*	*Level 1*	*Level 2*	*Level 3*	*Level 4*
3	Round 3		-0.65	-0.37	0.77
	Vertical Articulation		-0.65	-0.28	0.77
	Policy Meeting		-0.65	-0.28	0.77
4	Round 3		-0.55	0.01	0.82
	Vertical Articulation		-0.55	0.01	0.82
	Policy Meeting		-0.55	0.01	0.82
5	Round 3		-0.84	0.14	0.99
	Vertical Articulation		-0.84	0.14	0.99
	Policy Meeting		-0.84	-0.11	0.99
6	Round 3		-0.61	-0.10	0.31
	Vertical Articulation		-0.61	-0.10	0.53
	Policy Meeting		-0.61	-0.10	0.53
7	Round 3		-0.91	-0.25	0.24
	Vertical Articulation		-0.91	-0.25	0.77
	Policy Meeting		-0.91	-0.25	0.77
8	Round 3		-0.66	-0.18	0.44
	Vertical Articulation		-0.66	-0.18	0.44

	Policy Meeting	-0.66	-0.18	0.44
11	Round 3	-0.70	-0.19	0.44
	Vertical Articulation	-0.70	-0.19	0.44
	Policy Meeting	-0.70	-0.19	0.44

Final cut scores are shown below.

Chapter 7 Executive Summary Table 3. 2015 NCSC Standard Setting: Final Cuts—ELA

Grade	*Number of Students*	*Performance Levels*	*Theta Cut*	*Percent of Students*
3	3,968	Level 1		26.56
		Level 2	-0.70	17.99
		Level 3	-0.18	31.15
		Level 4	0.72	24.29
4	4,177	Level 1		34.26
		Level 2	-0.53	20.13
		Level 3	-0.01	35.60
		Level 4	1.43	10.01
5	4,257	Level 1		23.23
		Level 2	-0.84	29.95
		Level 3	-0.13	36.67
		Level 4	1.16	10.15
6	4,300	Level 1		33.00
		Level 2	-0.63	30.00
		Level 3	0.18	26.07
		Level 4	1.19	10.93
7	4,284	Level 1		32.21
		Level 2	-0.59	16.97
		Level 3	-0.20	35.64
		Level 4	0.95	15.17
8	4,489	Level 1		27.87
		Level 2	-0.75	28.25
		Level 3	0.04	25.62
		Level 4	0.78	18.27
11	4,023	Level 1		27.96
		Level 2	-0.77	18.25
		Level 3	-0.37	34.55
		Level 4	0.90	19.24

Chapter 7 Executive Summary Table 4. 2015 NCSC Standard Setting: Final Cuts—Mathematics

Grade	*Number of Students*	*Performance Levels*	*Theta Cut*	*Percent of Students*
3	3,969	Level 1		24.82
		Level 2	-0.65	19.60
		Level 3	-0.28	35.98
		Level 4	0.77	19.60
4	4,157	Level 1		32.09
		Level 2	-0.55	27.81
		Level 3	0.01	23.36
		Level 4	0.82	16.74
5	4,237	Level 1		22.14
		Level 2	-0.84	31.37
		Level 3	-0.11	32.24
		Level 4	0.99	14.26
6	4,279	Level 1		30.38
		Level 2	-0.61	28.60
		Level 3	-0.10	17.41
		Level 4	0.53	23.60
7	4,252	Level 1		16.49
		Level 2	-0.91	32.60
		Level 3	-0.25	33.94
		Level 4	0.77	16.98
8	4,425	Level 1		25.15
		Level 2	-0.66	23.12
		Level 3	-0.18	26.24
		Level 4	0.44	25.49
11	3,758	Level 1		19.29
		Level 2	-0.70	31.00
		Level 3	-0.19	25.28
		Level 4	0.44	24.43

DESCRIPTION OF STANDARD SETTING METHODOLOGY

OVERVIEW OF STANDARD SETTING PROCEDURES

The purpose of this report is to summarize the activities of the standard-setting meeting for the National Center and State Collaborative (NCSC) in English language arts (ELA) and mathematics (grades 3–8 and 11).10 The need for standard setting arises because NCSC is a new assessment that will be administered for the first time in 2015. For this new assessment, performance standards must be set. The primary goal of the standard setting was to determine the knowledge, skills, and abilities (KSAs) that are necessary for students to demonstrate in order to be classified into each of the performance levels. The NCSC standard setting meeting was held on August 10 through 13, 2015. In all, there were 8 panels with 81 panelists participating in the process. Within each panel, the panelists were organized into two tables

[10] A number of background supporting materials is available in the appendices, over and above what is cited throughout the report. See list in the Table of Contents; see accompanying pdf of appendices for Chapter 7.

of four to six panelists plus a table facilitator provided by Measured Progress. The configuration of the panels is shown in Table 7-1.

Table 7-1. 2015 NCSC Standard Setting: Configuration of Standard Setting Panels

Panel	*Number of Panelists*	*Content Area(s)*	*Grade*	*Days*
1	11	ELA	3	Mon–Tue
			4	Tue-Wed
2	10	ELA	5	Mon–Tue
			6	Tue–Wed
3	10	ELA	7	Mon–Tue
			8	Tue–Wed
4	10	ELA	11	Mon–Tue
5	9	Mathematics	3	Mon–Tue
			4	Tue–Wed
6	10	Mathematics	5	Mon–Tue
			6	Tue–Wed
7	10	Mathematics	7	Mon–Tue
			8	Tue–Wed
8	11	Mathematics	11	Mon–Tue

The standard setting process used was the bookmark procedure (see, e.g., Cizek & Bunch, 2007; Lewis et al., 1996; Mitzel et al., 2000). The main reason for choosing this method was that the assessment consists primarily of multiple-choice items but also includes some constructed-response items, and the bookmark procedure is appropriate for use with assessments that contain primarily or exclusively multiple-choice items, scaled using item response theory (IRT) (Cizek & Bunch, 2007).

Organization of This Report

This report is organized into three major sections, describing tasks completed (1) prior to, (2) during, and (3) after the standard setting meeting.

TASKS COMPLETED PRIOR TO THE STANDARD SETTING MEETING

Creation of Performance Level Descriptors

NCSC developed the performance level descriptors (PLDs) for mathematics and English language arts (ELA) at grades 3–8 and 11 through an iterative process involving multiple stakeholder groups (see Appendix 7-A). The NCSC partnership developed grade-level PLDs to summarize the knowledge, skills, and abilities (KSAs) prioritized for the NCSC assessments that students need to attain at each level of achievement (Level 1–Level 4).

Phase 1. Initial Draft

A committee of NCSC organizational partners with expertise in content, measurement, and/or significant disabilities developed the initial draft of the PLDs as described by Sireci, Hambleton, and Bahry (2013). Initial development culminated in a review by state and organizational partners focused on:

congruence with NCSC prioritized grade-level academic content; progression across performance levels; progression across grades; and consistency of grade-level expectations and specificity across mathematics and ELA.

Phase 2. Revision and Refinement of Initial Draft

NCSC content and measurement experts revised and refined the initial draft grade-level PLDs based on iterative reviews by NCSC state partners, organizational partners, and Technical Advisory Committee (TAC) members, including the input from Phase 1. Based on the comprehensive input received across these reviews, these experts implemented a three-step process to support ongoing development of the grade-level PLDs.

Step One. NCSC content and measurement experts analyzed the draft grade-level PLDs to account for the degree to which assessment characteristics impacted student performance descriptors including Depth of Knowledge (DOK), overall difficulty of content and concepts, item features, item supports, and passage complexity (ELA reading only).

Step Two. The second step in the process focused on articulating three types of expectations vertically across grades 3–8 and grade 11:

- Student Learning Expectations based on grade-specific learning outcomes focus on instructional content and describe end-of-year learning expectations for mathematics and ELA at each grade level.
- Measurement Expectations describe the knowledge and skills defined through the academic content prioritized for assessment at each grade level.
- Measurement Targets focus more narrowly, when applicable, on the subset of prioritized expectations used to develop items for the spring 2015 operational assessment. In some cases the Measurement Target represented the full Measurement Expectation.

The three sets of specifications resulting from this work showed NCSC's progression of expectations for learning and assessment within and across grade levels. The Student Learning and Measurement Expectations provided a context for interpreting student performance using the PLDs while the Measurement Targets provided a direct tool for refining and/or revising the grade-level PLDs.

Step Three. Developers used the Measurement Targets to examine the most current draft grade-level PLDs within and across grades. They refined and revised the descriptors at each performance level to ensure representation of intended expectations, differentiation across performance levels, and representation of the impact of graduated complexity on a student's ability to demonstrate the expected KSAs.

Phase 3. Finalizing the PLDs

The NCSC State Partners and TAC provided a final round of feedback on the PLDs. State partners provided feedback to specific questions that addressed descriptions of text and task complexity; inclusion of foundational skills; language used for the writing descriptors; and the format of the PLDs for use in standard setting. Input from these reviews was compiled and combined with item performance data from the NCSC pilots to prepare the final draft grade-level PLDs.

Summary

NCSC's iterative and comprehensive development process resulted in clarity with respect to how content and performance expectations as well as complexity and support change within and across grade levels; explication of the dimensionality embedded in the grade-level PLDs and how components of that dimensionality interact; development of a framework around the PLDs that provides a context for interpreting student performance; creation of an overall description of student learning expectations at each grade level; and streamlining the PLDs while ensuring the language is interpretable to the intended audiences. In addition, NCSC's in-depth examination of the grade-level PLDs within and across grade levels and content areas focused on ensuring the PLDs provided progressive descriptions of what students with the most significant cognitive disabilities are expected to know and be able to do as an outcome of progress across grades toward the end goal of college, career, and community readiness. The descriptors aligned with the KSAs prioritized for the NCSC assessments. NCSC's development process resulted in a context for interpretation and use of the grade-level PLDs that ensured clarity and supported the connection between the measurement and instructional contexts developed within the NCSC system. For key dates and tasks in the PLD development process, see Table 7-2 below. For additional information, see Appendix 7-B.

Table 7-2. 2015 NCSC Standard Setting: PLD Development Key Dates, Tasks, and Groups

Date	*Task/Meeting*	*Groups*
March and April 2013	Development of initial mathematics and ELA PLDs	▪ University of Massachusetts–Amherst (UMASS) ▪ NCSC Organizational Partners with content, measurement, and/or significant disabilities expertise
May 2013	NCSC PLD Review Meeting	▪ University of Massachusetts–Amherst (UMASS) ▪ NCSC State Partners ▪ NCSC Organizational Partners with content and significant disabilities expertise
August 2013	TAC Meeting: Review of Revised PLDs	▪ NCSC Technical Advisory Committee (TAC) ▪ NCSC Organizational Partners with content and measurement expertise
September 2013	Ad Hoc Presentation: Review of Revised, Draft PLDs	▪ NCSC Organizational Partners with content and measurement expertise ▪ NCSC State and Organizational Partners
January 2014	Ad Hoc Presentation: Clarifying the Grade-Level PLDs	▪ NCSC Organizational Partners with content and measurement expertise ▪ NCSC State and Organizational Partners
February 2014	Development of draft Student Expectations Profiles and draft prototype Measurement Expectations Profiles for mathematics and ELA for grades 3–8 and 11	▪ NCSC Organizational Partners including content, measurement, and significant disabilities experts
September 2014	Presentation of the activities/meetings contributing to the development and review of draft grade-level PLDs	▪ NCSC TAC ▪ NCSC Organizational Partners with content and measurement experts
November 2014	Review Meeting of the draft NCSC mathematics and ELA PLDs	▪ NCSC Organizational Partners with content and measurement experts ▪ State and Local Education Agency special educators and content experts
January 2015	TAC Meeting: Presentation of updated draft ELA PLDs	▪ NCSC TAC

Date	*Task/Meeting*	*Groups*
	and inclusion of writing	▪ NCSC State and Organizational Partners
January–March 2015	Evaluation and application of recommendations and feedback	▪ NCSC content and measurement experts ▪ NCSC Organizational Partners
March 2015	Review of the mathematics and ELA PLDs	▪ NCSC Organizational Partners content and measurement experts ▪ NCSC State and Organizational Partners
April 2015	Review of the revised draft mathematics and ELA PLDs and Front Matter and presentation of updated drafts	▪ NCSC Organizational Partners content and measurement experts ▪ NCSC Assessment Steering Committee ▪ NCSC TAC
May 2015	Presentation of the final draft PLDs and Front Matter	▪ NCSC Organizational Partners content and measurement experts ▪ NCSC Assessment Steering Committee
June 2015	Presentation and approval of the final mathematics and ELA PLDs and the Front Matter for Standard Setting	▪ NCSC Organizational Partners content and measurement experts ▪ NCSC State partners

MATERIALS FOR PANELISTS[11]

The following materials were assembled for presentation to the panelists at the standard setting meeting:

- Meeting agenda – included schedule for the duration of the standard setting.
- Non-Disclosure Agreement form – Each panelist signed to indicate none of the secure materials contents would be disclosed.
- PLD front matter – Provides background, interpretive information, and context for the PLDs
- PLDs – Performance level descriptors of grade and content appropriate for panelist
- Test booklets – Panelists were asked to take the test to give them a feel for the student experience.
- Answer keys – Keys for test booklets and ordered item booklets were provided.
- Ordered item booklets (OIBs) – Primary tool of the standard setting to gather evidence to determine cut scores. The ordered item booklet contained one item per page, ordered from the easiest item to the most difficult item. The primary task of the panelist is to place a bookmark between the items that divide one performance level from another.
- Item map forms - The item map form listed the items in the same order that they were presented in the ordered item booklet; the form included space for the panelists to write in the KSAs required to answer each item correctly. There was also space for the panelists to explain why they believed each item was more difficult than the previous one.
- Rating forms – Form used to record panelist ratings during each rating round.
- Evaluation forms – Training, procedural, and final evaluation forms were completed by the panelists.

OIBs were created from the NCSC item bank in consideration of both content and statistical representativeness of the items. None of them directly corresponded with the operational intact forms. NCSC selected approximately 30 items for ELA and 40 items for mathematics within each grade. In the subsequent results, the raw score cuts correspond to the raw scores examinees could have scored if they were administered the items in the OIBs.

PRESENTATION MATERIALS

The Standard Setting Process slide presentation used in the opening session was prepared prior to the meeting. A copy of the presentation is included in Appendix 7-H.

INSTRUCTIONS FOR FACILITATORS

Scripts were created for the group facilitators to refer to while working through each step of the standard setting process. This document is included in Appendix 7-I. NCSC reviewed and approved the facilitator scripts prior to the standard setting.

SYSTEMS AND MATERIALS FOR ANALYSIS DURING THE MEETING

The computational programming used to calculate cut scores and impact data during the standard setting meeting was completed and thoroughly tested prior to the standard setting meeting. See Section 3.10, Tabulation of Round 1 Results, for a description of the analyses performed during standard setting.

[11] The agenda for the standard setting meeting is provided in Appendix 7-C. See additional samples of meeting materials in Appendices 7-D-7-I. Also see panelist list in Appendix 7-J.

