The North Carolina Testing Program

Technical Report

2012–2015

Science Assessments

End-of-Grade 5, 8 and End-of-Course Biology

Public Schools of North Carolina
State Board of Education | Department of Public Instruction

Prepared by:

Thakur Karkee, Ph. D.
Kinge Mbella, Ph.D.
Min Zhu, Ph. D.
Hope Lung, Section Chief, Test Development
North Carolina Department of Public Instruction

March 2016

In compliance with federal laws, NC Public Schools administers all state-operated educational programs, employment activities and admissions without discrimination because of race, religion, national or ethnic origin, color, age, military service, disability, or gender, except where exemption is appropriate and allowed by law. Inquiries or complaints should be directed to:

Dr. Rebecca Garland, Deputy State Superintendent
Office of Accountability
6314 Mail Service Center
Raleigh, NC 27699-6314
Telephone (919) 807–3200; fax (919) 807–4065

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 Background and Overview

1. 1 Background

It is the intent of the North Carolina (NC) General Assembly to challenge each student in NC public schools with high expectations to learn, to achieve, and to fulfill his or her potential. To codify this, the General Assembly passed *GCS 115C-174.10* that states the following purposes for the testing program:

"(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results"

With that mission as its guide, the State Board of Education (SBE) developed a School-Based Management and Accountability Program to improve student performance in the early 1990s. In 1994, end-of-grade assessments designed to measure the SBE's adopted content standards were administered for the first time to all students in grades 3–8. Previously, assessments had not met alignment criteria, resulting in students not consistently receiving instruction on the content standards across the state. In 1996, the accountability system, referred to as Accountability, Basics, and Local Control (ABCs), used data from the end-of-grade assessments to inform parents, educators, and the public annually on the status of achievement at the school level. In the 1997–98 school year, five end-of-course tests were added to the ABCs school accountability model.

Since the 1990s, North Carolina has continually evolved its assessment system and its accountability system to increase academic expectations so students are prepared for success after high school. This was accomplished by reevaluating the content standards on a 5-year cycle and, based on these reviews, developing aligned assessments. Likewise, in keeping with continuous improvement, the ABCs model was amended to include additional end-of-course assessments and to fine-tune the model's business rules to ensure schools were being held accountable for all students.

The ABCs model continued until the 2012–13 school year when assessments aligned to the state's *Common Core Standards* in English Language Arts/Reading and Mathematics (adopted by the SBE in June 2010) and the NC *Essential Standards* (adopted by the SBE in February 2010) were implemented and the NC State Board of Education adopted a new accountability model. This document details the design, the development, and the outcomes of the assessments; and it provides evidence of the technical quality of the assessments. These attributes are evidence the test scores and the uses of the data are valid and reliable, and thus appropriate for reporting student achievement at the individual, school, district, and state levels. As with the ABCs, the test data are used for school accountability and for federal reporting. To provide additional context for the current edition of the assessments and the timeline for implementation, see Table 1-1.

Table 1-1 NCDPI Accountability and Testing Highlights

Year	*Action*
February 2010	The SBE adopted the NC Essential Standards for Science in February 2010.
June 2010	The SBE adopted the Standard Course of Study (based on the Common Core Standards for English Language Arts and Mathematics).
2011-12	Mathematics, Reading/English Language Arts and Science items field tested
2012 - 13	Mathematics, Reading/English Language Arts and Science assessments administered
July 2013	Mathematics, Reading/English Language Arts and Science standard setting conducted
October 2013	The SBE adopts academic achievement standards and performance level descriptors for Mathematics, Reading/English Language Arts and Science (revised by SBE action in March 2014).

1. 2 North Carolina Science Assessments

The End-of-Grade (EOG) assessments of Science in grades 5 and 8 are grade specific assessments aligned to the NC *Essential Standards* for Science *(*NC*ESS)* that measure NC students' Science skills. The standards are assessed again in high school with the Biology End-of-Course (EOC) assessment. The EOG and EOC assessments are administered to students in only English. Other native language translations are not yet available at this time.

The EOG and EOC Science assessments are available in both modes: paper-based fixed forms (A, B, and C) and computer-based fixed forms (M, N, and O). Each operational paper-based form has 60 operational multiple-choice (MC) items. Computer-based forms have 57 MC and 3 technology-enhanced (TE) items. Table 1-2 shows the summary of total operational items by item type and maximum score possible.

Table 1-2 Number of Operational Items and Maximum Possible Score by Item Type

Grade	Form	Total Score Points	MC Items		TE Items	
			No. of Items	Score Point	No. of Items	Score Points
Grade 5	A	60	60	60		
	B	60	60	60		
	C	60	60	60		
	M	60	57	57	3	3
	N	60	57	57	3	3
	O	60	57	57	3	3
Grade 8	A	60	60	60		
	B	60	60	60		
	C	60	60	60		
	M	60	57	57	3	3
	N	60	57	57	3	3
	O	60	57	57	3	3
Biology	A	60	60	60		
	B	60	60	60		
	C	60	60	60		
	M	60	57	57	3	3
	N	60	57	57	3	3
	O	60	57	57	3	3

Note: MC=Multiple-Choice; TE=Technology-Enhanced

The operational TE items include one text-identify (TI) and two drag-and-drop (DD) types. The DD items allows students to click and drag a response to a target location where students must outline and place words or phrases into text or label diagrams or graphs. The TI item type presents the student with a scrollable text and a question addressing information contained in the text. This type of item may provide the additional benefit of reducing the probability of guessing the correct answer to a negligible level.

North Carolina General Statute § 115C-174.12 mandates a statewide test administration window. Students on a semester schedule must be administered the EOG and EOC assessments during the final five (5) instructional days of the semester. For students enrolled in yearlong courses, EOG and EOC assessments must be administered during the final ten (10) instructional days of the school year. Students have up to four hours to complete each assessment.

1. 3 Report Summary

Chapter 1 provides a brief history of testing in North Carolina. The chapter also describes the main features of EOG Science and EOC Biology assessments, highlighting a description of each assessment, the intended population, and the administration window.

Chapter 2 presents an overview of the validation framework embedded throughout the design and development of the EOG and EOC assessments. Validity is a unifying and core concept in test development, and thus the gathering of evidence in support of proposed uses is fundamental and should be clearly documented. The first section provides a brief introduction of validity and an outline of key validity evidences as documented in this report. The second section discusses the main proposed uses of scores from EOG and EOC assessments.

Chapter 3 describes the 22-step test development outline adopted by the North Carolina Department of Public Instruction (NCDPI). Key steps described in this chapter include content standards, content specification and blueprints, item development, item-writer training, item review, and field test form assembly.

Chapter 4 describes the field test administration, including the sampling plan enacted to ensure that each form was administered to a representative sample of students. In addition, this chapter describes psychometric item analyses conducted on the field test data and the steps taken to construct the operational forms.

Chapter 5 of the technical report documents the procedures put in place by the NCDPI to assure the administrations of EOG and EOC assessments are standardized and fair and secured for all students across the state. The chapter also describes the accommodation procedures implemented to ensure all students with disabilities and English Language Learners are able to take EOG and EOC assessments.

Chapter 6 describes the processes used for scoring items and the procedure adopted to create final reportable scale scores. The first section of this chapter summarizes the automated scoring procedures used to transform students' responses into a number correct score for fixed response items. Sections two describes the procedures used to transform raw scores into a reportable scale across the different grades. The final section describes the data certification processes used by the NCDPI to ensure the quality of student data.

Chapter 7 describes the analyses of operational data after the first operational administrations of EOGs and EOCs assessments in 2012–13. The chapter begins with a description of the random spiraling process used to administer six parallel forms (three paper- and three computer- based) across North Carolina. This chapter also summarizes item analysis results from the operational administration in 2012–13, which includes CTT (p-value, biserial correlation, Cronbach's alpha) and IRT-based analysis (item calibration and scoring, test characteristics curves, test information functions, and conditional standard errors).

Chapter 8 presents a summary of the standard setting study that was conducted in July 2013 after the first operational administration of EOGs and EOC assessments. The NCDPI contracted with Pearson Inc. to conduct a standard-setting workshop to recommend cut scores and achievement levels for the newly developed EOG and EOC science assessments. This chapter is a condensed version of the final report prepared by Pearson, describing the full workshop and final cuts score recommendations.

Chapter 9 presents summary student performance results for EOG and EOC assessments from 2012 through the 2015 administration. This chapter is organized into two main sections. Section one highlights descriptive summary results of scale scores and reported achievement levels for EOG and EOC forms across major demographic variables. The second section of this chapter presents samples and summary descriptions of the various standardized reports created by the NCDPI and available to LEAs to share assessment results with various stakeholders.

Chapter 10 presents summary validity evidence collected in support of the interpretation of EOG and EOC test scores. The first couple of sections in this chapter present validity evidence in support of internal structures of EOG and EOC assessments. Evidences presented in these sections includes reliability, standard error estimates, classification consistency, summary of reported achievement levels, and exploratory Principal Component Analysis in support of the unidimensional analysis and interpretation of test scores. The final sections of the chapter document validity evidence based on content summarized from the alignment study and the relation to other variables summarized from correlation with external variables. The very last part of Chapter 10 presents a summary of procedures used to ensure EOG and EOC assessments are accessible and fair to all students.

Chapter 2 Validity Framework and Uses

This chapter presents an overview of the validation framework embedded throughout the design and development of the EOG and EOC assessments. Validity is a unifying and core concept in test development and thus the gathering of evidence in support of proposed uses is fundamental and should be clearly documented. The first section provides a brief introduction of validity and an outline of key validity evidences. The second section discusses the main uses of scores from EOG and EOC assessments.

2. 1 Summary Validation Framework for Science

A fundamental purpose of this technical report is to present and document validity evidences on the proposed inferences of EOG and EOC test scores as highlighted in *The Standards for Educational and Psychological Testing* (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014) hereafter referred to as the *Standards (AERA, APA, and NCME, 2014).*

> *"Validity refers to the degree to which evidences and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. . . . It is the interpretations of test scores for proposed uses that are evaluated, not the test itself."*

Standard 1.0 of the *Standards* states, "Clear articulation of each intended test score interpretation for the specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be presented" (p. 23). Throughout this technical report, the NCDPI will be constructing, evaluating, and documenting relevant evidences validating the proposed uses of test scores. From the test developer's perspective, validation is a fluid process of evidence gathering that begins with the declaration of the proposed test use and continues throughout the life cycle of the test.

As test developer of EOG and EOC assessments, the NCDPI has adopted a validation framework consistent with that prescribed in the *Standards (AERA, NCME & APA, 2014).* Under this framework, the NCDPI is committed to ongoing evaluation of the quality of its assessments and relevance of their intended uses by continuously collecting and updating validity evidences as new data become available. Linn (2002, p. 46) noted that serious planning and a great deal of effort is required to accumulate evidences needed to validate the intended uses and interpretations of state assessments. His recommendation is to prioritize so that the most critical validity questions can be addressed first: "…what are the arguments for and against the intended aims of the test? And what does the test do in the system other than what it claims? . . . For such questions, it is helpful to consider the level of stakes that are involved in the use or interpretation of results and then give the higher priority to those areas with highest stakes" (Linn, 2002).

Throughout this document, validity arguments and evidences have been summarized based on prioritization of components relevant to establishing the technical quality of EOG and EOC Science assessments. Even though each chapter highlights arguments and components related to particular source[s] of validity evidence, it is worth mentioning that the validation framework adapted by the NCDPI and endorsed by the *Standards* is a coherent process. A sound validity argument of the degree to which existing theory and evidence supports intended score interpretations is accomplished only by applying a holistic approach. **Error! Reference source not found.** presents an outline of the validation framework with relevant components as documented in this report.

Table 2-1 NCDPI Validation Framework for Science EOG and EOC Assessments

Sources of Validity Evidence	References	Data
Evidence based on intended uses	Chapter 2	Score Report Samples
Evidence based on content	Chapter 10	SEC alignment Part 1
Evidence of careful test construction	Chapter 3	Test construction steps, item review map
Evidence based on appropriate test administration	Chapter 5	Assessment Guides
Evidence based on internal structure and reliability	Chapter 10	Cronbach alpha and CSEM, Classification Consistency, Principal Component Analysis
Evidence based on appropriate scoring, scaling, and standard setting	Chapters 7, 8	Standard Setting Report
Evidence based on careful attention to fairness for all test takers	Chapters 3, 5, 10	Assessment Guides
Evidence based on appropriate reporting	Chapter 9	ISR, Goal Summary Reports, Scale Score Frequency Reports
Evidence based on relations to other variables	Chapter 10	Relationship with External Variables

2. 2 Uses of NC Science EOG/EOC Assessments

The NCDPI designs, develops, and administers customized high-quality North Carolina State Testing Program (NCSTP) assessments in grades 3–8 and high school that are aligned to College-and-Career Readiness standards for science, adopted by the North Carolina State Board of Education (NC*SBE*) in February 2010. These assessments provide valid and reliable information intended to serve two general purposes:

- Measure students' achievement and progress to readiness as defined by College-and-Career Readiness standards. Scores from EOG and EOC are transformed, grouped and reported into 1 of 5 achievement levels (in 2012-13 scores were reported using 4 achievement levels) corresponding to 1 of the 5 performance level descriptors adopted by the state to classify students based on their progress and readiness as defined by NC*ESS* College-and-Career Readiness standards.
- Assessment results are used for school and district accountability under the READY Accountability Model and for Federal reporting purposes. EOG and EOC students' score data are part of the quantitative indicators used in two main components of the new state READY accountability model: educator effectiveness and school performance grades. In the first component, the educator-effective model, which states teachers (standard 6) and school executives (standard 8) will contribute to the academic success of students and data from EOG and EOC assessments are used in a statewide value-added growth model to provide ratings for these respective standards. The second component is the use of score data from EOG and EOC assessments in the school report card for the calculation of school performance grade. Effective with the 2013–14 school year, each school was assigned a performance letter grade which included indicators of students' performance in EOG and EOC assessments.

In addition to these main uses, the NCSBE also mandates that at least 20 percent of the students' final grade in Biology must come from their EOC assessment scores. It is worth mentioning that the EOG assessments in grades 5 and 8 are not intended to be used as a main indicator for decisions on grade-level retention or promotion. To ensure all EOG and EOC assessment test scores are used as intended, the NCDPI provides score reports at the student, school, district, and state level. The North Carolina *Testing Code of Ethics* (see **Appendix 2-A**) dictates that educators use test scores and reports appropriately, which means that educators should recognize that a test score is only one piece of information and must be interpreted as intended. That is, the validity of a test refers to the valid interpretation[s] of test scores not the test itself (AERA, APA, & NCME, 2014).

In order to be consistent with standard 1.1 of the *Standard*, test "developers should set forth clearly how test scores are intended to be interpreted and consequently used. . . ." (p. 23).

The NCDPI WinScan software application is provided to school administrators at the district level to generate a variety of score reports for score interpretations: class roster reports, score frequency reports, achievement level frequency reports, and goal summary reports. Interpretive guides for the various score reports are published on the NCDPI website to help educators and decision makers at the classroom, school, and district levels understand the content and uses of these reports. These guides and reports are intended to help administrators and educators explain test results to parents and the general public. *Table 2-2* shows a list of reports described in subsequent sections and their intended audiences. The individual student reports (ISRs) are designed for students, parents, teachers, and school administrators. Class rosters are designed for teachers and school administrators. Score frequency reports, achievement level frequency reports, and goal summary reports are designed for teachers, school administrators, district administrators, and state administrators.

Table 2-2 WinScan Reports and Intended Audience

Report	*Audience*				
			Administrators		
	Parent	Teacher	School	District	State
Individual Student Report (ISRs)	✓	✓	✓		
Class Roster Reports		✓	✓		
Score and Achievement Level Frequency Reports		✓	✓	✓	✓
Goal Summary Reports		✓	✓	✓	✓

2. 3 Confidentiality of Student Test Scores

State Board of Education policy GCS-A-010 (j)(1) states, “Educators shall maintain the confidentiality of individual students. Publicizing test scores or any written material containing personally identifiable information from the student’s educational records shall not be disseminated or otherwise made available to the public by a member of the State Board of

Education, any employee of the State Board of Education, the State Superintendent of Public Instruction, any employee of the North Carolina Department of Public Instruction, any member of a local board of education, any employee of a local board of education, or any other person, except as permitted under the provisions of the Family Educational Rights and Privacy Act of 1974, 20 U.S.C.§1232g."

Chapter 3 Test Development Process

Standard 4.0 of the *Standards* states, “Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population” (p. 85). In adherence to the *Standards*, this chapter documents steps implemented by the NCDPI during design and development of EOG and EOC assessments. Key aspects of design and development described in this chapter include content standards, content specification and blueprints, item development, and item review. *Table 3-1* shows the sequence of events for the test development prescribed by the North Carolina State Board of Education (NCSBE; 2003, 2012). According to NCSBE policy (2012):

> *…the state-adopted content standards are periodically reviewed for possible revisions; however, test development is continuous. The NCDPI Accountability Services/Test Development Section test development staff members begin developing* ***operational*** *test forms for the North Carolina Testing Program when the State Board of Education determines that such tests are needed. The need for new tests may result from mandates from the federal government or the North Carolina General Assembly. New tests can also be developed if the SBE determines the development of a new test will enhance the education of North Carolina students. The test development process consists of six phases and takes approximately four years.   The phases begin with the development of test specifications and end with the reporting of operational test results.*

Additional information regarding the North Carolina State Assessment development process, including test specifications, items and form formats, alignment studies, test administrations for alternate assessments, and students with disabilities (SWD) and English Language Learner (ELL); standard setting; reporting; and uses of data for measuring growth can also be found in the technical brief (NCDPI, 2014) on the NCDPI website.