STAFFING ROLES

Measured Progress provided the staff displayed in Table 7-3 to conduct the standard setting meeting. Facilitators were selected based on their expertise in alternate assessment and prior standard setting experience. NCSC reviewed the qualifications of the facilitators and approved them based on their experience with either facilitating other standard setting meetings and/or working with states to facilitate their work with students with significant cognitive disabilities.

A facilitator training took place approximately one week prior to the standard setting meeting on August 4th, 2015. The purpose for the training was to prepare the standard setting facilitators for the committee activities, allow for minor modifications to the procedures and materials (if necessary), and allow for the NCSC staff and partner states to be aware of, and participate in, standard setting activities, should they so choose.

Psychometric staff from Measured Progress conducted the training. During the meeting, the written procedures for the standard setting activities were reviewed. Copies of all materials and handouts were distributed and explained, and the actual activities planned for the standard setting committees were used to ensure they are clear and workable. The training also described the responsibilities of the facilitators, which included leading and keeping the panelists on task throughout the standard setting process, ensuring that all panelists clearly understand the procedures, tracking the standard setting materials, and ensuring that the rating forms and evaluation forms are completed.

Table 7-3. 2015 NCSC Standard Setting: Measured Progress Staff

Role	*Name*	*Responsibility*
Lead Facilitator	Susan Izard	▪ Presentation of the NCSC Overview PowerPoint ▪ Provide communication between the facilitators and the Data Analysis Room ▪ Move from room to room to monitor for consistency
Lead Psychometrician	Jennifer Dunn	▪ Presentation of the Setting Standards for the NCSC Assessment PowerPoint ▪ Presentation of impact data and cut scores to small groups ▪ Move from room to room to monitor for consistency ▪ Oversee the data analysis process
Psychometrician	Han Yi Kim	▪ Presentation of impact data and cut scores to small groups ▪ Data entry and assistance to lead psychometrician and data analyst as needed
Data Analyst	Carly Gumpert	▪ Perform data analysis ▪ Data entry and assistance to psychometricians as needed
Data Entry and Logistics	Adriane Hoitt	▪ Data Entry ▪ Assist with QC of data ▪ Set up and staff the registration table ▪ Assist panelists with accommodations ▪ Interact with facility staff ▪ Provide communication between the facilitators and the Data Analysis Room as needed
Facilitator	Theresa Fulton	ELA 3/4
	Alicia Cuttle	ELA 5/6

Role	*Name*	*Responsibility*
	Tina Haley	ELA 7/8
	Charlene Newton	ELA 11
	Kristen Cole	Mathematics 3/4
	Chris Clough	Mathematics 5/6
	Betsy Rogers	Mathematics 7/8
	Sally Blake	Mathematics 11

Table 7-4. Facilitator Experience

Name	*NCSC Panel*	*List of specific standard setting experiences, including facilitation and/or panelist training*
Theresa Fulton	ELA 3/4	▪ Florida Alternate Assessment: 2012 and 2014, elementary ELA item review meeting facilitation ▪ Maryland: November 2014, February and March 2015, elementary ELA item review meeting facilitation
Alicia Cuttle	ELA 5/6	▪ Florida Alternate Assessment: April 2008, science, middle school; July 2008, reading, elementary ▪ New Mexico: 2011, reading, grade 5 ▪ New York Alternate Assessment: 2007, reading, middle school; 2014, floater
Tina Haley	ELA 7/8	▪ New Hampshire Alternate Assessment: 2011, Reading, Grades 2, 3, and 4
Charlene Newton	ELA 11	▪ Panelist Training: Developing Achievement Levels on the National Assessment of Educational Progress (NAEP): 2011, writing, grades 8 and 12 ▪ Washington Alternate Assessment: 2012, reading, high school
Kristen Cole	Mathematics 3/4	▪ Florida Alternate Assessment: 2009, science, grade 5 ▪ Nevada Alternate Assessment: 2010, ELA, grade 7 ▪ Maine Alternate Assessment: 2010, math, grade 5 ▪ New Mexico General Assessment: 2011, reading, grade 3 ▪ Washington Alternate Assessment: 2012, reading, grade 4
Chris Clough	Mathematics 5/6	▪ NAGB standard setting: 2011, facilitator and panelist training materials and Q&A ▪ New York Alternate Assessment: 2014, math, grade 6
Betsy Rogers	Mathematics 7/8	▪ New York Alternate Assessment: 2014, math, elementary ▪ Mississippi Alternate Assessment: 2015, science, grade 8
Sally Blake	Mathematics 11	▪ Kentucky General Assessment: Early 1990s,

Name	NCSC Panel	List of specific standard setting experiences, including facilitation and/or panelist training
		elementary math ▪ Massachusetts General Assessment: 1998, elementary math; 2006, math standards validation, elementary ▪ Wyoming General Assessment: 2000, elementary math

SELECTION OF PANELISTS

NCSC recruited 85 standard setting panelists from NCSC member states. Each NCSC state partner provided nominations for the standard setting panel, with a focus on providing a diverse group of potential panelists and paying particular attention to:

- Experience with special education
- Experience with general education
- Experience as a building administrator
- Experience with blind and/or deaf students
- Experience with ELLs
- Race
- Gender

NCSC made initial selections to balance these priorities, with the highest priority given to ensuring each state had representation on all panels. Following this, NCSC ensured panelist selection focused on nominees who represented a range of positions within their local educational agency, including special education teachers, general education teachers, and building administrators. NCSC's third selection priority focused on special education teachers with experience teaching special populations, including blind, deaf, or deaf-blind students and English language learners. Finally, NCSC attempted to balance the race and ethnicity of the panelists.

Once the initial selections were complete, NCSC notified nominees of their selection for the standard setting panel. If a nominee chose not to participate or did not reply, NCSC selected a replacement from the same state's nominee pool. If no other nominees were available, they chose a nominee from a different state. No more than eight panelists were chosen from a single state. NCSC followed this process until each panel had 11 confirmed panelists. Due to last minute cancelations, some panels had only 10 panelists.

Two panelists from each of the English language arts (ELA) and mathematics panels were asked to function as table leaders. Table leader responsibilities centered on facilitating, recording, and sharing panelist input from the small group discussions. Training was provided via webinar the week before standard setting. The goal of the training was to familiarize the table leaders with the standard setting process and to explain the responsibilities of and key considerations for the role. The table leaders from each panel also participated in the cross-grade articulation committee at the completion of the standard setting process (described below).

State

NCSC placed the highest priority on evenly recruiting panelists from all states. Each state was guaranteed at least six panelists so long as they provided enough nominees. Smaller member organizations, including PAC-6 and U.S. Virgin Islands, were permitted two panelists each. Table 7-5 shows the number of panelists by state.

Table 7-5. 2015 NCSC Standard Setting: State of Panelist by Grade Group

States	English Language Arts				Mathematics				Total
	3-4	5-6	7-8	11	3-4	5-6	7-8	11	
Arkansas	1	1	1	1	1		1		6
Arizona	1	1	1	1	1	1	1	1	8
PAC-6 [CNMI-GU)		1					1		2
Connecticut	1		1	1	1	1	1	1	7
Washington, DC	1	1	1		1	1	1	1	7
Idaho	1	1	1		1	1		1	6
Indiana		1				1		1	3
Maine	1	1	1	1	1	1	1	1	8
Montana	1		1	1	1	1	1	1	7
New Mexico	1	1	1	1	1	1	1	1	8
Rhode Island	1	1	1	1	1	1	1	1	8
South Carolina	1	1		1	1	1	1	1	7
South Dakota	1	1	1	1	1			1	6
Virgin Island				1			1		2
Total	11	11	10	10	11	10	11	11	85

Educator Type

NCSC organized panels to ensure each had a representation of special educators, general educators, and building administrators. Table 7-5 shows the number of administrators, general educators and special educators on each panel. Within this table, panelists are counted twice if they listed multiple types of experience. For example, NCSC counted an administrator who was also a special educator twice in Table 7-6. During recruiting, NCSC asked nominees to indicate their experience with regard to the types of classroom in which they have taught (e.g., general education, special education, etc.), grade-level(s) they have taught (e.g., 3rd, 4th, 5th, etc.), and content-area(s) they have taught (ELA or mathematics). This resulted in an uneven distribution of nominees across grades and content areas. In some nominee groups, panelists indicated a more diverse background than in others. For example, more Grade 3-4 mathematics nominees listed experience in both general and special education classrooms as compared to Grade 3-4 ELA nominees. The results in an uneven distribution of across grade level panels based on educator type. Importantly, each content area panel was comprised of educators representing a range of roles and experiences relevant to the task of a panelist.

Table 7-6. 2015 NCSC Standard Setting: Educator Type by Grade Group

Educator Type	English Language				Mathematics			
	3-4	5-6	7-8	11	3-4	5-6	7-8	11
Administrator	1	2	1	3	1	1	4	3
General	1	2	1	1	4	1	1	3
Special	10	10	9	8	8	9	7	9
Missing		1					1	

Experience with Special Populations

NCSC also selected panelists with experience with blind, deaf, or deaf-blind students to ensure that each panel had at least two nominees with this experience. Table 7-7 shows the number of panelists with this type of experience. The next priority was panelist experience with ELLs. Again, NCSC selected panelists so that each

panel had one member with this type of background. Table 7-7 shows the number of panelists with this experience.

Table 7-7. 2015 NCSC Standard Setting: Panelist Experience with Special Populations by Grade Group

Special Population Experience	*English Language Arts*				*Mathematics*			
	3-4	*5-6*	*7-8*	*11*	*3-4*	*5-6*	*7-8*	*11*
Blind, Deaf, DB	2	4	4	3	4	2	3	6
English Language	6	6	5	4	2	1	4	5

Gender and Race

States provided lists of nominees that were mostly white females so it was not possible to create panels balanced for gender. Table 7-8 shows the numbers of females and males on each panel. Table 7-9 shows the number of white and non-white panelists on each panel.

Table 7-8. 2015 NCSC Standard Setting: Panelist Gender by Grade Group

Gender	*English Language Arts*				*Mathematics*				*Total*
	3-4	*5-6*	*7-8*	*11*	*3-4*	*5-6*	*7-8*	*11*	
Female	9	10	8	7	10	8	8	8	68
Male	1	1	1	2			1		6
Missing	1		1	1	1	2	2	3	11
Total	11	11	10	10	11	10	11	11	85

Table 7-9. 2015 NCSC Standard Setting: Panelist Race by Grade Group

Race	*English Language Arts*				*Mathematics*				*Total*
	3-4	*5-6*	*7-8*	*11*	*3-4*	*5-6*	*7-8*	*11*	
White	7	7	8	7	8	7	5	5	54
Non-White	3	2	1	2	2	1	4	2	17
Missing	1	2	1	1	1	2	2	4	14
Total	11	11	10	10	11	10	11	11	85

TASKS COMPLETED DURING THE STANDARD SETTING MEETING

OVERVIEW OF BOOKMARK METHOD

The bookmark method (Cizek & Bunch, 2007; Lewis et al., 1996; Mitzel et al., 2000) involves rank ordering the items by difficulty and asking the panelists to identify the point in the ordered set of items at which the students at the borderline of two performance levels would no longer answer the item correctly.

The ordered item booklet (OIB) contained one item per page, ordered from the easiest item to the most difficult item. Since none of the four operational forms administered in 2014-15 appropriately reflected the NCSC test blueprints, a separate standard setting form was created. Items included in the OIBs were selected by NCSC content experts to represent the intended blueprints as closely as possible. The ordered item booklet was created by sorting the items according to their item response theory (IRT)-based difficulty values. A two-parameter logistic IRT model was used to calculate the response probability (RP) values for dichotomous items. The two-parameter model was chosen as it provided the best fit to the data (over, for example, the one-parameter Rasch model). Also, while some ELA passage-based clusters of items are scored polytomously, the component items were treated separately for the purposes of the developing the OIB. In addition, ELA grades 3 and 4 have "Foundational Items", which are clusters of 3 or 5 selected-response items which are collectively scored dichotomously, and were displayed in an identical manner in the OIB. As such, there was no need for the use of polytomous IRT models in developing the OIB.

In developing the OIB two different response probabilities were used. For ELA, an RP value of 0.67 was used and mathematics used an RP value of 0.50. Each of these RP values signify the ability level students needed to have to achieve a 67% (or 50% for mathematics) chance of correctly answering an item as discussed below.

The basic procedure for the bookmark method is as follows: Beginning with the first ordered item and considering the KSAs needed to complete it, panelists ask themselves, "Would at least two out of three (or 50% of the) students performing at the borderline of Level 3 answer this question correctly?" Panelists considered each ordered item in turn, asking themselves the same question. They placed the bookmark between the two items where their answer changed from "yes" (or predominantly "yes") to "no" (or predominantly "no"). Panelists then repeated the process for Level 2 and Level 4 cuts and used the rating form to record their ratings for each cut.

The RP values were selected to best facilitate the panelists asking the above question. In general, the mathematics items were more difficult and an RP value of 0.67 would have required them to place bookmarks very early in the OIB (perhaps even before the first page for the first cut). Generally speaking this is not desirable, so an RP value of 0.50 was selected to allow bookmark placement later in the OIB. Note, however, that some of the items only had two choices for the student to choose. As a result, these items would have a guessing chance of 50%, which makes answering the standard Bookmark question, "Would an examinee just barely at this level have a 50% chance of answering this item correctly?", awkward for the panelists to answer. As such, the question was restructured and slightly reframed to include "at least 50%", so that conceptually the question was more straightforward to answer. ELA items were of more moderate difficulty, so an RP value of 0.67 provided better fit to the items in terms of bookmark placement. While ELA also had items with only two choices, since the RP value of 0.67 was greater than the guess chance of 0.50, there was not the same conceptual issue as mathematics. However, for the purposes of consistency the same question structure ("at least 67%") was employed.

Each panel was responsible for recommending standards for two grade levels, with the exception of those recommending standards for grade 11. The grade 11 panelists set standards for only a single grade and content area. Therefore, the results presented in Tables below represent a repetition of the process by each panel. In each case, a panel would complete the process for its first grade level, starting with the review of the assessment materials and ending with the Round 3 ratings, and then repeat the entire process one more time for the remaining grade level.

ORIENTATION

With regard to panelist training, Standards for Educational and Psychological Testing states the following:

Care must be taken to assure these persons understand what they are to do and that their judgments are as thoughtful and objective as possible. The process must be such that well-qualified participants can apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions (AERA/APA/NCME, 2014, p. 101).

The training of the panelists began with a general orientation at the start of the standard setting meeting. The purpose of the orientation was to ensure that all panelists received the same information about the need for and goals of standard setting and about their part in the process. First, Ms. Susan Izard, Director of Special Education at Measured Progress, greeted the panels and introduced NCSC member state representatives and Measured Progress staff on-site. Then, Dr. Phyllis Lynch, Director of Office of Instruction, Assessment and Curriculum at Rhode Island Department of Education (member of NCSC Steering Committee), provided some pertinent context about the timeline and development of the NCSC assessment and an introduction to the issues of standard setting. Dr. Lynch also provided an overview of the NCSC assessment design, item types, and participation criteria. Next, Dr. Jennifer Dunn, Director of Psychometrics at Measured Progress presented an overview of the bookmark procedure and the activities that would occur during the standard setting meeting. This included a walkthrough of the training procedures, a step-by-step overview of the how the panelists will apply the Bookmark method, and details on what will occur during each round of the method. Additional details can be found in the PowerPoint slides displayed in Appendix 7-H. After the general orientation was complete, each panel convened in a breakout room, where the panelists received more detailed training from their facilitator and completed the standard setting activities.