Even though the NCSBE (2012) policy states that the “test development process consists of six phases and take(s) approximately four years,” only two years were allotted to NCDPI to develop and administer the first operational assessments aligned to NC*ESS*. To accommodate the shortened timeline, NCDPI made three modifications to the SBE assessment development flow chart depicted in Table 3-1**:**

I. The NCDPI waived the full-scale “item tryout” component (Steps 3–8) and implemented a smaller scale of item tryout for the newly developed innovative technology-enhanced item types.

II. The NCDPI also waived pilot testing (Step 18) because pilot tests are administered only for newly developed items, not for assessments revised from a preceding test (GCS-A-013, Phase 4: Pilot/Operational Test Development, Step 18: Administer Test as Pilot, footnote 5).

III. The NCDPI used operational data (Step 21) instead of field test data for the Standard Setting process (Step 20).

Table 3-1 Flow Chart of Test Development of North Carolina Assessments

Adopt Content Standards	Step 8 Develop New Items	Step16 Review Assembled Test
Step 1[a] Develop Test Specifications (Blueprint)	Step 9[b] Review Items for Field Test	Step17 Final Review of Test
Step 2[b] Develop Test Items	Step 10 Assemble Field Test Forms	Step 18[ab] Administer Test as Pilot
Step 3[b] Review Items for Tryouts	Step 11 Review Field Test Forms	Step19 Score Test
Step 4 Assemble Item Tryout Forms	Step 12[b] Administer Field Test	Step 20[ab] Establish Standards
Step 5 Review Item Tryout Forms	Step 13 Review Field Test Statistics	Step 21[b] Administer Test as Fully Operational
Step 6[b] Administer Item Tryouts	Step14[b] Conduct Bias Reviews	Step 22 Report Test Results
Step 7 Review Item Tryout Statistics	Step15 Assemble Equivalent and Parallel Forms	

[a] Activities done only at implementation of new curriculum

[b] Activities involving NC teachers

3.1 Content Standards and Curriculum Connectors

As stated in Chapter 1 (see Table 1-1), the NC*SBE* adopted the revised NC*ESS* in June 2010. Operational test forms aligned to the NC*ESS* were administered in 2012–13 testing administration (READY initiative). Testing of North Carolina students' skills relative to the standards and objectives in the NC*ESS* is one component of the NCSTP. To ensure items written for the EOG and EOC assessments met the cognitive rigor as specified in the adopted standards, NCSTP worked with curriculum to provide training workshops on Revised Bloom Taxonomy (RBT), Webb's Depth of Knowledge (DOK), and overall alignment of assessments to content standards.

3.1.1 Revised Bloom Taxonomy and Depth of Knowledge

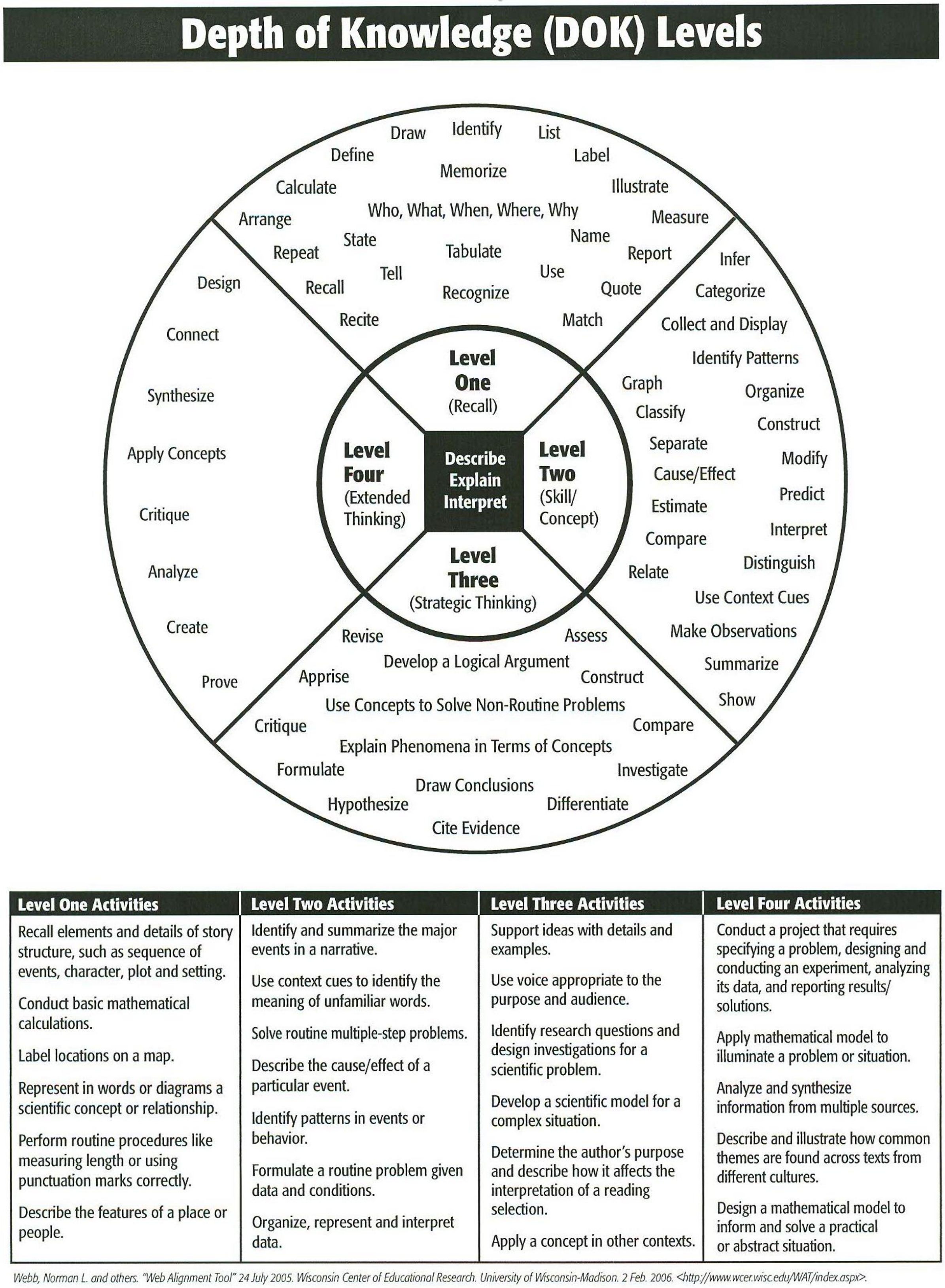
As part of pre-item development training for the new EOG and EOC assessments, NCSTP, with collaboration from the NCDPI's Curriculum Division, organized two main workshops on RBT and Webb's DOK. The first workshop was organized on July 8, 2010, and the focus was to get NCDPI Test Measurement Specialists (TMSs), North Carolina State University-Technical Outreach for Public Schools (NCSU-TOPS) content leads, and NCDPI Curriculum Content Specialists familiarized with Hess's matrix, which the NCDPI had decided to use for alignment purposes because it relates RBT to Webb's alignment scheme. Dr. Karin Hess (The National Center for the Improvement of Educational Assessment, Inc. (NCIEA) also known as Center for Assessment) developed a 4 by 6 table containing Webb's DOK levels across the top and RBT process dimension down the side (see Table 3-2). During the workshop participants received training and started to classify NC*ESS* using Hess's matrix.

On July 26, 2010, NCDPI organized a one-day, face-to-face training session on Webb's Alignment. Norm Webb was invited and served as lead facilitator on alignment and DOK training. During the first four hours of the training, Webb presented an overview of his alignment model (Webb et. al. 2005) and his definitions of Depth-of-Knowledge (see Figure 3-1**)**. Slides used for the training are in Appendix 3-A Norm Webb Training – Content Complexity. This workshop built on the July 8th workshop in which participants were able to classify standards using Hess's matrix. During the July 26th workshop, participants received training on aligning items using the RBT framework and how to classify items based on their cognitive complexity using the Webb alignment tool, which organizes verbs into general DOK categories.

Table 3-2 Hess's Cognitive Rigor Matrix with Curricular Examples

Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 Recall & Reproduction	Level 2 Skills & Concepts	Level 3 Strategic Thinking/Reasoning	Level 4 Extended Thinking
Remember Retrieve knowledge from long-term memory, recognize, recall, locate, identify	○ Recall, recognize, or locate basic facts, ideas, principles ○ Recall or identify conversions between representations, numbers, or units of measure ○ Identify facts/details in texts			
Understand Construct meaning, clarify, paraphrase, represent, translate, illustrate, give examples, classify, categorize, summarize, generalize, infer a logical conclusion (such as from examples given), predict, compare/contrast, match like ideas, explain, construct models	○ Compose & decompose numbers ○ Evaluate an expression ○ Locate points (grid/ number line) ○ Represent math relationships in words, pictures, or symbols ○ Write simple sentences ○ Select appropriate word for intended meaning ○ Describe/explain how or why	○ Specify and explain relationships ○ Give non-examples/examples ○ Make and record observations ○ Take notes; organize ideas/data ○ Summarize results, concepts, ideas ○ Make basic inferences or logical predictions from data or texts ○ Identify main ideas or accurate generalizations	○ Explain, generalize, or connect ideas using supporting evidence ○ Explain thinking when more than one response is possible ○ Explain phenomena in terms of concepts ○ Write full composition to meet specific purpose ○ Identify themes	○ Explain how concepts or ideas specifically relate to other content domains or concepts ○ Develop generalizations of the results obtained or strategies used and apply them to new problem situations
Apply Carry out or use a procedure in a given situation; carry out (apply to a familiar task), or use (apply) to an unfamiliar task	○ Follow simple/routine procedure (recipe-type directions) ○ Solve a one-step problem ○ Calculate, measure, apply a rule ○ Apply an algorithm or formula (area, perimeter, etc.) ○ Represent in words or diagrams a concept or relationship ○ Apply rules or use resources to edit spelling, grammar, punctuation, conventions	○ Select a procedure according to task needed and perform it ○ Solve routine problem applying multiple concepts or decision points ○ Retrieve information from a table, graph, or figure and use it solve a problem requiring multiple steps ○ Use models to represent concepts ○ Write paragraph using appropriate organization, text structure, and signal words.	○ Use concepts to solve non-routine problems ○ Design investigation for a specific purpose or research question ○ Conduct a designed investigation ○ Apply concepts to solve non-routine problems ○ Use reasoning, planning, and evidence ○ Revise final draft for meaning or progression of ideas	○ Select or devise an approach among many alternatives to solve a novel problem ○ Conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results ○ Illustrate how multiple themes (historical, geographic, social) may be interrelated
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	○ Retrieve information from a table or graph to answer a question ○ Identify or locate specific information contained in maps, charts, tables, graphs, or diagrams	○ Categorize, classify materials ○ Compare/contrast figures or data ○ Select appropriate display data ○ Organize or interpret (simple) data ○ Extend a pattern ○ Identify use of literary devices ○ Identify text structure of paragraph ○ Distinguish: relevant-irrelevant information, fact/opinion	○ Compare information within or across data sets or texts ○ Analyze and draw conclusions from more complex data ○ Generalize a pattern ○ Organize/interpret data: complex graph ○ Analyze author's craft, viewpoint, or potential bias	○ Analyze multiple sources of evidence or multiple works by the same author, or across genres or time periods ○ Analyze complex/abstract themes ○ Gather, analyze, and organize information ○ Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			○ Cite evidence and develop a logical argument for concepts ○ Describe, compare, and contrast solution methods ○ Verify reasonableness of results ○ Justify conclusions made	○ Gather, analyze, & evaluate relevancy & accuracy ○ Draw & justify conclusions ○ Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	○ Brainstorm ideas, concepts, or perspectives related to a topic or concept	○ Generate conjectures or hypotheses based on observations or prior knowledge	○ Synthesize information within one source or text ○ Formulate an original problem given a situation ○ Develop a complex model for a given situation	○ Synthesize information across multiple sources or texts ○ Design a model to inform and solve a real-world, complex, or abstract situation

Figure 3-1 Webb alignment Tool



Depth of Knowledge (DOK) Levels

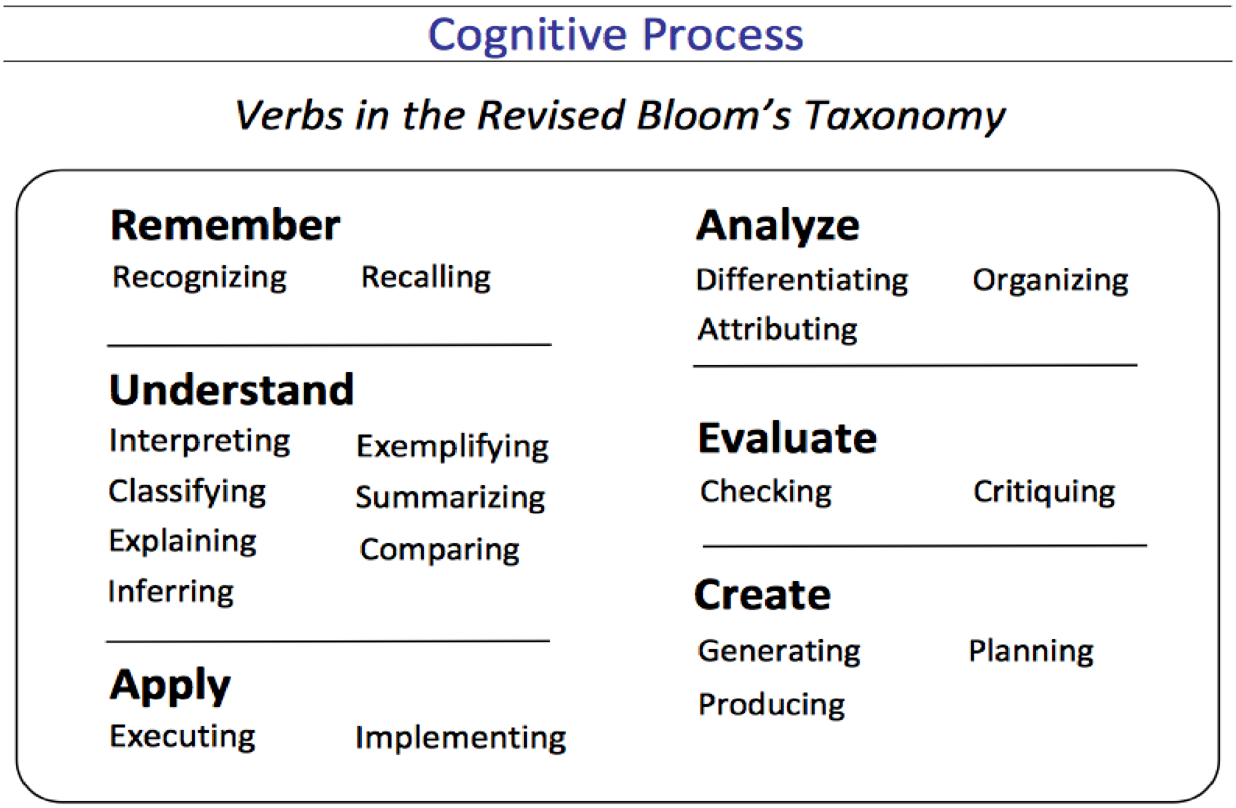
Level One Activities	Level Two Activities	Level Three Activities	Level Four Activities
Recall elements and details of story structure, such as sequence of events, character, plot and setting. Conduct basic mathematical calculations. Label locations on a map. Represent in words or diagrams a scientific concept or relationship. Perform routine procedures like measuring length or using punctuation marks correctly. Describe the features of a place or people.	Identify and summarize the major events in a narrative. Use context cues to identify the meaning of unfamiliar words. Solve routine multiple-step problems. Describe the cause/effect of a particular event. Identify patterns in events or behavior. Formulate a routine problem given data and conditions. Organize, represent and interpret data.	Support ideas with details and examples. Use voice appropriate to the purpose and audience. Identify research questions and design investigations for a scientific problem. Develop a scientific model for a complex situation. Determine the author's purpose and describe how it affects the interpretation of a reading selection. Apply a concept in other contexts.	Conduct a project that requires specifying a problem, designing and conducting an experiment, analyzing its data, and reporting results/ solutions. Apply mathematical model to illuminate a problem or situation. Analyze and synthesize information from multiple sources. Describe and illustrate how common themes are found across texts from different cultures. Design a mathematical model to inform and solve a practical or abstract situation.

Webb, Norman L. and others. "Web Alignment Tool" 24 July 2005. Wisconsin Center of Educational Research. University of Wisconsin-Madison. 2 Feb. 2006. <http://www.wcer.wisc.edu/WAT/index.aspx>.