REVIEW OF ASSESSMENT MATERIALS

The first step after the opening session was for the panelists to become familiar with the NCSC assessment. The facilitators provided an overview of the assessment. Then, each item was projected and read aloud to the panelists from the Directions for Test Administration (DTA) to closely mirror student experience. The answer key for each item was provided after each item was presented. The purpose of the step was to help the panelists establish a good understanding of the test items and to gain an understanding of the experience of the students who take the assessment.

REVIEW OF PERFORMANCE-LEVEL DESCRIPTOR (PLD) FRONT MATTER AND PLDS

After taking the test, panelists reviewed the PLD front matter and the PLDs. This important step was designed to ensure that panelists thoroughly understood the knowledge, skills, and abilities (KSAs) needed for students to be classified into performance levels (Level 1, Level 2, Level 3, and Level 4). Panelists first reviewed the PLD front matter and the PLDs on their own and then participated in group discussion, clarifying each level. The PLD front matter and PLDs are provided in Appendix 7-B.

COMPLETION OF THE ITEM MAP FORM

Panelists then completed the item map form. The item map form listed the items in the same order that they were presented in the ordered item booklet; the form included space for the panelists to write in the KSAs

required to answer each item correctly. There was also space for the panelists to explain why they believed each item was more difficult than the previous one.

The purpose of this step was to ensure that panelists became familiar with the ordered item booklet and understood the relationships among the ordered items. Each panelist reviewed the ordered item booklet item by item, considering the KSAs students needed to answer each one. The panelists recorded this information on the item map form along with reasons why each item was more difficult than the previous one. After they finished working individually, panelists had the opportunity to discuss the item map form as a group and make necessary additions or adjustments.

Discussion of PLDs and Borderline Students

Panelists had another opportunity to individually review the PLDs as needed. Afterward, panelists developed consensus definitions of borderline students—that is, students who have only barely qualified for a particular performance level. Bulleted lists of characteristics for each level were generated based on the whole-group discussion and posted in the room for reference throughout the bookmark process. Note that the purpose of this step was to clarify and add specificity to the PLDs based on the KSAs identified for each item in the previous step (Completion of the Item Map Form), with particular attention to the definitions of the borderline students. The bulleted lists were developed as working documents to be used by the panelists for the purposes of standard setting. They supplemented the PLDs, which provide the official definition of what it means for a student to be classified into each performance level, by specifically addressing the KSAs that define the borderline of each level.

Practice Round

Next, the panelists completed a practice round of ratings. Note that at this point, panelists were grouped into two table groups of approximately five panelists each. The purpose of the practice round was to familiarize the panelists with all the materials they would be using for the standard setting process and to walk them through the process of placing bookmarks. In addition to the PLDs and borderline descriptions, panelists were given a practice ordered item booklet, which consisted of five items (two easy, two difficult, and one moderately difficult), and a practice rating form.

The facilitator explained what each of the materials was and how panelists would use it to make their ratings. Then, beginning with the first ordered item and considering the skills and abilities needed to complete it, panelists were instructed to ask themselves, “Would at least two out of three students performing at the borderline of Level 3 answer this question correctly?” for ELA, and “Would at least 50% of the students performing at the borderline of Level 3 answer this question correctly?” for mathematics. Panelists considered each ordered item in turn, asking themselves the same question until their answer changed from “yes” (or predominantly “yes”) to “no” (or predominantly “no”). Each panelist practiced placing the Level 3 bookmark in the practice ordered item booklet. The facilitator then led the panelists in a readiness discussion, asking panelists to share the reasoning behind their bookmark placements with the group and assessing each panelist’s understanding of the rating task, borderline students, and the two-thirds (or 50%) rule.

TRAINING EVALUATION

At the end of the practice round, panelists completed the training evaluation form. The evaluation form was designed to ascertain whether the panelists were comfortable moving ahead to the rating task or whether there were lingering questions or issues that needed to be addressed before proceeding to the Round 1 ratings. Facilitators were instructed to glance over each panelist's evaluation as he or she completed it to make sure panelists were ready to move on. Table 7-10 below provides summary results collapsed across grade and content area.

Table 7-10. 2015 NCSC Standard Setting: Training Evaluation Results Summary

	N	*Mean*	*% SD*	*% D*	*% A*	*% SA*
I understand the goals of the standard setting meeting.	82	3.70	0.00%	0.00%	30.49%	69.51%
I understand the procedures we are using to set standards.	82	3.70	0.00%	0.00%	30.49%	69.51%
I understand how to use the standard setting materials.	82	3.61	0.00%	0.00%	39.02%	60.98%
I understand the differences between the performance levels.	82	3.37	0.00%	4.88%	53.66%	41.46%
I understand how to make the cut score judgment.	82	3.49	0.00%	0.00%	51.22%	48.78%
I am confident in my conceptualization of better than 50% (or 67% for ELA) of the borderline students answering questions correctly.	82	3.37	0.00%	2.44%	58.54%	39.02%
I know what tasks to expect for the remainder of the meeting.	82	3.44	0.00%	3.66%	48.78%	47.56%
I am confident in my understanding of the standard setting task.	82	3.51	0.00%	1.22%	46.34%	52.44%

The full results by grade and content area of the training evaluation can be found in Appendix 7-K.

ROUND 1 JUDGMENTS

In the first round, panelists worked individually with the PLD front matter, the PLDs, the item map form, and the ordered item booklet. Beginning with the first ordered item and considering the KSAs needed to complete it, they asked themselves, "Would at least two out of three (or 50% of the) students performing at the borderline of Level 3 answer this question correctly?" Panelists considered each ordered item in turn, asking themselves the same question. They placed the bookmark between the two items where their answer changed from "yes" (or predominantly "yes") to "no" (or predominantly "no"). Panelists then repeated the process for Level 2 and then Level 4, and used the rating form to record their ratings for each cut.

TABULATION OF ROUND 1 RESULTS

After the Round 1 ratings were complete, the Measured Progress staff members calculated the median cut scores for the tables based on Round 1 bookmark placements. Cut scores were calculated using Statistical Analysis Software (SAS). First, each panelist's cutpoints were found on the theta scale by averaging the RP0.67 (or RP0.50) values of the items on either side of the bookmark placed by that panelist for each cut. For a given cutpoint, the median was taken across all panelists. Using this methodology, all cutpoints were determined on the theta scale. Because the NCSC assessment is constructed and equated using IRT analyses, use of an IRT-based standard setting method and calculating cuts on the theta metric is the technically sound choice (Cizek & Bunch, 2007). The theta scale established for the 2015 NCSC forms will be the reference scale for equating future test forms, and thus the cutpoints on the theta scale will represent a comparable level of performance across forms and years.

The results of the panelists' Round 1 ratings and associated impact data are outlined in Tables 3-2 and 3-3. Note that all impact data were computed using the available data at the time of standard setting. The median theta cuts for each performance level along with the associated raw score cuts are shown. In addition, the median absolute deviation (MAD) for the assessment was provided for each of the scale cuts, which gives an indication of the extent to which judgments were consistent across panelists and reflects the level of agreement among the ratings with each successive round of ratings. Finally, impact data—reflecting the percentage of students across the NCSC states that would fall into each performance level category according to the Round 1 total group median cutpoints—were calculated. Table-level results for all rounds are displayed in Appendix 7-L. The Round 1 results are outlined in Tables 7-11 and 7-12.

Table 7-11. 2015 NCSC Standard Setting: Round 1 Results—ELA

Grade	*Performance Levels*	*Median Theta Cut*	*Median Absolute Deviation*	*Percent of Students*
3	Level 1			31.70
	Level 2	-0.57	0.02	18.50
	Level 3	-0.06	0.12	25.50
	Level 4	0.72	0.18	24.29
4	Level 1			34.26
	Level 2	-0.53	0.03	20.13
	Level 3	-0.01	0.05	35.60
	Level 4	1.33	0.06	10.01
5	Level 1			39.75
	Level 2	-0.51	0.08	13.44
	Level 3	-0.17	0.16	36.67
	Level 4	1.27	0.13	10.15
6	Level 1			33.00
	Level 2	-0.63	0.07	36.72
	Level 3	0.38	0.07	24.35
	Level 4	1.59	0.17	5.93
7	Level 1			32.21
	Level 2	-0.59	0.02	16.97
	Level 3	-0.24	0.03	20.54
	Level 4	0.39	0.10	30.28

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
8	Level 1			27.87
	Level 2	-0.75	0.04	28.25
	Level 3	0.04	0.05	25.62
	Level 4	0.68	0.09	18.27
11	Level 1			23.02
	Level 2	-0.83	0.04	14.27
	Level 3	-0.51	0.05	16.83
	Level 4	-0.02	0.13	45.89

Table 7-12. 2015 NCSC Standard Setting: Round 1 Results—Mathematics

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
3	Level 1			24.82
	Level 2	-0.65	0.08	15.97
	Level 3	-0.37	0.03	26.25
	Level 4	0.28	0.11	32.96
4	Level 1			28.10
	Level 2	-0.59	0.05	26.68
	Level 3	-0.06	0.07	28.48
	Level 4	0.82	0.15	16.74
5	Level 1			29.22
	Level 2	-0.61	0.09	36.54
	Level 3	0.14	0.06	22.85
	Level 4	1.05	0.05	11.40
6	Level 1			30.38
	Level 2	-0.61	0.04	28.60
	Level 3	-0.10	0.03	11.50
	Level 4	0.34	0.03	29.52
7	Level 1			11.41
	Level 2	-0.93	0.08	37.68
	Level 3	-0.25	0.11	17.87
	Level 4	0.20	0.13	33.04
8	Level 1			25.15
	Level 2	-0.66	0.07	28.00
	Level 3	-0.11	0.05	21.36
	Level 4	0.44	0.06	25.49
11	Level 1			19.29
	Level 2	-0.70	0.04	38.03
	Level 3	-0.03	0.03	18.25
	Level 4	0.44	0.04	24.43

ROUND 2 JUDGMENTS

The purpose of Round 2 was for panelists to discuss their Round 1 placements and, if necessary, to revise their ratings. As noted earlier, panelists were grouped into two table groups of approximately five panelists each. Prior to beginning their discussions, the panelists at each table were presented with the table-level median cut points based on the Round 1 ratings for the panelists at that table in addition to the overall group median. The median cutpoints were presented in terms of location in the ordered item booklet. The panelists at each table then shared their individual rationales for their bookmark placements in terms of the necessary knowledge and skills for each classification. Panelists were asked to pay particular attention to how their individual ratings compared to those of the others at their table and get a sense for whether they were unusually stringent or lenient within the group. Panelists were asked to consider the whole-group median cutpoints as well. Psychometricians presented the information to the group with projected tables and figures, and explained how to use it as they completed their Round 2 discussions. Sample tables and figures that were shown to the panelists after each of the rounds are displayed in Appendix 7-N.

Panelists were told to set bookmarks according to their individual best judgments; consensus among the panelists was not necessary. Panelists were encouraged to listen to the points made by their colleagues but not to feel compelled to change their bookmark placements. Once the discussions were complete, panelists were given the opportunity to revise their Round 1 ratings on the rating form.

TABULATION OF ROUND 2 RESULTS

When Round 2 ratings were complete, the Measured Progress data analysis team calculated the median cut scores for the room and associated impact data. The results of the panelists' Round 2 ratings are outlined in Tables 7-13 and 7-14.

Table 7-13. 2015 NCSC Standard Setting: Round 2 Results—ELA

Grade	*Performance Levels*	*Median Theta Cut*	*Median Absolute Deviation*	*Percent of Students*
3	Level 1			31.70
	Level 2	-0.57	0.01	18.50
	Level 3	-0.05	0.08	25.50
	Level 4	0.72	0.06	24.29
4	Level 1			34.26
	Level 2	-0.53	0.02	15.90
	Level 3	-0.05	0.03	39.84
	Level 4	1.33	0.05	10.01
5	Level 1			39.75
	Level 2	-0.51	0.06	7.66
	Level 3	-0.29	0.06	35.94
	Level 4	1.03	0.16	16.65
6	Level 1			33.00
	Level 2	-0.63	0.02	30.00
	Level 3	0.24	0.05	26.07
	Level 4	1.19	0.16	10.93
7	Level 1			32.21
	Level 2	-0.59	0.01	16.97

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
	Level 3	-0.17	0.01	25.16
	Level 4	0.59	0.09	25.65
8	Level 1			27.87
	Level 2	-0.75	0.00	35.02
8	Level 3	0.17	0.02	18.85
	Level 4	0.67	0.03	18.27
11	Level 1			27.96
	Level 2	-0.77	0.01	18.25
	Level 3	-0.37	0.02	25.88
	Level 4	0.52	0.08	27.91

Table 7-14. 2015 NCSC Standard Setting: Round 2 Results—Mathematics

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
3	Level 1			24.82
	Level 2	-0.65	0.00	15.97
	Level 3	-0.37	0.00	39.61
	Level 4	0.77	0.06	19.60
4	Level 1			32.09
	Level 2	-0.55	0.01	22.68
	Level 3	-0.06	0.05	28.48
	Level 4	0.82	0.02	16.74
5	Level 1			22.14
	Level 2	-0.84	0.08	43.62
	Level 3	0.14	0.01	22.85
	Level 4	1.05	0.05	11.40
6	Level 1			26.90
	Level 2	-0.66	0.02	32.09
	Level 3	-0.10	0.02	11.50
	Level 4	0.31	0.03	29.52
7	Level 1			16.49
	Level 2	-0.91	0.02	32.60
	Level 3	-0.25	0.04	17.87
	Level 4	0.17	0.03	33.04
8	Level 1			25.15
	Level 2	-0.66	0.01	23.12
	Level 3	-0.18	0.03	26.24
	Level 4	0.44	0.02	25.49
11	Level 1			19.29
	Level 2	-0.70	0.04	31.00
	Level 3	-0.19	0.04	25.28
	Level 4	0.44	0.01	24.43

ROUND 3 JUDGMENTS

The purpose of Round 3 was for panelists to discuss their Round 2 placements and, if necessary, to revise their ratings. Round 3 discussions were conducted at the whole-group level, rather than at table groups. Prior to the discussions, the panelists were presented with the median cuts of the whole group—in addition to table-level results—based on Round 2 results. During this round, the group was also presented with the impact data (i.e., the percentage of students classified into each performance level based on the group median cuts). The psychometrician presented the information to the group on chart paper and explained how to use it as they completed their Round 3 discussions. Sample tables and figures that were shown to the panelists after each of the rounds are displayed in Appendix 7-N.

The lead facilitator then held an extended discussion of the Round 2 results. The discussion walked the panelists through the ordered item booklet, focusing on the KSAs needed for each item and how they related to the performance level descriptors. In addition, the discussion explored the differences in where each panelist and table placed the cuts. Finally, after the discussions, panelists were given a final opportunity to revise their bookmark placements. Once again, the facilitator reminded the panelists that they should place the bookmarks according to their individual best judgment and that it was not necessary for them to reach a consensus.