3.1.2 Curriculum Development

North Carolina uses the RBT to help educate students in the complex thinking skills expected of 21st Century graduates. The RBT was chosen because it has well-defined verbs and is based on modern cognitive research. RBT categorizes both the **cognitive process** (Figure 3-2) and the **knowledge dimension** of the standard. The cognitive process is delineated by the verb used in the standard. The chart below illustrates the verbs used in the RBT and their specific definitions.

Figure 3-2 Cognitive Process: Verbs in the Revised Bloom's Taxonomy



Cognitive Process

Verbs in the Revised Bloom's Taxonomy

Remember
Recognizing Recalling

Understand
Interpreting Exemplifying
Classifying Summarizing
Explaining Comparing
Inferring

Apply
Executing Implementing

Analyze
Differentiating Organizing
Attributing

Evaluate
Checking Critiquing

Create
Generating Planning
Producing

From *Anderson, Lorin and David Krathwohl, A Taxonomy For Learning, Teaching and Assessing. New York: Longman, 2001.*

A common understanding of these verbs by teachers is the backbone of professional development around the new standards. The knowledge dimension is a way to categorize the type of knowledge to be learned. For instance, in the standard "the student will understand the concept of equality as it applies to solving problems with unknown quantities," the knowledge to be learned is *"the concept of equality as it applies to solving problems with unknown quantities."*

Knowledge in the RBT falls into four categories:

- Factual Knowledge
- Conceptual Knowledge
- Procedural Knowledge
- Meta-Cognitive Knowledge

3.2 Step 1–Content Domain Specification and Blueprints

Test specifications[c] for the NCSTP were developed in accordance with the standards and objectives specified in the NC*ESS*. AERA/APA/NCME Standard 4.1 states:

> *Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s) (p. 85).*

In addition, AERA/APA/NCME Standard 4.12 states, "Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications" (p. 89).

The NCDPI invited teachers to collaborate and develop recommendations for a prioritization of the standards indicating the relative importance of each standard, the anticipated instructional time, and the appropriateness of the standard to different item types. Subsequently, curriculum and test development staff from the NCDPI met and reviewed the results from the teacher panels and developed weighted distributions of the number of items sampled across domains for each grade level.

Table 3-3 through *Table* 3-5 show the adopted content domain specification as well as item types for EOG Science Grades 5 and 8 and EOC Biology assessments by form. Based on the content domain specification, test blueprints were developed that matched the number of items from each standard to be represented on each test form. The tables show that at the domain level and in terms of the relative emphasis of the standards coverage, all test forms (paper- and

[c] The EOG and EOC assessment specifications information can be found in the following website: http://www.ncpublicschools.org/accountability/testing/technicalnotes

computer- based) conform closely to the content domain specification and blueprints (*see* Appendix 3-B Content Domain Specification and Blueprints).

The paper form consisted of all MC items. Computer based forms have two new additional item types; drag-and-drop (DD) and text identify (TI). The Computer based forms' content domain by item types are shown in *Table 3-4* for grade 5, Table 3-6 for grade 8, and *Table 3-8* for Biology. Each Computer based form consisted of 57 MC, 2 DD, and 1 TI except for grade 8 science Form O where there are 58 MC, 1 DD, and 1 TI items. *Section 3.3.3* describes the characteristics of the DD and TI item types.

Table 3-3 Content Standards and Weight Distribution, Grade 5 Science

Domain	Blue Print (%)	Form A		Form B		Form C		Form M		Form N		Form O	
		No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%
Forces and Motion (5.P.1)	13–15	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3
Matter: Properties and Change (5.P.2)	12–14	9	15	9	15	9	15	9	15.0	8	13.3	9	15.0
Energy: Conservation and Transfer (5.P.3)	11–13	5	8.3	6	10	5	8.3	6	10.0	6	10.0	4	6.7
Earth Systems, Structures and Processes (5.E.1)	15–17	10	16.7	11	18.3	11	18.3	10	16.7	11	18.3	11	18.3
Structures and Functions of Living Organisms (5.L.1)	14–16	10	16.7	9	15	9	15	10	16.7	9	15.0	9	15.0
Ecosystems (5.L.2)	14–16	10	16.7	10	16.7	9	15	10	16.7	11	18.3	10	16.7
Evolution and Genetics (5.L.3)	13–15	8	13.3	7	11.7	9	15	7	11.7	7	11.7	9	15.0
Total	100	60	100	60	100	60	100	60	100	60	100	60	100

Table 3-4 Computer Forms Content Standards by Item Type, Grade 5 Science

Domain	Form M				Form N				Form O			
	DD	MC	TI	Total	DD	MC	TI	Total	DD	MC	TI	Total
Forces and Motion (5.P.1)	0	7	1	8	0	8	0	8	0	8	0	8
Matter: Properties and Change (5.P.2)	0	9	0	9	0	8	0	8	0	9	0	9
Energy: Conservation and Transfer (5.P.3)	1	5	0	6	0	5	1	6	0	4	0	4
Earth Systems, Structures and Processes (5.E.1)	0	10	0	10	1	10	0	11	0	11	0	11
Structures and Functions of Living Organisms (5.L.1)	1	9	0	10	0	9	0	9	0	9	0	9
Ecosystems (5.L.2)	0	10	0	10	1	10	0	11	2	7	1	10
Evolution and Genetics (5.L.3)	0	7	0	7	0	7	0	7	0	9	0	9
Total	2	57	1	60	2	57	1	60	2	57	1	60

Table 3-5 Content Standards and Weight Distribution, Grade 8 Science

Domain	Blue Print (%)	Form A		Form B		Form C		Form M		Form N		Form O	
		No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%	No. of Items	%
Matter: Properties and Change (8.P.1)	14-16	10	16.7	10	16.7	10	16.7	10	16.7	10	16.7	10	16.7
Energy: Conservation and Transfer (8.P.2)	10-12	6	10	6	10	6	10	6	10.0	6	10.0	6	10.0
Earth Systems, Structures and Processes (8.E.1)	13-15	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3
Earth History (8.E.2)	11-13	7	11.7	7	11.7	7	11.7	7	11.7	7	11.7	7	11.7
Structures and Functions of Living Organisms (8.L.1/8.L.2)	19-23	11	18.3	11	18.3	9	15	11	18.3	11	18.3	9	15.0
Ecosystems (8.L.3)	9-11	6	10	6	10	8	13.3	6	10.0	6	10.0	8	13.3
Evolution and Genetics (8.L.4)	11-13	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3	8	13.3
Molecular Biology (8.L.5)	8-10	4	6.7	4	6.7	4	6.7	4	6.7	4	6.7	4	6.7
Total	100	60	100	60	100	60	100	60	100	60	100	60	100

Table 3-6 *Computer Forms Content Standards by Item Type, Grade 8 Science*

Domain	Form M				Form N				Form O			
	DD	MC	TI	Total	DD	MC	TI	Total	DD	MC	TI	Total
Matter: Properties and Change (8.P.1)	1	8	1	10	0	9	1	10	1	8	1	10
Energy: Conservation and Transfer (8.P.2)	0	6	0	6	0	6	0	6	0	6	0	6
Earth Systems, Structures and Processes (8.E.1)	0	8	0	8	1	7	0	8	0	8	0	8
Earth History (8.E.2)	0	7	0	7	1	6	0	7	0	7	0	7
Structures and Functions of Living Organisms (8.L.1/8.L.2)	0	11	0	11	0	11	0	11	0	9	0	9
Ecosystems (8.L.3)	0	6	0	6	0	6	0	6	0	8	0	8
Evolution and Genetics (8.L.4)	1	7	0	8	0	8	0	8	0	8	0	8
Molecular Biology (8.L.5)	0	4	0	4	0	4	0	4	0	4	0	4
Total	2	57	1	60	2	57	1	60	1	58	1	60

Table 3-7 Content Standards and Weight Distribution, Biology

Domain	Code	Blue Print (%)	Form A/M		Form B/N		Form C/O	
			No. of Items	%	No. of Items	%	No. of Items	%
Structures and Functions of Living Organisms	Bio.1.1/1.2	18-22	12	20.0	12	20.0	12	20.0
Ecosystems	Bio.2.1/2.2	18-22	12	20.0	12	20.0	12	20.0
Evolution and Genetics	Bio.3.1/3.2 /3.3/ 3.4/3.5	43-53	27	45.0	27	45.0	27	45.0
Molecular Biology	Bio.4.1/4.2	15-19	9	15.0	9	15.0	9	15.0
Total		100	60	100	60	100	60	100

Table 3-8 Computer Form Content Standards by Item Type, Biology

		Form M				Form N				Form O			
Domain	Code	DD	MC	TI	Total	DD	MC	TI	Total	DD	MC	TI	Total
Structures and Functions of Living Organisms	Bio.1.1/1.2	1	11	0	12	0	12	0	12	1	11	0	12
Ecosystems	Bio.2.1/2.2	1	11	0	12	0	11	1	12	0	12	0	12
Evolution and Genetics	Bio.3.1/3.2 /3.3/ 3.4/3.5	0	26	1	27	1	26	0	27	0	26	1	27
Molecular Biology	Bio.4.1/4.2	0	9	0	9	1	8	0	9	1	8	0	9
Total		2	57	1	60	2	57	1	60	2	57	1	60

DD=Drag-and-drop, MC=Multiple-Choice, TI=Text identify

3.3 Step 2-Item Development

In Step 2, NCDPI began the process of writing and aligning items to NC grade-level assessments blueprints. This section as well as *Sections 3.4* and *3.5* discuss item development in order to comply with AERA/APA/NCME Standard 4.7, which states, “The procedures used to develop, review, and try out items and to select items from the item pool should be documented” (p. 87).