TABULATION OF ROUND 3 RESULTS

When Round 3 ratings were complete, the Measured Progress staff members once again calculated the median cut scores for the room and the associated impact data. The results of the panelists' Round 3 ratings are outlined in Tables 7-15 and 7-16. Panelists were also shown the final results including median cuts and impact data. Sample tables and figures that were shown to the panelists after each of the rounds are displayed in Appendix 7-N.

Table 7-15. 2015 NCSC Standard Setting: Round 3 Results—ELA

Grade	*Performance Levels*	*Median Theta Cut*	*Median Absolute Deviation*	*Percent of Students*
3	Level 1			31.70
	Level 2	-0.57	0.01	18.50
	Level 3	-0.05	0.04	25.50
	Level 4	0.72	0.00	24.29
4	Level 1			34.26
	Level 2	-0.53	0.03	20.13
	Level 3	-0.01	0.04	35.60
	Level 4	1.43	0.05	10.01
5	Level 1			39.75
	Level 2	-0.51	0.04	7.66
	Level 3	-0.29	0.05	42.45
	Level 4	1.16	0.12	10.15
6	Level 1			33.00
	Level 2	-0.63	0.00	30.00
	Level 3	0.18	0.03	26.07
	Level 4	1.19	0.12	10.93
7	Level 1			32.21
	Level 2	-0.59	0.01	16.97

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
	Level 3	-0.18	0.01	35.64
	Level 4	0.95	0.06	15.17
8	Level 1			27.87
	Level 2	-0.75	0.00	28.25
8	Level 3	0.04	0.01	25.62
	Level 4	0.66	0.00	18.27
11	Level 1			27.96
	Level 2	-0.77	0.00	18.25
	Level 3	-0.37	0.01	25.88
	Level 4	0.52	0.00	27.91

Table 7-16. 2015 NCSC Standard Setting: Round 3 Results—Mathematics

Grade	Performance Levels	Median Theta Cut	Median Absolute Deviation	Percent of Students
3	Level 1			24.82
	Level 2	-0.65	0.00	15.97
	Level 3	-0.37	0.00	39.61
	Level 4	0.77	0.01	19.60
4	Level 1			32.09
	Level 2	-0.55	0.01	27.81
	Level 3	0.01	0.01	23.36
	Level 4	0.82	0.02	16.74
5	Level 1			22.14
	Level 2	-0.84	0.04	43.62
	Level 3	0.14	0.01	19.99
	Level 4	0.99	0.03	14.26
6	Level 1			30.38
	Level 2	-0.61	0.01	28.60
	Level 3	-0.10	0.02	11.50
	Level 4	0.31	0.04	29.52
7	Level 1			16.49
	Level 2	-0.91	0.00	32.60
	Level 3	-0.25	0.03	21.33
	Level 4	0.24	0.02	29.59
8	Level 1			25.15
	Level 2	-0.66	0.01	23.12
	Level 3	-0.18	0.01	26.24
	Level 4	0.44	0.02	25.49
11	Level 1			19.29
	Level 2	-0.70	0.04	31.00
	Level 3	-0.19	0.03	25.28
	Level 4	0.44	0.01	24.43

Figure 7-1. 2015 NCSC Standard Setting: Round 3 Results—ELA

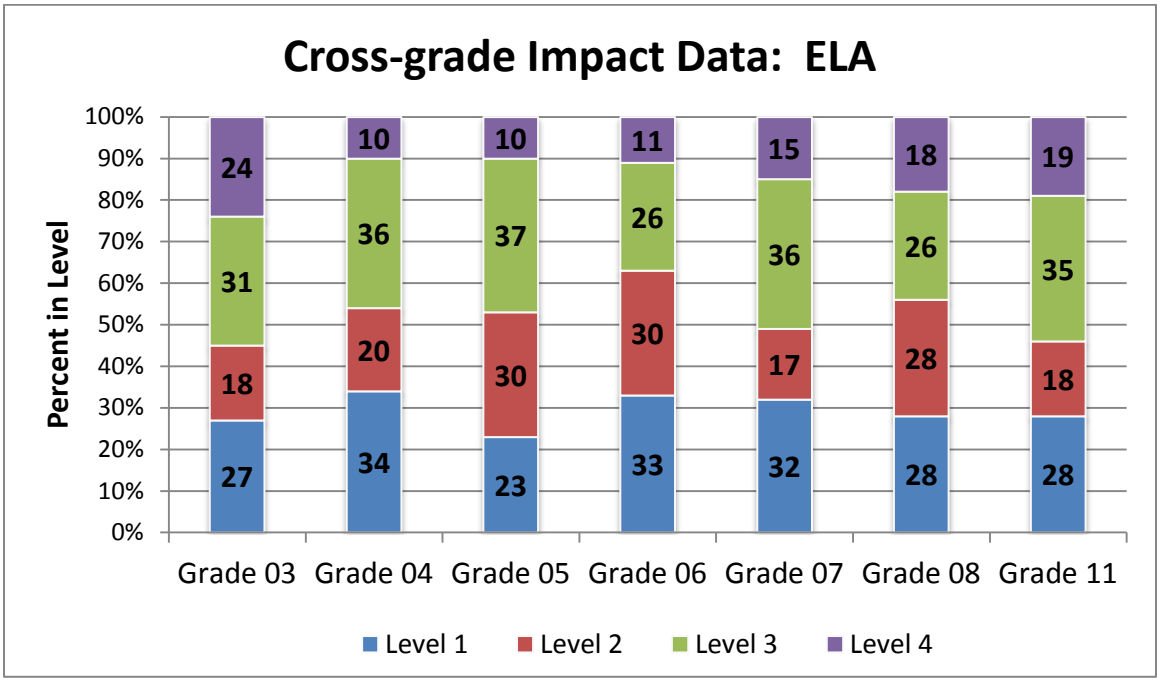
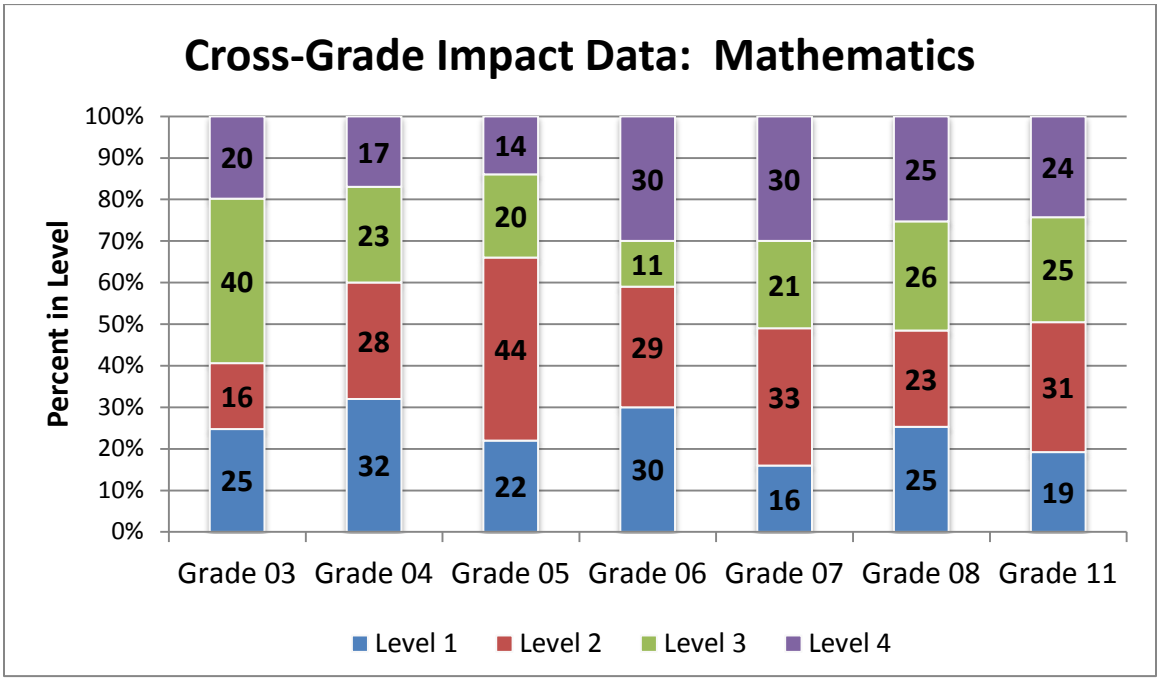


Figure 7-2. 2015 NCSC Standard Setting: Round 3 Results—Mathematics



RESULTS SUMMARY

It is important to summarize the pattern of judgments by round attending in particular to the deviation by round. When judgments are characterized by relative consistency in each round and deviation is reduced in each subsequent round, this lends credibility to the claim that panelists are converging on a shared judgment about the location of each standard. Tables 7-18 and 7-19 reveal that this desirable pattern is observed. Further, Figures 7-3 and 7-4 show that the mean absolute deviation is reduced for each performance level cut in each round. In fact, the deviation for most round 3 judgments is less than .05 for nearly all grades and levels.

Table 7-18. NCSC STANDARD SETTING: SUMMARY OF ELA RESULTS BY ROUND

		Round 1		Round 2		Round 3	
Grade	*Performance Levels*	*Median Theta Cut*	*Median Absolute Deviation*	*Median Theta Cut*	*Median Absolute Deviation*	*Median Theta Cut*	*Median Absolute Deviation*
3	Level 1						
	Level 2	-0.57	0.02	-0.57	0.01	-0.65	0
	Level 3	-0.06	0.12	-0.05	0.08	-0.37	0
	Level 4	0.72	0.18	0.72	0.06	0.77	0.06
4	Level 1						
	Level 2	-0.53	0.03	-0.53	0.02	-0.55	0.01
	Level 3	-0.01	0.05	-0.05	0.03	-0.06	0.05
	Level 4	1.33	0.06	1.33	0.05	0.82	0.02
5	Level 1						
	Level 2	-0.51	0.08	-0.51	0.06	-0.84	0.08
	Level 3	-0.17	0.16	-0.29	0.06	0.14	0.01
	Level 4	1.27	0.13	1.03	0.16	1.05	0.05
6	Level 1						
	Level 2	-0.63	0.07	-0.63	0.02	-0.66	0.02
	Level 3	0.38	0.07	0.24	0.05	-0.1	0.02
	Level 4	1.59	0.17	1.19	0.16	0.31	0.03
7	Level 1						
	Level 2	-0.59	0.02	-0.59	0.01	-0.91	0.02
	Level 3	-0.24	0.03	-0.17	0.01	-0.25	0.04
	Level 4	0.39	0.1	0.59	0.09	0.17	0.03
8	Level 1						
	Level 2	-0.75	0.04	-0.75	0	-0.66	0.01
	Level 3	0.04	0.05	0.17	0.02	-0.18	0.03
	Level 4	0.68	0.09	0.67	0.03	0.44	0.02
11	Level 1						
	Level 2	-0.83	0.04	-0.77	0.01	-0.7	0.04
	Level 3	-0.51	0.05	-0.37	0.02	-0.19	0.04
	Level 4	-0.02	0.13	0.52	0.08	0.44	0.01

Table 7-19. NCSC STANDARD SETTING: SUMMARY OF ELA RESULTS BY ROUND

		Round 1		Round 2		Round 3	
Grade	*Performance Levels*	*Median Theta Cut*	*Median Absolute Deviation*	*Median Theta Cut*	*Median Absolute Deviation*	*Median Theta Cut*	*Median Absolute Deviation*
3	Level 1						
	Level 2	-0.65	0.08	-0.57	0.01	-0.65	0
	Level 3	-0.37	0.03	-0.05	0.08	-0.37	0
	Level 4	0.28	0.11	0.72	0.06	0.77	0.06
4	Level 1						
	Level 2	-0.59	0.05	-0.53	0.02	-0.55	0.01
	Level 3	-0.06	0.07	-0.05	0.03	-0.06	0.05
	Level 4	0.82	0.15	1.33	0.05	0.82	0.02
5	Level 1						
	Level 2	-0.61	0.09	-0.51	0.06	-0.84	0.08
	Level 3	0.14	0.06	-0.29	0.06	0.14	0.01
	Level 4	1.05	0.05	1.03	0.16	1.05	0.05
6	Level 1						
	Level 2	-0.61	0.04	-0.63	0.02	-0.66	0.02
	Level 3	-0.1	0.03	0.24	0.05	-0.1	0.02
	Level 4	0.34	0.03	1.19	0.16	0.31	0.03
7	Level 1						
	Level 2	-0.93	0.08	-0.59	0.01	-0.91	0.02
	Level 3	-0.25	0.11	-0.17	0.01	-0.25	0.04
	Level 4	0.2	0.13	0.59	0.09	0.17	0.03
8	Level 1						
	Level 2	-0.66	0.07	-0.75	0	-0.66	0.01
	Level 3	-0.11	0.05	0.17	0.02	-0.18	0.03
	Level 4	0.44	0.06	0.67	0.03	0.44	0.02
11	Level 1						
	Level 2	-0.7	0.04	-0.77	0.01	-0.7	0.04
	Level 3	-0.03	0.03	-0.37	0.02	-0.19	0.04
	Level 4	0.44	0.04	0.52	0.08	0.44	0.01

Figure 7-3. 2015 NCSC Standard Setting: ELA Deviation by Grade, Performance Level, and Round

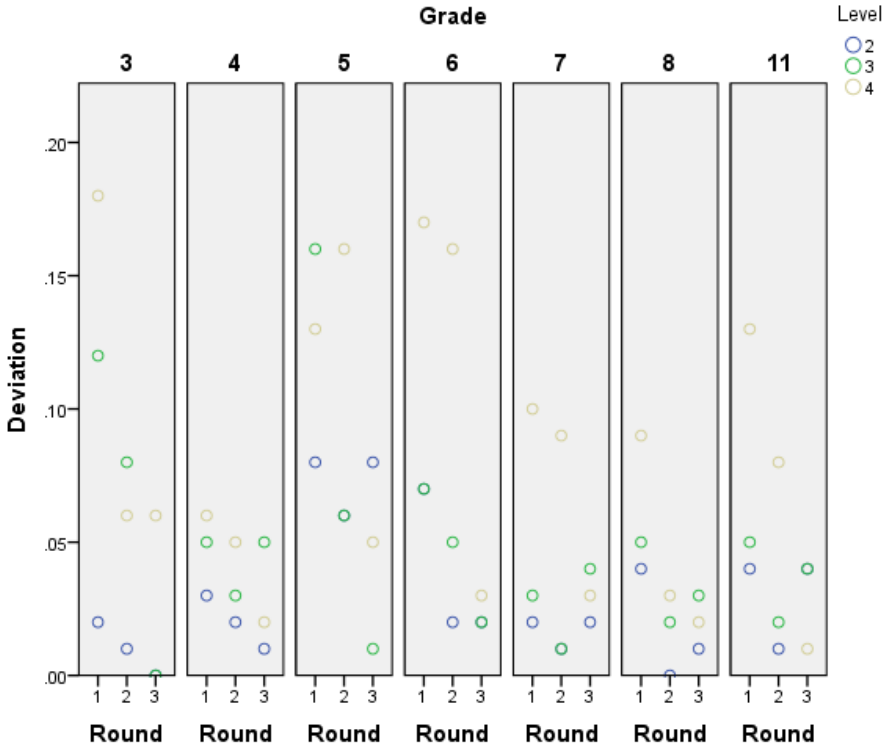
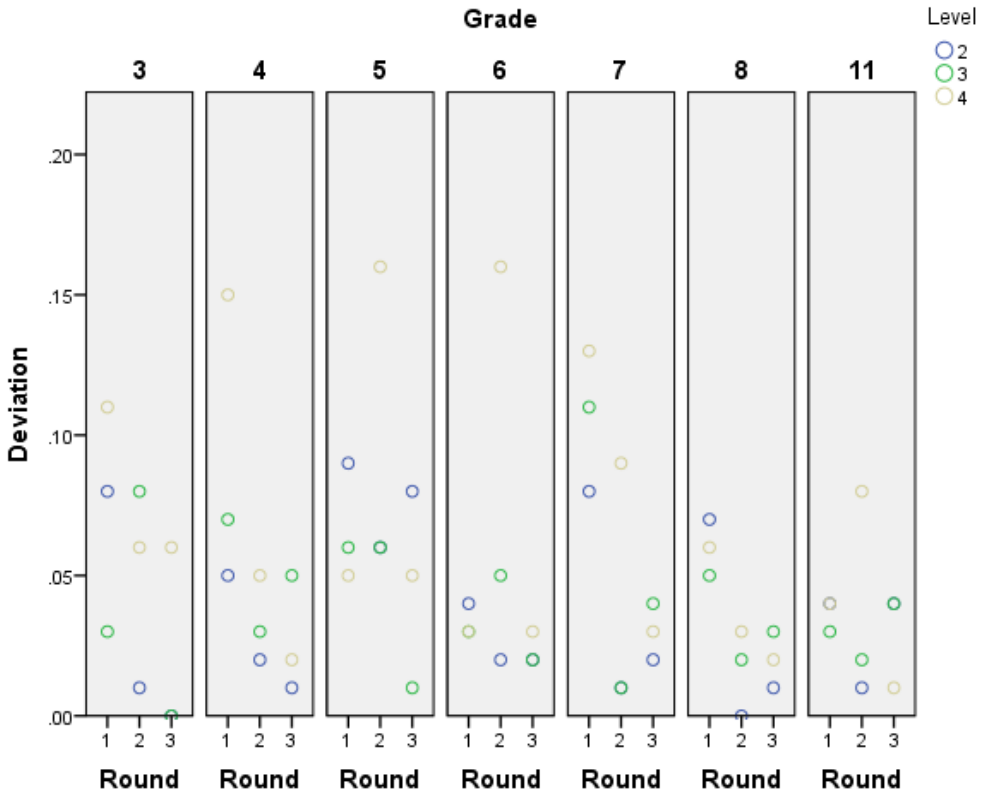


Figure 7-4. 2015 NCSC Standard Setting: Mathematics Deviation by Grade, Performance Level, and Round



EVALUATION

The measurement literature sometimes considers the evaluation process to be another product of the standard setting process (e.g., Reckase, 2001), as it provides important validity evidence supporting the cut scores that are obtained. To provide evidence of the participants' views of the standard setting process, panelists were asked to complete an evaluation about the general session presentations, the practice round, and also about the standard setting process itself. These evaluations were separated into a procedural evaluation that was completed after setting standards for the first grade for each panel (i.e., after grades 3, 5, 7, and 11) and a final evaluation completed at the end of the meeting (i.e. after setting standards for grades 4, 6, 8, and 11). Summary procedural evaluation results collapsed across grade and content are presented in Tables 7-20 and 7-21, and summary final evaluation results are displayed in Table 7-22.

Table 7-20. 2015 NCSC Standard Setting: Procedural Evaluation Results Summary

Please rate the usefulness of each of the following:	*N*	*Mean*	*% SD*	*% D*	*% A*	*% SA*
I understood how to make the cut score judgments.	80	3.65	0.00%	0.00%	35.00%	65.00%
I understood how to use the materials provided.	80	3.79	0.00%	0.00%	21.25%	78.75%

Please rate the usefulness of each of the following:	*N*	*Mean*	*% SD*	*% D*	*% A*	*% SA*
I understood how to record my judgments.	80	3.78	0.00%	0.00%	22.50%	77.50%
I think the procedures make sense.	79	3.61	0.00%	1.27%	36.71%	62.02%
I am sufficiently familiar with the assessment.	80	3.68	0.00%	0.00%	32.50%	67.50%
I understand the differences between the performance levels.	80	3.59	0.00%	0.00%	41.25%	58.75%

Table 7-21. 2015 NCSC Standard Setting: Procedural Evaluation Results Summary

Do you believe the final recommended cut score for each of the performance levels is too low, about right, or too high?	*N*	*Mean*	*Too Low*	*Somewh at Low*	*About Right*	*Somewh at High*	*Too High*
Level 4/Level 3	77	3.01	2.60%	6.49%	77.92%	12.99%	0.00%
Level 3/Level 2	77	2.97	2.60%	5.19%	84.42%	7.79%	0.00%
Level 2/Level 1	77	3.06	1.30%	1.30%	88.31%	7.79%	1.30%

Table 7-22. 2015 NCSC Standard Setting: Final Evaluation Results Summary

Please mark the appropriate box for each statement.	*N*	*Mean*	*% SD*	*% D*	*% A*	*% SA*
I understood the goals of the standard setting meeting.	85	3.60	0.00%	0.00%	40.00%	60.00%
I understood the procedures we used to set standards.	85	3.65	0.00%	1.18%	32.94%	65.88%
The facilitator helped me understand the process.	85	3.57	1.18%	3.53%	32.94%	62.35%
The materials contained the information needed to set standards.	84	3.36	0.00%	7.14%	50.00%	42.86%
I understood how to use the materials provided.	85	3.44	0.00%	3.53%	48.24%	48.24%
The borderline performance level definitions were clear.	85	3.16	0.00%	8.23%	67.06%	24.70%
I understood how to make the cut score judgments.	85	3.60	0.00%	0.00%	40.00%	60.00%
I understood how to use	85	3.68	0.00%	0.00%	31.76%	68.24%

Please mark the appropriate box for each statement.	*N*	*Mean*	*% SD*	*% D*	*% A*	*% SA*
the feedback provided after each round.						
I understood how to use the impact data.	85	3.52	1.18%	2.35%	40.00%	56.47%
I understood how the cut scores were calculated.	85	3.38	0.00%	8.23%	45.88%	45.88%
The facilitator was able to get answers to my questions.	85	3.35	1.18%	3.53%	54.12%	41.18%
Sufficient time was allotted for training on the standard setting tasks.	85	3.57	0.00%	4.71%	34.12%	61.18%
Sufficient time was allotted to complete the standard setting tasks.	85	3.68	0.00%	2.35%	27.06%	70.59%
The facilitator helped the standard setting process run smoothly.	85	3.57	0.00%	5.88%	31.76%	62.35%
Overall the standard setting process produced credible results.	85	3.32	0.00%	11.76 %	44.71%	43.53%

The full results of the procedural and final evaluations broken down by grade and content area are presented in Appendix 7-K.

Upon completion of the evaluation forms, panelists' responses were reviewed by the psychometricians. This review did not indicate any reason that a particular panelist's data should not be included when the final cutpoints were calculated. In general, participants felt that the recommended cutpoints were appropriate and that their judgments were based on appropriate information and decision making (see Appendix 7-K).

However, the final evaluation results (Appendix 7-K) did show problematic ratings for some of the statements for the ELA Grade 3-4 group:

- 36% disagreed or strongly disagreed with the statement, "The facilitator helped me understand the process"
- 36% disagreed with the statement, "Sufficient time was allotted to training on standard setting tasks"
- 45% disagreed with the statement, "The facilitator helped the standard setting process run smoothly"

However, the results from the ELA Group 3-4 show levels of panelist consistency (as measured by Mean Absolute Deviation) that did not look out of line with those from other grade level groups. It might be noted that the cross-grade articulation panel (described in Section 4.1) made two adjustments for the Grade 3 results in ELA (Level 2 and Level 3 cut scores), lowering the theta-cuts slightly for each.

TASKS COMPLETED AFTER THE STANDARD SETTING MEETING

Upon conclusion of the standard setting meeting, several important tasks were completed. These tasks centered on the following: convening a cross-grade articulation committee to review the cut scores for all grades and content areas; reviewing the standard setting process and addressing issues presented by the outcomes; presenting the results to NCSC; making any final revisions or adjustments based on policy considerations under the direction of the NCSC states; and preparing the standard setting report.

CROSS-GRADE ARTICULATION COMMITTEE MEETING

Upon completion of the standard setting process, a cross-grade articulation committee was convened. Two panelists from each of the English language arts (ELA) and mathematics panels were asked to be a part of this meeting. The ELA and Mathematics panelists met in two separate meetings. Panelists were given an overview of the process, which involved: (1) reviewing the impact data that result from the Round 3 ratings, (2) completing a rating form to indicate if they think each cut score is too high, about right, or two low, and (3) discussing any concerns or observations they have about the data (a sample of the evaluation forms is included in Appendix G). The discussions started with the Level 2/Level 3 cut for the lowest grade, followed by discussions of the Level 2/Level 3 cut for the next grade. If the panelists were uncomfortable with a particular cut score, and wanted to investigate it further, they were presented with the ordered item booklet, performance level descriptors, borderline performance level descriptors and the location of the bookmark for the grade of interest. Their task was to review the content of the items that surround the bookmark and make a recommendation for a revised placement. Once the group made a recommendation, the impact data results were updated and shared with the group for further discussion. This process continued until the committee discussed all cut scores that were of concern.

For both subject areas a few of the cuts were decided to be inconsistent with other grades' bookmark placement reasoning and adjustments were made. A summary of these bookmark adjustments are displayed in Tables 7-23 and 7-24. The post-articulation cross-grade impact data are shown in Figures 7-5 and 7.6.

Table 7-23. 2015 NCSC Standard Setting: Summary of Articulation Changes—ELA

Grade	*Performance Levels*	*Round 3 Median Theta Cut*	*Post-Articulation Median Theta Cut*
3	Level 2	-0.57	-0.73
3	Level 3	-0.05	-0.18
5	Level 2	-0.51	-0.84
8	Level 4	0.66	0.78
11	Level 4	0.52	0.90

Figure 7-5. 2015 NCSC Standard Setting: Post-Articulation Results—ELA

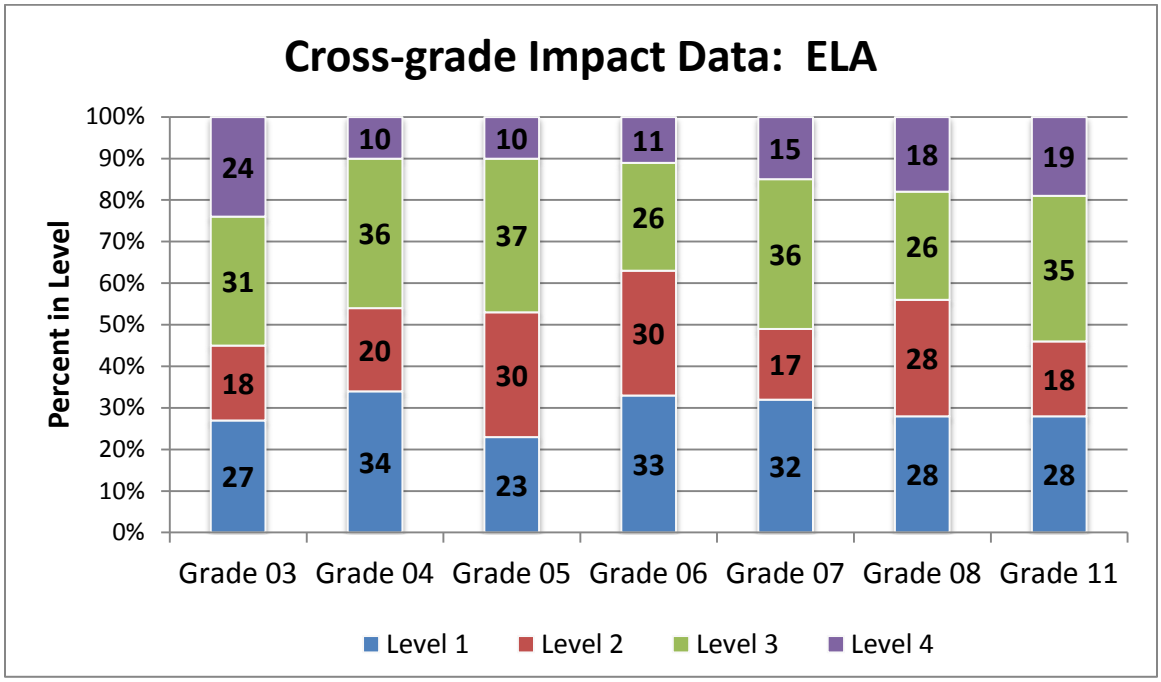
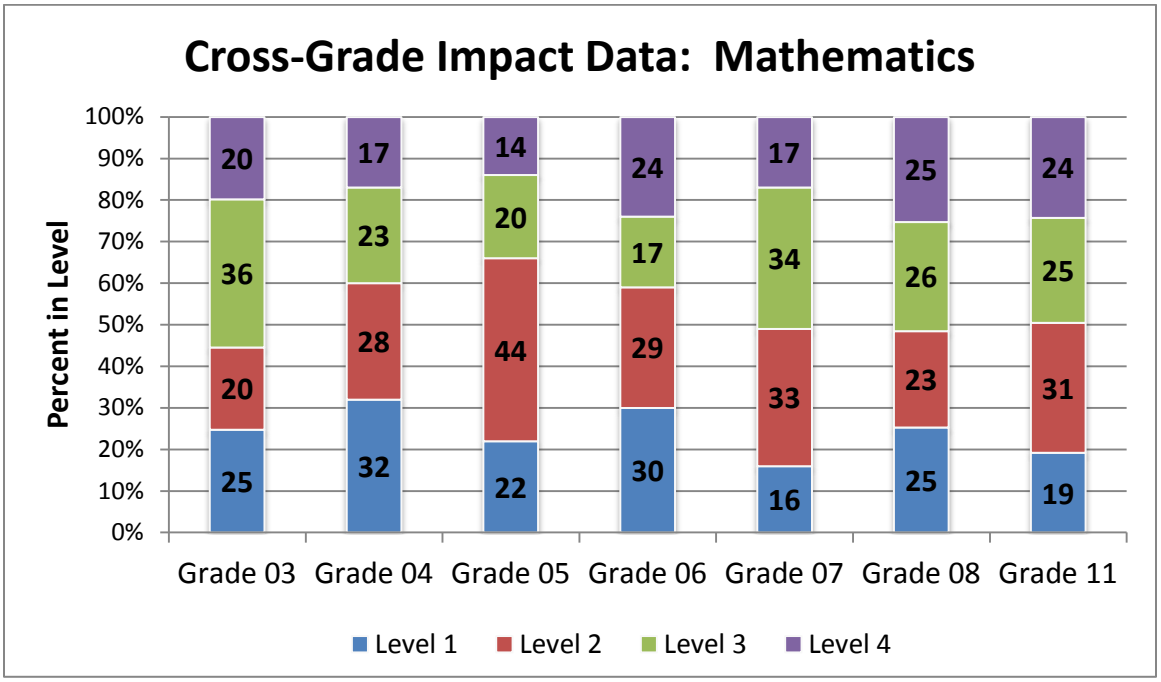


Table 7-24. 2015 NCSC Standard Setting: Summary of Articulation Changes—Mathematics

Grade	*Performance Levels*	*Round 3 Median Theta Cut*	*Post-Articulation Median Theta Cut*
3	Level 3	-0.37	-0.28
6	Level 4	0.31	0.53
7	Level 4	0.24	0.77

Figure 7-6. 2015 NCSC Standard Setting: Post-Articulation Results—Mathematics



POLICY ADJUSTMENTS

At the completion of the standard setting activities, including review of the cut scores by a cross-grade articulation panel, representatives from the partner states evaluated the recommended cut scores and related data in light of the Performance Level Descriptors, Borderline Descriptors, and Ordered Item Books. The states also included the results and judgments of the cross-grade articulation meeting as part of their decision making process.