3.3.1 Plain English Approach

Before the development of items, the NCDPI on April 28, 2011, conducted a workshop on the use of “Plain English” practices in test construction. The workshop was facilitated by Dr. Edynn Sato, Director of Research and English Learner Assessment with the Assessment and

Standard Development Services Program at West Ed. Target participants for this workshop included personnel from NCDPI Accountability Division (also includes the test development section), the K-12 Curriculum and Instruction Division, and NCSU-TOPS staff. The one day training workshop focused on the latest research in the area of plain English practices and examined its use in the NCDPI training of item writers and reviewers. Lessons learned from this training were used to reevaluate how items for the new assessments were developed following the plain English framework, which emphasizes clarity without altering the construct being assessed. In general, the goal was to develop items that assess the construct without adding in the construct- irrelevant variance that may come into play if the students cannot access and interpret what is being required of them.

The training emphasized aspects of the test items, such as presentation of material, socio-cultural contexts, and culture-specific references, which may interfere with the measurement of the students' ability to demonstrate their knowledge of the content. This is also known as construct-irrelevant variance. Such construct-irrelevant variance can lead to an underestimation of the students' true ability levels. Strategies such as Universal Design and Plain English have been found to increase access by reducing unnecessary linguistic and cultural complexities, thus reducing construct-irrelevant variance for students for whom these factors may exist while yet maintaining appropriate measurement of the construct for the entirety of the student population.

The concept of Universal Design originated in architecture with the goal of providing the maximum accessibility and usability of buildings, outdoor spaces, and living environments. This concept centered on the belief that our environments should be accessible and usable by everyone, regardless of their age, ability, or circumstance. When applied to learning and assessment, Universal Design centers around development and creation of learning environments and assessments that are accessible and usable by students of all abilities, including SWD and ELL. These core principles are emphasized in the item writer training courses designed by the NCDPI and required to be taken by all potential item writers/reviewers. The complete workshop materials, including the workshop agenda, are available in Appendix 3-C Exhibit 307 Plain English Training_042811.

3.3.2 Item Writer Training

North Carolina educators from across the state were recruited and trained to develop new items. The diversity among the item writers and their knowledge of the current NCE*SS* was addressed during recruitment. Educators with expertise and experience with students with disabilities, English language learners, and other student populations such as visually impaired are recruited to write and review items. The use of North Carolina educators to develop items strengthened the instructional and face validity of the items. Teachers and educators are recruited as needed. Item writing training for the item tryout and field test administrations occurred using a face-to-face format.

The NC Education Moodle system was introduced in 2011–12 allowing for virtual training. Depending on the event and the experience of the group that is being asked to write and review, training may be best applied in a face-to-face session. However, the majority of training is designed to be delivered in self-directed online training modules. To be included in the potential item writer or reviewer pool, teachers and educators from North Carolina were asked to visit https://center.ncsu.edu/nc/x_courseNav/index.php?id=21and take the appropriate subject-area "A" level Content Standards Overview course and the "B" level Test Development Basics course in the Moodle system. The "A" level subject course cover two main topics. The first section presents an overview tutorial that unpacks the NC*ESS* for the specific content area. This is intended to broaden their understanding of the content standards and the areas of interest. The second section of the tutorial provides trainees with an overview of Webb's DOK and Webb's alignment model adopted by the NCDPI as a tool to help them develop test questions that closely agree with the NC*ESS*. The "B" level course is designed as the next-level course for potential item writer/reviewers who have successfully completed the "A" level course. This course is presented under six main sections:

1. Test Development Process
2. Multiple-Choice Item Writing Basics
3. Fairness and Sensitivity
4. Security and Copyright
5. Using the Test Development System (TDS)
6. Next Steps

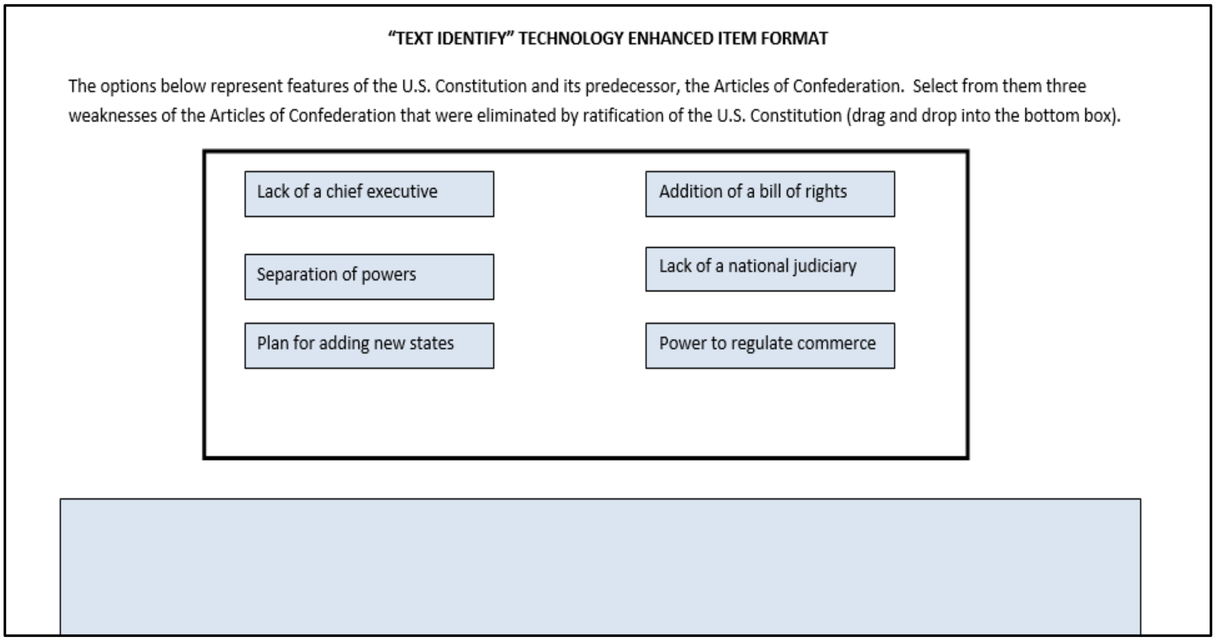
Once the online training courses are completed, teachers are directed to go to an online interest form at http://goo.gl/forms/wXv4Imh0ko. Here they can register to let the North Carolina Testing Program know they are interested in writing or reviewing items. Teachers who submit interest forms will be contacted when item writing or reviewing is needed in their subject area. For a complete description of the item writer training process and links to the training courses see Appendix 3-D Test Development Process_Teachers_6-2-15.

3.3.3 Usability Study for Technology-Enhanced Items

As a part of the Accountability and Curriculum Reform Effort (ACRE) initiative and the redesign of the EOG and EOC assessments, the NCDPI conducted a usability study on new item types with the goal of making assessments more authentic and engaging to students. The usability study for science was on computer based TE items. The evaluation criteria centered on aspects of accessibility, user-friendliness, and authenticity of construct measured. During the exploratory phase of science online tests, the NCSTP looked at two types of TE items, DD and TI, and their functions.

While the TE items hold promise to improve student engagement and the appeal of the assessment, they do require extra development safeguards to ensure that the items appear and function as intended while minimizing the introduction of construct-irrelevant variance. Also, there needs to be evidence that the scoring protocol is accurate and all responses are scored properly and that students with fewer computer skills are not disadvantaged. Figure 3-3 shows an example of a TI item with a stem and multiple options. Students are instructed to read the stem then identify the correct text provided by clicking on all correct options.