The Partner States revised the cut scores in two areas:

- Mathematics, grade 5—the cut score between Level 2 and Level 3 was lowered from Item 19 to Item 17.
- English Language Arts, grade 5—the cut score between Level 2 and Level 3 was raised from Item 12 to Item 14.

Final acceptance of the cut scores was determined by a majority vote of participating Partner States. Final cut scores and thetas are as follows:

Table 7-25. 2015 NCSC Standard Setting: Final Cuts—ELA

Grade	*Performance Levels*	*Theta Cut*	*Percent of Students*
3	Level 1		26.56
	Level 2	-0.70	17.99
	Level 3	-0.18	31.15
	Level 4	0.72	24.29
4	Level 1		34.26

Grade	Performance Levels	Theta Cut	Percent of Students
	Level 2	-0.53	20.13
	Level 3	-0.01	35.60
	Level 4	1.43	10.01
5	Level 1		23.23
	Level 2	-0.84	29.95
	Level 3	-0.13	36.67
	Level 4	1.16	10.15
6	Level 1		33.00
	Level 2	-0.63	30.00
	Level 3	0.18	26.07
	Level 4	1.19	10.93
7	Level 1		32.21
	Level 2	-0.59	16.97
	Level 3	-0.20	35.64
	Level 4	0.95	15.17
8	Level 1		27.87
	Level 2	-0.75	28.25
	Level 3	0.04	25.62
	Level 4	0.78	18.27
11	Level 1		27.96
	Level 2	-0.77	18.25
	Level 3	-0.37	34.55
	Level 4	0.90	19.24

Table 7-26. NCSC Standard Setting: Final Cuts—Mathematics

Grade	Performance Levels	Theta Cut	Percent of Students
3	Level 1		24.82
	Level 2	-0.65	19.60
	Level 3	-0.28	35.98
	Level 4	0.77	19.60
4	Level 1		32.09
	Level 2	-0.55	27.81
	Level 3	0.01	23.36
	Level 4	0.82	16.74
5	Level 1		22.14
	Level 2	-0.84	31.37
	Level 3	-0.11	32.24
	Level 4	0.99	14.26
6	Level 1		30.38
	Level 2	-0.61	28.60
	Level 3	-0.10	17.41
	Level 4	0.53	23.60
7	Level 1		16.49
	Level 2	-0.91	32.60
	Level 3	-0.25	33.94

Grade	Performance Levels	Theta Cut	Percent of Students
	Level 4	0.77	16.98
8	Level 1		25.15
	Level 2	-0.66	23.12
	Level 3	-0.18	26.24
	Level 4	0.44	25.49
11	Level 1		19.29
	Level 2	-0.70	31.00
	Level 3	-0.19	25.28
	Level 4	0.44	24.43

Figure 7-7. 2015 NCSC Standard Setting: Post-Policy Adjustment Results—ELA

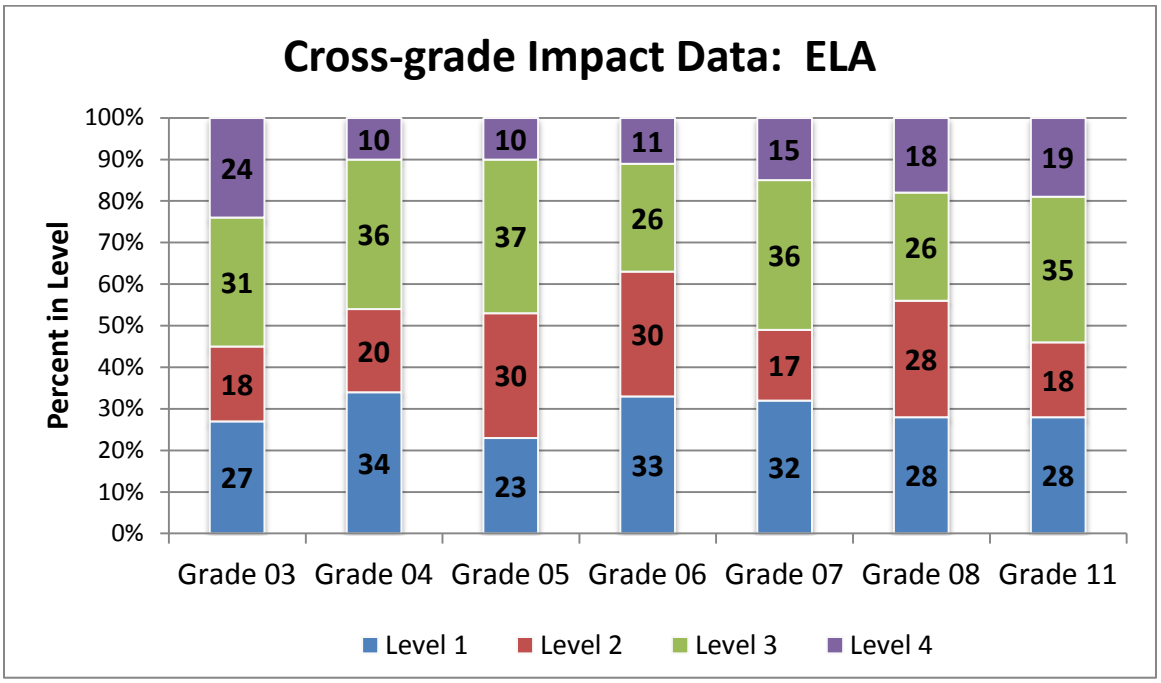
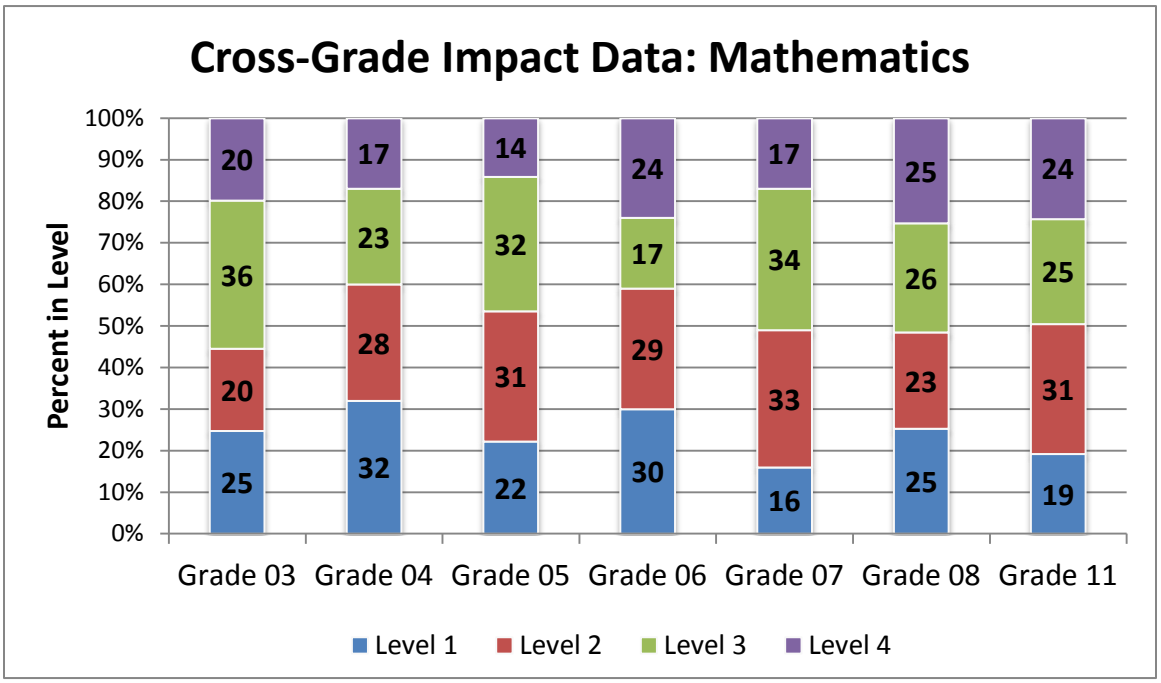


Figure 7-8. 2015 NCSC Standard Setting: Post-Policy Adjustment Results—Mathematics



Preparation of Standard Setting Report

Following the final compilation of the standard setting results, Measured Progress prepared this report, which documents the procedures and results of the 2015 standard setting meeting held in order to establish performance standards for the NCSC assessment in ELA and mathematics.

External Evaluation of Standard Setting

Dr. Barbara Plake, Professor Emeritus of the University of Nebraska-Lincoln, served as the external evaluator for the standard setting activities. She reviewed materials and procedures prior to the meeting, observed the entire meeting, and wrote a report on those observations called *Synopsis of Validity Evidence for the Cutscores Derived from the Grades 3 – 8 and 11 Standard Setting for NCSC ELA and Mathematics Assessments* (see Appendix 7-O). She also reviewed the test administration vendor's Standard Setting Report, and produced a summary report called *Review of the Standard Setting Report Prepared by Measured Progress On the Mathematics and ELA Standard Setting Process for Grades 3 – 8 and 11* (see Appendix 7-P).

Using these two sources of validity evidence for the cutscores derived from the Grades 3 – 8 and 11 standard setting on NCSC's Assessments of Mathematics and English Language Arts, she concluded in her final validity memo, *Report on Evidence for Validity for the Cutscores Derived from the Grades 3 – 8 and 11 Standard Setting* (see Appendix 7-Q):

Based on the validity evidence provided in these documents, the following evidence supports the validity of the interpretations and uses of the cutscores:

- The panelists selected to participate in the standard setting process matched the targeted characteristics for panel composition (representing the partner states, appropriate range of educational experiences, experience with special needs students)
- Facilitators were trained to facilitate the standard setting panels
- Panelists were adequately trained in the process (they took the test, familiarized themselves with the PLDs, completed a item map to identify the competencies measured by the test questions, developed Borderline Performance Level Descriptors (BPLDs), completed a practice task and indicated that they felt ready to make their item ratings)
- Panelists indicated that they understood the feedback presented between rounds of ratings
- Panelists indicated that they considered the BPLDs in making their item ratings
- Panelists indicated that had sufficient time to make their item ratings
- Panelists indicated they felt confident in making their item ratings
- Panelists Mean Absolute Deviations showed cohesion in their item ratings
- Panelists Mean Absolute Deviations narrowed over rounds of ratings
- Cross-grade articulations were based on a "reasonableness" criterion, indicating that external criteria were used in making these adjustments
- Policy adjustments were based on a "reasonableness" criterion, indicating that external criteria were used in making any adjustments

Together these sources of validity evidence provide a strong case for supporting the overall validity for the interpretation and use of the cutscores derived from the standard setting for NCSC's assessments in Mathematics and ELA for Grades 3 – 8 and 11. (Plake, November 2015, see full report in Appendix P.)

CHAPTER 8: STUDIES OF RELIABILITY AND CONSTRUCT-RELATED VALIDITY

STATISTICAL RELIABILITY ANALYSES

Although an individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together. Tests that function well provide a dependable assessment of the student's level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student's score being either higher or lower than his or her true ability. For example, a student may misread an item, or mistakenly fill in the wrong bubble when he or she knew the answer. Collectively, extraneous factors that affect a student's score are referred to as "measurement error." Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores, or vice versa. Consequently, one cannot reliably measure a student's true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average; and students' scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment's reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, the extraneous factors affecting performance are small and the test is reliable. (This is referred to as "test-retest reliability.") A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the remembering items problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly, the test is considered reliable. (This is known as "alternate forms reliability" because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address the latter two problems is to split the test in half and then correlate students' scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval and with creating and administering two parallel forms of the test are alleviated. This is known as a "split-half estimate of reliability." If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test into halves will result in a different correlation. Another problem with the split-half method of calculating reliability is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, α (alpha), that

eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach's α was used to assess the reliability of the 2014–15 NCSC tests:

$$\alpha \equiv \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma^2_{(Y_i)}}{\sigma^2_x}\right]$$

where
i indexes the item,
n is the total number of items,
$\sigma^2_{(Y_i)}$ represents individual item variance, and
σ^2_x represents the total test variance.

Tables 8-1 and 8-2 present descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement for ELA and mathematics by grade and form. (Statistics are based on core items only, which were those that counted toward students' reported scores.)

Table 8-1. 2014–15 NCSC: Reliability by Form—ELA

Grade	*Form*	*Number of Students*	*Raw Score*			*Alpha*	*SEM*
			Maximum	*Mean*	*Standard Deviation*		
3	1	1,029	30	18.54	5.85	0.85	2.29
	2	960	29	18.76	5.61	0.85	2.19
	3	1,024	30	18.55	5.96	0.86	2.24
	4	958	30	19.46	6.09	0.87	2.22
4	1	1,052	31	18.73	6.28	0.86	2.39
	2	1,011	31	18.98	6.09	0.85	2.40
	3	1,058	31	19.05	6.58	0.87	2.34
	4	1,055	31	18.80	6.69	0.88	2.34
5	1	1,113	30	17.53	5.63	0.82	2.37
	2	1,102	30	17.79	5.35	0.80	2.39
	3	1,031	32	18.83	6.17	0.84	2.46
	4	1,005	29	16.94	5.71	0.83	2.32

Cont.