Figure 3-3 Text Identify TE Item Example



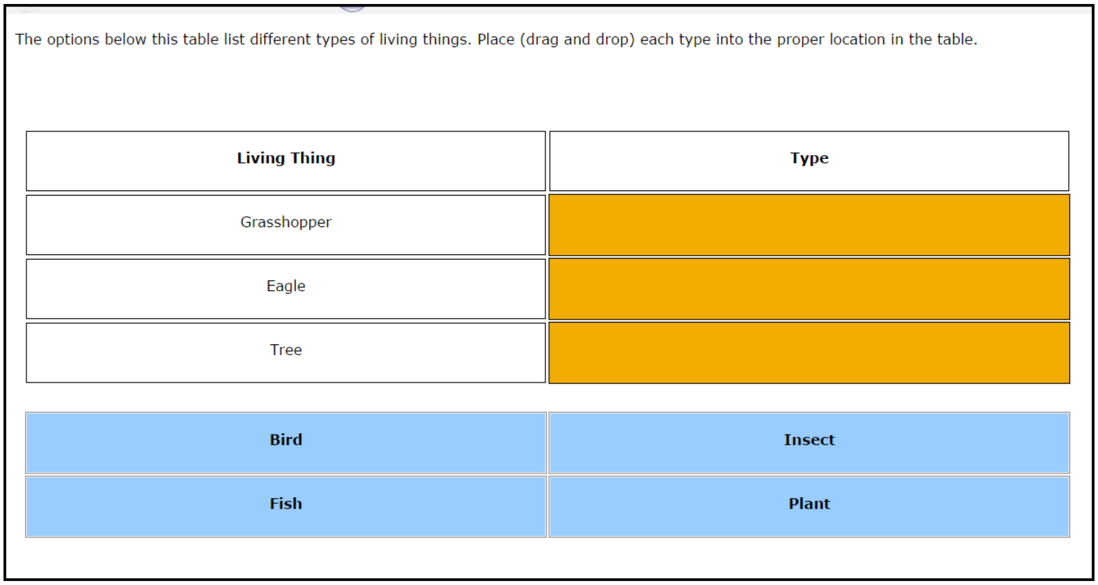
“TEXT IDENTIFY” TECHNOLOGY ENHANCED ITEM FORMAT

The options below represent features of the U.S. Constitution and its predecessor, the Articles of Confederation. Select from them three weaknesses of the Articles of Confederation that were eliminated by ratification of the U.S. Constitution (drag and drop into the bottom box).

Lack of a chief executive	Addition of a bill of rights
Separation of powers	Lack of a national judiciary
Plan for adding new states	Power to regulate commerce

An example of the drag-and-drop item is shown in Figure 3-4. In this type of item, students drag-and-drop correct options as answer into different containers.

Figure 3-4 Drag-and Drop TE Item Example



The options below this table list different types of living things. Place (drag and drop) each type into the proper location in the table.

Living Thing	Type
Grasshopper	
Eagle	
Tree	

Bird	Insect
Fish	Plant

A TE Item Usability Study (TEUS) for science was conducted by observing a sample of students in a class involving ten students in Grade 5, six in Grade 8, and five in high school Biology. Table 3-9 shows the usability study process in detail. At the end of each session, evaluators went over a set of survey questions with each student. Evaluators also completed a second survey at the end of the study. The complete survey instrument is presented in Appendix 3-E TEUS Survey Questions_2011.

The observation results showed that most grade 5 students spent 1 to 2 minutes in reading directions. However, grade 8 and high school students spent 1 minute or less. Five out of ten students (50%) in Grade 5, three students (50%) in Grade 8, and one student (20%) in Biology reported directions being unclear or wordy, and did not follow directions correctly. Only three students (30%) in Grade 5 knew how to indicate answers, and the rest needed help in figuring out the drag-and-drop function in text identifier items and to know how to deselect a choice. In Grade 8 and Biology class, fewer students (two from each grade level) turned to facilitators for help.

Table 3-9 Technology-Enhanced Items Usability Process

Step	Purpose	Time (minutes)
1. Introductions	Introduce student to evaluator.	3–5
2. Ice breaker activity	Set the student at ease and establish a friendly atmosphere.	4–5
3. Overview of session	Preview the session. Provide directions.	3–5
4. Present item 1	Protocol 1. Evaluator begins recording 2. Present item and ask student to read directions and answer question 3. Student interacts with test question 4. Evaluator observes and takes notes 5. Evaluator stops recording when student is finished	7–10
5. Present items 2–4	• *Repeat protocol with question 2–4*	7–10
6. Conclusion	• Present survey questions. • Replay recording of interaction and ask the student what they were thinking during certain parts of the interaction. • Thank the student for their feedback and participation.	5–15
	TOTAL	35–60

During the test, most students reacted well to the scroll bar, and only a few students (around 30 to 40% from each grade level) either did not realize there was a scroll bar or did not know how to use it to see all the choices. Most intervention was provided when students were dealing with the drag-and-drop function as well as the scroll bar.

On the survey question that asked whether the test questions were confusing or unclear, some minor technical issues were reported. One student from Grade 5 reported an issue with the drag-and-drop function, and the other one from Grade 8 reported the mouse jumping around and causing unintentional scrolling. When accessing the items, the only problem reported was the use of the scroll bar. Other than that, answers stored correctly and scoring worked correctly in these three grades.

The results from the survey showed that, in general, most students reacted positively to the TE items. Some of them thought the TE items were easier than the other multiple-choice items. One Grade 8 student thought that "it was cool how they moved," and one in Grade 5 said it was a good way to take a test. Some treated the new item type the same as other test questions. Only one student from Grade 5 reacted to the TE items impatiently, because it slowed him down with scrolling issues. The usability study allowed NCDPI to observe students interacting with these new items and provided valuable feedback on the improvement, design, and selection of TE items.

3.3.4 Item Tryout

In spring 2011, the NCDPI conducted an online item tryout for EOG Science Grade 5 and Grade 8 as well as EOC Biology with a purpose to evaluate new item types and assessments delivered via the new computer platform. As a part of the item tryout, at the end of the assessment, students were asked to respond to a short survey about their experience interacting with the test questions, their preferences regarding online assessments, and their online experiences outside of summative assessments. The gender and ethnicity distributions of the respondents are shown in *Table 3-10*. The survey recorded 4202 respondents for grade 5, 3734 for grade 8, and 2331 for Biology.

The grades 5 and 8 Science and high school Biology computer-based assessments consisted of traditional MC and TE item types. Results of the student survey questions dealing specifically with TE item types were mixed (see *Table 3-11* and *Table 3-12*). In general, students reported that their experience with computer tests was positive (69% agreed in grade 5, 58% in grade 8, and 54% in Biology). Less than half of the students responded positively when asked if they liked the new item types (45% in grade 5, 37% in grade 8, and 36% in Biology). The balance of responses were distributed across "Neutral," "Disagree," and "Did Not Respond" categories. When students were asked if the new types of test questions on this test were easy to understand, responses varied, but 36% in grade 8 and 45% in Biology agreed that they were. In grade 5, however, the largest proportion of students (42%) disagreed with it. For the Biology assessments (*Table 3-12*), the largest proportion of the students (44%) liked the new item types better than multiple-choice, and clicking and dragging worked well for 73% of the students.

Table 3-10 Demographic Characteristics of the Students Who Took the Survey

Grade	Demographic Characteristics		Frequency	Percent
Grade 5 (Total = 4,202)	Ethnicity	White	2283	54%
		Black	970	23%
		Hispanic	542	13%
		Asian	133	3%
		American Indian	132	3%
		Multiple	142	3%
	Gender	Female	2127	51%
		Male	2075	49%
Grade 8	Ethnicity	White	1517	41%
(Total = 3,734)		Black	1346	36%
		Hispanic	548	15%
		Asian	121	3%
		American Indian	61	2%
		Multiple	139	4%
		Pacific Islander	2	0.05%
	Gender	Female	1895	51%
		Male	1839	49%
Biology	Ethnicity	White	1326	57%
		Black	650	28%
(Total = 2,331)		Hispanic	180	8%
		Asian	92	4%
		American Indian	22	1%
		Multiple	61	3%
	Gender	Female	1184	51%
		Male	1147	49%

Table 3-11 Preference of Item Types / Test Modes – EOG Science

	Grade 5				Grade 8			
	Agree	Neutral	Disagree	Did Not Respond	Agree	Neutral	Disagree	Did Not Respond
I liked taking this kind of test on the computer.	2901 (69%)	677 (16%)	550 (13%)	74 (2%)	2176 (58%)	656 (18%)	842 (23%)	60 (2%)
I liked the new types of test questions that were introduced on this test.	1905 (45%)	1289 (31%)	820 (20%)	188 (4%)	1365 (37%)	1074 (29%)	1154 (31%)	141 (4%)
The new types of test questions on this test were easy to understand.	1197 (29%)	1060 (25%)	1775 (42%)	170 (4%)	1358 (36%)	1088 (29%)	1148 (31%)	140 (4%)

Table 3-12 Preference of Item Types / Test Modes – EOC Biology

	Agree	Neutral	Disagree	Did Not Respond
I liked taking this kind of test on the computer.	1253 (54%)	559 (24%)	479 (21%)	40 (2%)
I liked the new types of test questions that were introduced on this test.	850 (36%)	709 (30%)	708 (30%)	64 (3%)
The new types of test questions on this test were easy to understand.	1053 (45%)	674 (29%)	541 (23%)	63 (3%)
I liked the new types of questions on this test more than the usual multiple-choice type questions.	1036 (44%)	652 (28%)	579 (25%)	64 (3%)
Test questions that required clicking and dragging a word to a location on the screen worked well.	1698 (73%)	294 (13%)	275 (12%)	64 (3%)

Regarding students' spending time on electronic devices, most students reported that they spend about one to four hours a day (65% of grade 5, 73% of grade 8, and 76% of Biology students) using a computer or related products in all three grades (see *Table 3-13*). Students who did not spend time on any electronic devices amounted to 21% or fewer across grades.