Grade	Form	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
6	1	1,074	30	18.90	5.98	0.85	2.33
	2	1,119	29	18.45	5.16	0.81	2.25
	3	1,042	30	18.91	6.10	0.86	2.30
	4	1,071	29	18.94	5.68	0.85	2.20
7	1	1,072	29	18.56	5.62	0.84	2.25
	2	1,071	32	19.54	6.19	0.85	2.39
	3	1,028	31	19.43	6.00	0.84	2.37
	4	1,128	29	17.86	5.30	0.81	2.31
8	1	1,180	31	19.97	5.72	0.83	2.34
	2	1,056	32	20.22	6.04	0.84	2.43
	3	1,163	31	19.28	5.60	0.82	2.39
	4	1,087	32	19.64	5.95	0.84	2.41
11	1	1,019	28	19.36	5.40	0.85	2.08
	2	921	29	19.42	5.58	0.85	2.17
	3	1,097	29	19.44	5.46	0.84	2.16
	4	981	31	20.58	6.57	0.89	2.23

Table 8-2. Reliability by Form—Mathematics

Grade	Form	Number of Students	Raw Score			Alpha	SEM
			Maximum	Mean	Standard Deviation		
3	1	1,045	35	18.37	6.63	0.84	2.68
	2	960	35	18.55	7.12	0.87	2.62
	3	1,021	35	18.26	7.13	0.86	2.64
	4	964	35	18.23	7.15	0.86	2.64
4	1	1,056	33	15.81	5.83	0.80	2.60
	2	1,010	33	15.78	5.51	0.78	2.60
	3	1,061	33	15.90	5.91	0.81	2.59
	4	1,064	33	16.10	5.36	0.76	2.61
5	1	1,123	35	16.50	5.64	0.77	2.71
	2	1,108	35	16.68	5.45	0.75	2.70
	3	1,038	34	15.66	5.46	0.76	2.67
	4	1,008	33	15.66	5.62	0.78	2.61
6	1	1,074	35	18.76	6.41	0.83	2.67
	2	1,124	35	18.81	6.22	0.81	2.68
	3	1,041	34	17.52	6.66	0.84	2.64
	4	1,085	35	18.63	6.77	0.85	2.65
7	1	1,077	34	17.78	6.23	0.82	2.62
	2	1,067	35	18.81	6.87	0.85	2.62
	3	1,026	34	17.77	6.61	0.85	2.60
	4	1,139	34	17.70	6.29	0.83	2.63
8	1	1,179	35	17.56	6.50	0.83	2.71
	2	1,064	34	17.36	6.57	0.84	2.66

	3	1,163	35	17.75	6.69	0.84	2.69
	4	1,094	35	16.62	6.43	0.82	2.70
11	1	962	34	16.12	6.72	0.85	2.61
	2	848	35	16.98	6.59	0.83	2.69
	3	1,034	34	16.18	5.97	0.80	2.67
	4	916	34	16.05	6.61	0.84	2.65

All of the reliability coefficients fell within acceptable ranges for large scale summative assessments. An alpha coefficient that is .7 or larger is acceptable according to Kline (2000). Because different grades have different test designs, it is inappropriate to make inferences about the quality of one test by comparing its reliability to that of another test from a different grade.

Subgroup Reliability

The reliability coefficients discussed in the previous section were based on the overall population of students who took the 2014–15 NCSC. Appendix 8-A presents reliabilities for various subgroups of interest (e.g., gender, ethnicity, Learner Characteristics Inventory (LCI) categories). Subgroup Cronbach's α's were calculated using the formula defined above, based only on the members in the subgroup of interest in the computations; values were only calculated for subgroups with 100 or more students. For several reasons, the results of this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test but also on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix 8-A that subgroup sample sizes varied considerably, which resulted in natural variation in reliability coefficients. Additionally, α, which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability in performance (Draper & Smith, 1998).

Reliability of Performance Level Categorization

While related to reliability, the accuracy and consistency of classifying students into performance categories are even more important statistics in a standards-based reporting framework (Livingston & Lewis, 1995). After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For NCSC, students were classified into one of four performance levels: Level 1, Level 2, Level 3, and Level 4. This section of the report explains the methodology used to assess the reliability of classification decisions; results are provided. Livingston and Lewis (1995) method was used, with the technical adjustments described in the same article.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist. Consistency measures the extent to which classification decisions based

on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2014–15 NCSC AA-AAS because it is easily adaptable to all types of testing formats, including mixed format tests.

The accuracy and consistency estimates reported in Appendix 8-B make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their "true" classifications.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments, per Livingston and Lewis (1995), a new four-by-four contingency table was created for ELA and mathematics by grade and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell [i, j] of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where i = 1 to 4) and whose observed score on the second form would fall into classification j (where j = 1 to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.} C_{.i}}{1 - \sum_i C_{i.} C_{.i}},$$

where

$C_{i.}$ is the proportion of students whose observed performance level would be Level i (where i = 1–4) on the first hypothetical parallel form of the test;

$C_{.i}$ is the proportion of students whose observed performance level would be Level i (where i = 1–4) on the second hypothetical parallel form of the test;

C_{ii} is the proportion of students whose observed performance level would be Level i (where i = 1–4) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

Decision Accuracy and Consistency Results

The decision accuracy and consistency analyses described above are provided in Table 8-B-1 of Appendix 8-B. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon performance level are also given. For these calculations, the denominator is the proportion of students associated with a given performance level. For example, the

conditional accuracy value is 0.85 for Level 1 for ELA grade 4. This table indicates that among the students whose true scores placed them in this classification, 85% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.81 indicates that 81% of students with observed scores in Level 1 would be expected to score in this classification again if a second, parallel test form were used. The relatively lower accuracy and consistency values conditional upon performance levels, in particular, the ones for Level 2, result from the relatively small number of students that fall into this category and the narrower score ranges of the performance level. However, it is also a clear indication of where test development should focus to improve the quality of the tests, even though the overall indexes are satisfactory.

For some testing situations, the greatest concern may be decisions close to level thresholds. For example, in testing done for No Child Left Behind accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. For the 2014–15 NCSC AA-AAS, Table 8B-2 in Appendix 8-B provides accuracy and consistency estimates at each cutpoint, as well as false positive and false negative decision rates. A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, decision accuracy and consistency statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Appendix 8-B should be interpreted with caution. In addition, it is important to remember that it is inappropriate to compare decision accuracy and consistency statistics between grades and content areas.

STATISTICAL VALIDITY ANALYSES

DIF

The *Code of Fair Testing Practices in Education* (2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are due to construct-relevant, rather than irrelevant, factors. Chapter 3 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME,

2014) includes similar guidelines. As part of the effort to identify such problems, NCSC items were evaluated in terms of differential item functioning (DIF) statistics.

For the NCSC AA-AAS, the standardization DIF procedure (Dorans & Kulick, 1986) was employed to evaluate subgroup differences. The standardization DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. The DIF procedure calculates the difference in item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two groups.

When differential performance between two groups occurs on an item (i.e., a DIF index in the "low" or "high" categories, explained below), it may or may not be indicative of item bias. Course-taking patterns or differences in school curricula can lead to DIF but for construct-relevant reasons. On the other hand, if subgroup differences in performance can be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items should be reconsidered.

For the 2014–15 NCSC, six subgroup comparisons were evaluated for DIF:

- Male vs. female
- White vs. Black
- White vs. Hispanic
- White vs. American Indian
- Not low socioeconomic status (SES) vs. low SES
- Not LEP vs. LEP (including current, exited 1 year, and exited 2 year)

The tables in Appendix 8-C present the number of items classified as either "low" or "high" DIF, overall and by group favored. Computed DIF indices have a theoretical range from –1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for polytomous items. Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible. The preponderance of NCSC items fell within this range (see Tables 8C-1 and 8C-2 in Appendix 8-C). Dorans and Holland further stated that items with values between –0.10 and –0.05 and between 0.05 and 0.10 (i.e., "low" DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the –0.10 to 0.10 range (i.e., "high" DIF) are more unusual and should be examined very carefully.

Table 8-3 summarizes the DIF results across all grades and content areas, showing the number of items classified as "High" DIF using the criterion described.

Table 8-3. 2014–15 NCSC: Number of Items Classified as High DIF

	ELA		Mathematics	
	Total DIF Tests	Total High DIF Items	Total DIF Tests	Total High DIF Items
Grade 3	360	2	376	7
Grade 4	348	3	390	11
Grade 5	358	2	298	3
Grade 6	356	3	348	5
Grade 7	358	3	370	3
Grade 8	356	0	390	1
Grade 11	358	8	276	1

Dimensionality Analyses

Because tests are constructed with multiple content area subcategories, and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional IRT models that are used for calibrating, linking, scaling, and equating the 2014-15 NCSC test forms.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2014-15 NCSC AA-AAS core items for ELA and mathematics are in this section. (Only core items used for score reporting were included in these analyses.)

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected total test scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and local dependence implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first divided into a training sample and a cross-validation sample. Then an exploratory analysis of

the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first divided into a training sample and a cross-validation sample. The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances: within-cluster conditional covariances are summed; from this sum the between-cluster conditional covariances are subtracted; this difference is divided by the total number of item pairs; and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality), values of 0.2 to 0.4 weak to moderate multidimensionality; values of 0.4 to 1.0 moderate to strong multidimensionality, and values greater than 1.0 very strong multidimensionality (Roussos & Ozbek, 2006).

DIMTEST and DETECT were applied to each form on the 2014-15 NCSC ELA and mathematics tests. The data for each form were split into a training sample and a cross-validation sample. Every form had at least 847 student examinees, so every training sample and cross-validation sample had at least 423 students. DIMTEST was then applied to every form. DETECT was applied to each dataset for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

Even though the sample sizes were not large for the NCSC test forms, the DIMTEST null hypothesis was rejected at a significance level of 0.01 for every dataset, suggesting that the violations of local independence were sizeable. Thus, it was important to use DETECT to estimate the effect size of the violations of local independence found by DIMTEST. Table 8-4 displays the multidimensional effect size estimates from DETECT.

Table 8-4. 2014–15 NCSC: Multidimensionality Effect Sizes by Form, Content Area, and Grade

Content Area	*Grade*	*Form*	*Multidimensionality Effect Size*
ELA	3	1	1.58
		2	0.98
		3	1.03
		4	1.03
		Average	1.16
	4	1	1.33
		2	0.87
		3	1.15
		4	1.20
		Average	1.14

Content Area	*Grade*	*Form*	*Multidimensionality Effect Size*
	5	1	0.86
		2	0.83
		3	0.79
		4	0.66
		Average	0.79
	6	1	0.80
		2	0.66
		3	0.83
		4	0.46
		Average	0.69
	7	1	0.89
		2	0.84
		3	0.71
		4	0.94
		Average	0.85
	8	1	0.95
		2	1.16
		3	0.96
		4	0.79
		Average	0.97
	11	1	1.08
		2	0.79
		3	0.53
		4	0.68
		Average	0.77
Total Average			0.91
Mathematics	3	1	1.36
		2	0.86
		3	0.96
		4	0.76
		Average	0.99
	4	1	1.19
		2	1.13
		3	1.02
		4	1.31
		Average	1.16
	5	1	1.64
		2	1.05
		3	1.01
		4	0.67
		Average	1.09
	6	1	0.88
		2	1.02
		3	1.06
		4	0.99
		Average	0.99
	7	1	1.01

Content Area	*Grade*	*Form*	*Multidimensionality Effect Size*
		2	1.13
		3	0.83
		4	1.21
		Average	1.05
Mathematics	8	1	0.66
		2	0.91
		3	0.83
		4	0.94
		Average	0.84
	11	1	0.67
		2	0.40
		3	0.66
		4	0.62
		Average	0.59
Total Average			0.96

All the DETECT values for 2014-15 indicated moderate to strong and very strong multidimensionality. Given the unusually large DETECT indices, it was important to identify the source(s) of the violations of local independence. Hence, we investigated how DETECT divided the tests into clusters to see if there were any discernable patterns with respect to known substantive item characteristics, such as, item type (e.g., foundational items, multi-part selected-response items, and Tier 1 writing prompt selected-response items), item position, cognitive load (e.g., tier level), and assorted other item content considerations. From our investigation we found no evidence that any of these characteristics were related to the DETECT clusters. However, we did find a strong and consistent, though unusual, pattern related to the clusters – the placement of the correct-response key option was a very strong indicator of the cluster membership of nearly every item. As an example, consider Form 1 of the grade 11 ELA test. This test form had 32 items, and the DETECT analysis reported a two-cluster solution. The first cluster contained 18 items, and the second had 14 items. The first cluster included all 14 items for which "A" was the correct response option, and the remaining 4 items in that cluster were three-option items for which "B" was the correct option. The second cluster contained all 18 items for which the last response option ("B" for two-option items and "C" for three-option items) was the correct response option. Thus, the first cluster contained all the items for which the correct option was not the last option, while the second cluster contained all the items for which the correct option was indeed the last option. All the test forms for all the grades for both mathematics and ELA showed similar clustering. In addition to this type of clustering, one other type of clustering was observed. For grades 3 and 4 on the ELA tests, the foundational items formed their own separate clusters.

These dimensionality analysis results indicate two types of violations of local independence: one having to do with how some student scores are related to the placement of the correct response options and the other having to do with the construct underlying the foundational items. In general, it is important that violations of local independence be understood, monitored, and controlled on tests. The violations of local independence having to do with the foundational items are well controlled by the combination of strict test specifications in regard to content, psychometric characteristics, and scoring so that their influence on student scores can be kept consistent from form to form and year to year.

The results of these analyses suggest further research is warranted to better understand the nature and implications of the findings. Since the NCSC 2015 administration is the first time a large enough sample of students with students with the most significant cognitive disabilities taking two and three response items has been available to enable these types of analyses, it was anticipated that further research and refinement of both the test and individual test items would be required. Based on results thus far, studies have been designed to continue dimensionality analyses through the second year of the program.

CHAPTER 9: REPORTING, INTERPRETATION, AND USE

Introduction

To ensure that reported results for the NCSC AA-AAS were accurate relative to collected data, a decision rules participation status hierarchy document delineating processing rules was prepared and approved by all participating states prior to preparing the results. The decision rules, which included a participation status of a "Tested, In Progress, Submitted and Closed" structure, provided the framework for the reporting requirements that were defined for each unique report at the State, District, school and individual student level, and approved by all participating states prior to reporting.

Development & Approval

The decision rules document (see above) was developed by the test administration vendor in collaboration with the NCSC Steering Committee. The decision rules document contains the hierarchy by which the participation statuses were assigned for each individual test incorporating data elements collected by the test platform and directly from the states. The reporting requirements, and corresponding report design templates, were developed by the vendor with the guidance of the NCSC Reporting Committee. Both documents underwent iterative review processes that included draft reviews by the appropriate committee, incorporation of edits, draft reviews by all participating states, and committee review and integration of feedback, until final revisions were approved by all participating states. The approved decision rules and reporting requirements are in Appendix 9-A.

To develop the report design templates, the vendor worked with the NCSC Reporting Committee to identify the data elements that were meaningful for reporting the NCSC results. Once identified, the vendor generated several different content displays to communicate each statistic. Using an interdisciplinary team consisting of reporting specialists, data analysts, psychometricians, and client services, and the NCSC Reporting Committee, each display was tailored to the specific needs of the report audience. Once content displays were chosen and the layout was finalized, the Reporting Committee provided the report text that was incorporated by the vendor's design team, culminating in a final draft. The results of this collaborative process were presented to participating state leads for final approval.

After reports were approved, the test administration vendor commenced development of the NCSC Guide for Score Report Interpretation. The vendor and NCSC collaborated to determine the layout and information that would be most helpful to teachers, as well as administrators and district and school staff; they also reviewed and discussed reports with parents and guardians. Due to varying state needs, NCSC requested that the test administration vendor generate a base document, delineating areas where states could include state-specific content. The guide included an overview of the NCSC AA-AAS, score reporting, and the various types of reports available to schools and districts. Guidelines were provided to inform the interpretation and use of NCSC AA-AAS scores. The guide also included explanations for all special reporting codes and messages, as well as explanations of the scores/codes for the constructed response writing items. States were permitted to remove codes not used in their state. The test administration vendor revised the base document through an iterative process with State Coordinators and

organizational partners, and the final, approved document was delivered to State NCSC Coordinators for state-specific revisions and distribution (see Appendix 9-B).

Primary Reports

The test administration vendor, in collaboration with the NCSC Reporting Committee created the following primary reports for the NCSC alternate assessment:

- Student Reports
- School Roster Reports
- School, District, and State Summary Reports

These confidential reports, along with student results data files, were posted online via the NCSC Assessment System's secure data and reporting portal. Access was controlled through user permissioned accounts, as illustrated in the following matrix:

Table 9-1. 2014–15 NCSC: Matrix by Users by Report

	State TC	*District TC*	*School TC*
Student Reports	Yes	Yes	Yes
School Roster Reports	Yes	Yes	Yes
School Summary Reports	Yes	Yes	Yes
District Summary Reports	Yes	Yes	No
State Summary Reports	Yes	No	No

As determined by NCSC, only Test Coordinators (TCs) were granted access to the online reports. For the purposes of the assessment system, State Coordinators were regarded as State TCs. As such, they were able to add new district and school TCs to the online system and to block users no longer in the TC role from accessing the system. Reports were generated for each school, district, and state that had results, as defined by the NCSC decision rules and reporting requirements.

The primary results reported were the student's scale score and performance level classification for Mathematics and English Language Arts. The performance level classifications, with cuts determined through the standard setting process, were reported under the generic titles of: Level 1, Level 2, Level 3, and Level 4, with Level 1 as the lowest level, and Level 4 as the highest attainable performance level.

The average scale score and percent of students in each performance level were summarized by school, district, and state on both the Roster and Summary reports. This allowed for the comparison of individual student performance in relation to the state, as well as for comparison of school and district results against the overall state results.

Student Report

The student report was a two-sided single-page document generated for each student eligible to receive a performance level in at least one content area, as defined by the student report requirements.

The report contained results for both content areas, and was developed for parents and guardians of students who participated in the NCSC AA-AAS. Reports were organized by school and posted via the secure-access portal for permissioned users to download, print, and disseminate to parents and guardians as appropriate. Each report contained the student name, test grade, and school on the front and back of the report. The back page also included the state student ID for additional confirmation of the student's identification. Additionally, some states chose to print and distribute paper versions of these reports to districts/schools for distribution to students' parents/guardians.

The front page of the report contained a brief overview of the NCSC alternate assessment, including examples of some of the built-in supports available during testing, and highlighted the compatibility of the assessment with various modes of communication. The front also contained a short overview of the results included on the back page, as well as a link to where more information could be accessed online. Parents and guardians were encouraged to communicate with their child's teacher about their child's specific mode of communication and performance.

The back page of the report contained the scaled score, performance level, and associated performance level descriptor for the level obtained by the student for each content area. A sentence below the graphical display encapsulated the standard error of measurement in an easy to understand manner by providing the expected range of scores the student would likely earn if tested again.

For students unable to show an observable mode of communication, the lowest scaled score was assigned and displayed along with the Level 1 performance level. This was annotated, and in place of the Level 1 performance level descriptor, the following text was displayed: Your child did not show a consistent observable mode of communication during the test and the test was closed by the teacher. Since your child did not complete the test the results may not be an accurate representation of your child's skills. If you have additional questions, please contact your child's teacher.

In the event that a student received a student report but did not receive results for one of the two content areas, results for the missing content area were replaced with text encouraging the parent or guardian to contact the child's teacher or school for more information.

Student Roster

The student roster was organized at the school level and provided a by-grade list of all students enrolled in the NCSC AA-AAS, with a snapshot of their participation status and results for both content areas. The number of tested students, the average scale score, and the percent of students by performance level were summarized for the school, district, and state at the top of the roster. Roster reporting requirements identified which of the participation status codes were included on the roster and which of the participation status codes were included in each calculation (see Appendix 9-C).

The summary information at the top of the student roster supported interpretation of results by users, typically those at the school and district levels. Given that many schools have a relatively small number of students in this population, NCSC did not suppress information when the number of students participating was small. This practice placed a burden on users to understand the data in the context of

small numbers and to use all information provided to understand the results, as explained in the NCSC Guide for Score Report Interpretation (see Appendix 9-B).

Student results were listed below the summary section and were identified by name and state student identification number. For each content area the following student level elements were reported:

- Participation Status
- Scale Score
- Performance Level
- Comparison to the State Average
- Optional by State: Writing Prompt Scores for Trait 1, Trait 2, and Trait 3 – See the section below on Variations and Amendments to the Primary Reports.

It is intended that these data points are to be used in conjunction with the 2015 Guide for Score Report Interpretation (see Appendix 9-B).

Summary Report

Summary reports were organized at the school, district, and state levels, for each entity with at least one student included in summary report calculations. Inclusion in these calculations was defined by the decision rules and summary report requirements. The following information was summarized by grade and content area, and displayed for the school, district, and state, based on the level of the report:

- Number of students enrolled
- Number of valid student tests
- Number of invalidated student tests
- Number of students who did not test
- Average scale score
- Number and percent of students at each performance level

This summary provided a comparative snapshot of results and participation information at a high level. It included both participation and performance summary information, allowing users to evaluate both aspects of their assessment results as guided by the 2015 Guide for Score Report Interpretation (see Appendix 9-B).

Variations and Amendments to the Primary Reports

To support individual requirements and needs of the operational states, the test administration vendor made the following variations and amendments to the primary reports.

Reporting of the Writing Prompt (Field Test):

Tier 2 writing prompts were field tested in ELA in each grade this year to enable further research and examination of results. Further development is in progress with the intention of including Tier 2 writing prompts in the overall ELA score for students in the future. Writing SRs and Tier 1 prompts were included in the overall ELA score for 2015. In order to provide feedback in support of instructional decisions, the vendor developed a second version of the student roster report to include the raw writing trait scores for the field tested Tier 2 writing prompts at the student level.

Only the student roster report was modified to include writing information based on students' responses to the writing prompt at Tier 2. If states opted to share this information on writing, schools received the version of the report that included writing trait scores for organization, idea development, and conventions. In the event a response was unable to be scored, a score code was provided instead for each of the three traits. The interpretation of scores or score codes was included in the 2015 Guide for Score Report Interpretation (see Appendix 9-B).

States opting to not report writing received the Roster Report without mention of writing. All states received the writing prompt scores for their students in the state level student results data file, regardless of whether they opted to include writing information on the student roster report.

State-Specific Variations:

The vendor worked with several states to accommodate their individual standard setting and reporting needs. These differences are delineated in Appendix 9-D (State-Specific Reporting Variations).

Quality Assurance

The vendor's proprietary quality assurance measures were embedded throughout the entire process of analysis and reporting. The data processors and data analysts that worked on the project implemented quality control checks of their respective computer programs. Moreover, when data were handed off to different functions within the Data and Reporting Services department, the sending function verified that the data were accurate prior to handoff. Additionally, when a function received a data set, the first step was to verify the data for accuracy.

A second level of quality assurance measure was parallel processing. One data analyst was responsible for writing all programs required to populate the student and aggregate reporting tables for the administration. Each reporting table was assigned to another data analyst on staff who used the decision rules to independently program the reporting table. The production and quality assurance tables were compared, and only after there was 100% agreement were the tables released for report generation.

The third aspect of the vendor's quality control involved the procedures implemented by the quality assurance group to check the accuracy of reported data. Using a sample of schools and districts, the quality assurance group verified that reported information was correct. The selection of sample schools and districts for this purpose was very specific and could have affected the success of the quality control efforts. There were three sets of samples selected that may not have been mutually exclusive. The first set included those that satisfied the following criteria:

- one-school district
- multischool district
- school(s) with at least one student meeting the criteria for each participation status defined in the decision rules

The second set of samples included districts or schools that had unique reporting situations as indicated by decision rules. This set was necessary in order to check that each rule was applied correctly.

The third set included districts and schools identified by NCSC states for its review and approval before reports were produced for distribution.

The quality assurance group used a checklist to implement its procedures. Once the checklist was completed, an internal parallel verification process was conducted, and then sample reports were circulated for psychometric checks and program management review. Samples of the final reports were then sent for NCSC state review and signoff. Simultaneously, several states, including Arizona, Connecticut, and New Mexico ran successful independent confirmations of the results contained in their state data files. After signoff was received from all states, the final reports were uploaded into the NCSC assessment system reporting portal.

References

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

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Browder, D. M., Flowers, C., & Wakeman, S. Y. (2008). Facilitating participation in assessments and the general curriculum: Level of symbolic communication classification for students with significant cognitive disabilities. *Assessment in Education: Principles, Policy, and Practice, 15*(2), 137-151.

Browder, D. M., Trela, K., & Jimenez, B. (2007). Training teachers to follow a task analysis to engage middle school students with moderate and severe developmental disabilities in grade appropriate literature. *Focus on Autism and Other Developmental Disabilities, 22*, 206–219.

Cameto, R., Haertel, G., Morrison, K., & Russell, M. (2010). *Synergistic use of evidence-centered design and universal design for learning for improved assessment design.* Menlo Park, CA: SRI International.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: Establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

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

Dolan, R. P., Rose, D. H., Burling, K. S., Harms, M., & Way, W. (2007). *The universal design for computer-based testing framework: A structure for developing guidelines for constructing innovative computer-administered tests*. Paper presented at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description. In P. W. Holland & H. Wainer (Eds.). *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons, Inc.

Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179–197.

Flowers, C., Wakeman, S., Browder, D., & Karvonen, M. (2009). An alignment protocol for alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice, 28(1),* 25-37.

Forte, E. (2013a). *Evaluating alignment for assessments developed using evidence-centered design*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Forte, E. (2013b). *The next generation of alignment.* Paper presented at the National Conference on Student Assessment, Washington, DC.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., & Van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Hess, K. K., (Ed.). (2010, December). *Learning progressions frameworks designed for use with the Common Core State Standards in mathematics k-12*. Lexington, KY: National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment, Dover, N.H. Retrieved from http://www.nciea.org/publication_PDFs/Math_LPF_KH11.pdf

Hess, K. K., (2011, December). *Learning progressions frameworks designed for use with the Common Core State Standards in English language arts.* Lexington, KY: National Alternate Assessment Center at the University of Kentucky and the National Center for the Improvement of Educational Assessment, Dover, N.H. Retrieved from http://www.nciea.org/publication_PDFs/ELA_LPF_12%202011_final.pdf

Hudson, M.E. & Test, D.W. (2011, March). Evaluating the evidence base of shared story reading to promote literacy for students with extensive support needs. *Research and Practice for Persons with Severe Disabilities*, *36*(1-2), 34-45.

Johnstone, C. J., Altman, J., & Thurlow, M. (2006). *A state guide to the development of universally designed assessments*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Joint Committee on Testing Practices. Retrieved from http://www.apa.org/science/programs/testing/fair-code.aspx

Kearns, J., Towles-Reeves, E., Kleinert, H., & Kleinert, J. (2006). *Learning characteristics inventory report*. Lexington, KY: University of Kentucky, National Alternate Assessment Center.

Kearns, J., Towles-Reeves, E., Kleinert, H., Kleinert, J., & Thomas, M. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *Journal of Special Education*, *45*(1), 3-14.

Kleinert, H. L., Browder, D. M., & Towles-Reeves, E. A. (2009). Models of cognition for students with significant cognitive disabilities: Implications for assessment. *Review of Educational Research, 79*, 301–326.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). Florence, KY: Taylor & Frances/Routledge.

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

Koppenhaver, D.A., Erickson, K.A., & Skotko, B.G. (2001). Supporting communication of girls with Rett syndrome and their mothers in storybook reading. *International Journal of Disability, Development and Education, 48(*4), 395-410.

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

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*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marion, S. F., & Pellegrino, J. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47-57.

Marion, S. F. & Perie, M. (2009). Validity arguments for alternate assessments. In W. Schafer & R. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 115-127). Baltimore, MD: Brooks Publishing.

Measured Progress Department of Psychometrics and Research. (2015). *Standard setting report*. Unpublished report.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education.

Mims, P., Hudson, M., & Browder, D. (2012). Using read alouds of grade-level biographies and systematic prompting to promote comprehension for students with moderate and severe developmental disabilities. *Focus on Autism and Developmental Disabilities, 27,* 67-80.

Mislevy, R. J. (2009). *Validity from the perspective of model-based reasoning* (CRESST Report 752). Los Angeles, CA: National Center for Research on Evaluation, Standards, & Student Testing.

Mislevy, R. J. and Haertel, G. D. (2006), Implications of Evidence-Centered Design for educational testing. *Educational Measurement: Issues and Practice*, 25: 6–20.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Washington, DC: National Academy Press.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2000). The bookmark procedure: Cognitive perspectives on standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

New Hampshire Enhanced Assessment Grant, National Alternate Assessment Center, & National Center for the Improvement of Educational Assessment. (2006a, October). *Documenting the technical quality of your state's alternate assessment system: An annotated workbook: Volume I: "Nuts and bolts"*. Lexington, KY: University of Kentucky, National Alternate Assessment Center.

New Hampshire Enhanced Assessment Grant, National Alternate Assessment Center, & National Center for the Improvement of Educational Assessment. (2006b, October). *An annotated workbook for documenting the technical quality of your state's alternate assessment system: Volume II: The validity evaluation.* Lexington, KY: University of Kentucky, National Alternate Assessment Center.

Quenemoen, R. F. (2009). The long and winding road of alternate assessments: Where we started, where we are now, and the road ahead. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 1227-153). Baltimore, MD: Brookes.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.). *Setting performance standards: concepts, methods, and perspectives* (pp. 159–173). Mahwah, NJ: Lawrence Erlbaum.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement, 43,* 215-243.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–617.

Sireci, S., Hambleton, R., & Bahry, L. (2013). *Developing performance level descriptors for the National Center and State Collaborative Alternate Assessment: Literature review and progress report*. Paper prepared for the National Center and State Collaborative Technical Advisory Committee, Chicago, IL.

Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M., A. J. van Duign, & T. A. B. Snijders (Eds.). *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.

Thurlow, M. L., Lazarus, S. S., & Christensen, L. L. (2013). Accommodations for assessment. In B. G. Cook & M. Tankersley (Eds.). *Research-based practices in special education* (pp. 311-327). Boston, MA: Pearson.

Webb, N. L. (2005, November). *Alignment, depth of knowledge, and change*. Paper presented at the annual meeting of the Florida Educational Research Association, Miami, Florida.

Wood, L., Browder, D. M., & Flynn, L. (2015). Teaching students with intellectual disability to use a self-questioning strategy to comprehend social studies text for an inclusive setting. *Research and Practice for Persons with Severe Disabilities*. Advance online publication.

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249.