Table 3-13 About how many hours per day do you usually spend using a computer and/or video game console?

Hours Spent in Computer Related Activities	Grade 5		Grade 8		Biology	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
0	890	21%	240	13%	265	11%
1 to 4	2744	65%	1363	73%	1772	76%
5 to 10	326	8%	169	9%	169	7%
Greater than 10	134	3%	38	2%	64	3%
Did Not Respond	108	3%	48	3%	61	3%

Students were also asked to provide information about any prior experience with computers for academic use (Table 3-14 and Table 3-15). The majority of grade 8 (54%) and HS Biology (69%) students indicated that they turned in their homework using a computer. Similarly, 90% or more of the students have used handheld electronic devices in all three grade levels. HS Biology students also frequently used social networking services (86%) and online courses (46%) as a part of their experience with electronic devices (see Table 3-15).

Table 3-14 Past Experience with Computer – EOG Science

Survey Questions	Grade 5			Grade 8		
	Yes	No	Did Not Respond	Yes	No	Did Not Respond
Have you turned in classwork or homework assignments using a computer?	1656 (39%)	2425 (58%)	121 (3%)	2019 (54%)	1608 (43%)	107 (3%)
Have you used any handheld electronics at school such as clickers, calculator, etc.?	3799 (90%)	290 (7%)	113 (3%)	3427 (92%)	209 (6%)	98 (3%)

Table 3-15 Past Experience with Computer – EOC Biology

Survey Questions	Yes	No	Did Not Respond
Have you turned in classwork or homework assignments using a computer?	1604 (69%)	649 (28%)	78 (3%)
Have you used any handheld electronics at school such as clickers, calculator, etc.?	2135 (92%)	119 (5%)	77 (3%)
Have you used social network services (e.g., Facebook, MySpace, etc.)?	2007 (86%)	246 (11%)	78 (3%)
Have you taken a course online or do you plan to take one in the near future?	1080 (46%)	1175 (50%)	76 (3%)

Table 3-16 summarizes technical issues students experienced while completing TE items during the tryout. Thirty-five percent or less of the students indicated experiencing some sort of technical issues. Highlighting text was the most common issue reported in 5th grade (35%), followed by "Clicking on answer choice" (18%) and "Clicking on buttons or using tools" (16%). The same pattern is true for grade 8 with 21%, 16%, and 15% respectively. In HS Biology, the highest proportion of students (17%) indicated "Moving between pages/questions" as the biggest technical issue followed by "Highlighting text" (16%).

Table 3-16 Please check any of the features you had problems using.

Technical Issues	Grade 5		Grade 8		Biology	
	N	Percent	N	Percent	N	Percent
Moving between pages /questions	635	15%	771	21%	389	17%
Clicking on buttons or using tools	689	16%	556	15%	211	9%
Clicking on answer choices	739	18%	611	16%	209	9%
Scrolling within a question	359	9%	346	9%	175	8%
Highlighting text	1462	35%	787	21%	368	16%

Among the respondents, 61% of Grade 5, 58% of grade 8, and 47% of HS Biology students preferred online tests over paper and pencil tests for Science (*Table 3-17*). Only 19% or less of students in each grade indicated “No”.

Table 3-17. For this subject, do you feel that online tests are better than paper-and-pencil tests?

	Grade 5		Grade 8		Biology	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Yes	2552	61%	2152	58%	1096	47%
No	678	16%	599	16%	451	19%
Did Not Respond	972	23%	983	26%	784	34%

3.3.5 Item Difficulty

For the purposes of guiding item writers to provide a variety of items, they were instructed to classify items into three expected levels of difficulty: easy, medium, and hard. Easy items are defined as items that the item writers expect will be answered correctly by approximately 70% or more examinees. Medium items are expected to be answered correctly by 40–70% of the examinees. Hard items are expected to be answered correctly by approximately < 40% of the examinees.

The item writers were further instructed to write approximately 25% of their items at the hard level, 25% at the easy level, and the remaining 50% at the medium level of difficulty. These targets are used to replenish item pools ensuring an adequate range of difficulty. It is important to note that these levels of difficulty are based solely on the judgment of item writers and are not empirically derived. Actual item difficulty as defined by the actual proportion correct under field test and operational test conditions will be presented in Chapter 4.

In addition to expected difficulty, item writers also considered the cognitive rigor or DOK in terms of recall and reproduction, skills and concepts, strategic thinking, and extended thinking required to answer each item. This ensures a balance of difficulty as well as a balance across the different cognitive levels among the items in the North Carolina EOG and EOC assessments.

3.3.6 Item Alignment

A critical aspect of item quality is alignment. Alignment refers to the extent to which an item agrees with and represents the content standard it is designed to measure. Assessments composed of items that are misaligned will generate scores that do not measure the breadth and depth of the intended construct. Scores from a misaligned assessment are characterized by high construct-irrelevance variance and will underestimate or overestimate students' achievements. For this reason, alignment evidence is one of the most important sources of content validity.

During the item development phase, two groups were responsible for item alignment: 1) content specialists at the North Carolina State University-Technical Outreach for Public Schools (NCSU-TOPS) and 2) members of the NCDPI/K-12 Curriculum and Instruction Division[d]. These groups independently reviewed proposed items through NC's online item writing system, the Test Development System (TDS) and classified them by the NCE*SS* and DOK levels. Any items with discrepant classifications were prevented from continuing through item development until the discrepancy was resolved.

3.3.7 Item Format

The Grades 5 and 8 Science and Biology assessments consist of traditional four-foil MC items in Paper forms and MC as well as two types of TE items in computer-based forms. The two types of TE items referenced in the usability studies that were developed for the EOG and EOC forms are: TI and DD. For examples of these item types, please refer to Figure 3-3 and Figure 3-4 in Section 3.3.3.

3.4 Step 9–Field Test Item Review

To ensure that items were developed in alignment with the NC*ESS* standards, each item went through a detailed review process before being placed on a field test. The following *Standards* (AERA/APA/NCME, 2014) state the need of testing process and minimizing construct irrelevant variance:

[d] The NCDPI/test development created an alignment plan in 2010 before the development of any items. The alignment plan was reviewed by an expert in content alignment, Dr. Karen Hess, from the Center for Assessment. Based on her recommendations, an alignment plan was devised that would pre-align test items to the NC*SCS* content standards.

> Standard 3.1—Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population.
>
> Standard 3.2—Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

A separate group of North Carolina educators were recruited to review all items. Once items had gone through educator review, test development staff members, with input from curriculum specialists also reviewed each item. Items were further reviewed by educators and/or staff familiar with the needs of students with disabilities, English Language Learners and students with visual impairments. This review addresses concerns due to bias or sensitivity issues, such as contexts that may elicit an emotional response, inhibit a student's ability to respond, or may be unfamiliar to a student for cultural or socio-economic reasons.

The criteria for evaluating each written item included the following:

1. **Conceptual**

   - Objective match (curricular appropriateness)
   - Webb's Depth-of-Knowledge match
   - Fair representation
   - Lack of bias or sensitivity
   - Clear statement
   - One best answer
   - Common context in foils
   - Credible foils
   - Technical correctness

2. **Language**

   - Appropriate for age
   - Correct punctuation
   - Spelling and grammar
   - Lack of excess words

- No stem or foil clues
- No negative in foils (unless it fits the objective)
- Readability is grade-level appropriate
- Idioms and two-word verbs do not inhibit accessibility for ELL students

3. Format

- Logical order of foils
- Familiar presentation style, print size, and type
- Correct mechanics and appearance
- Equal/balanced length foils

4. Diagram/Graphics

- Necessary
- Clean
- Relevant
- Unbiased
- Accessibility for visually impaired students
- Ability to be Brailed

3.5 Steps 10/11–Field Test Forms Assembly and Review

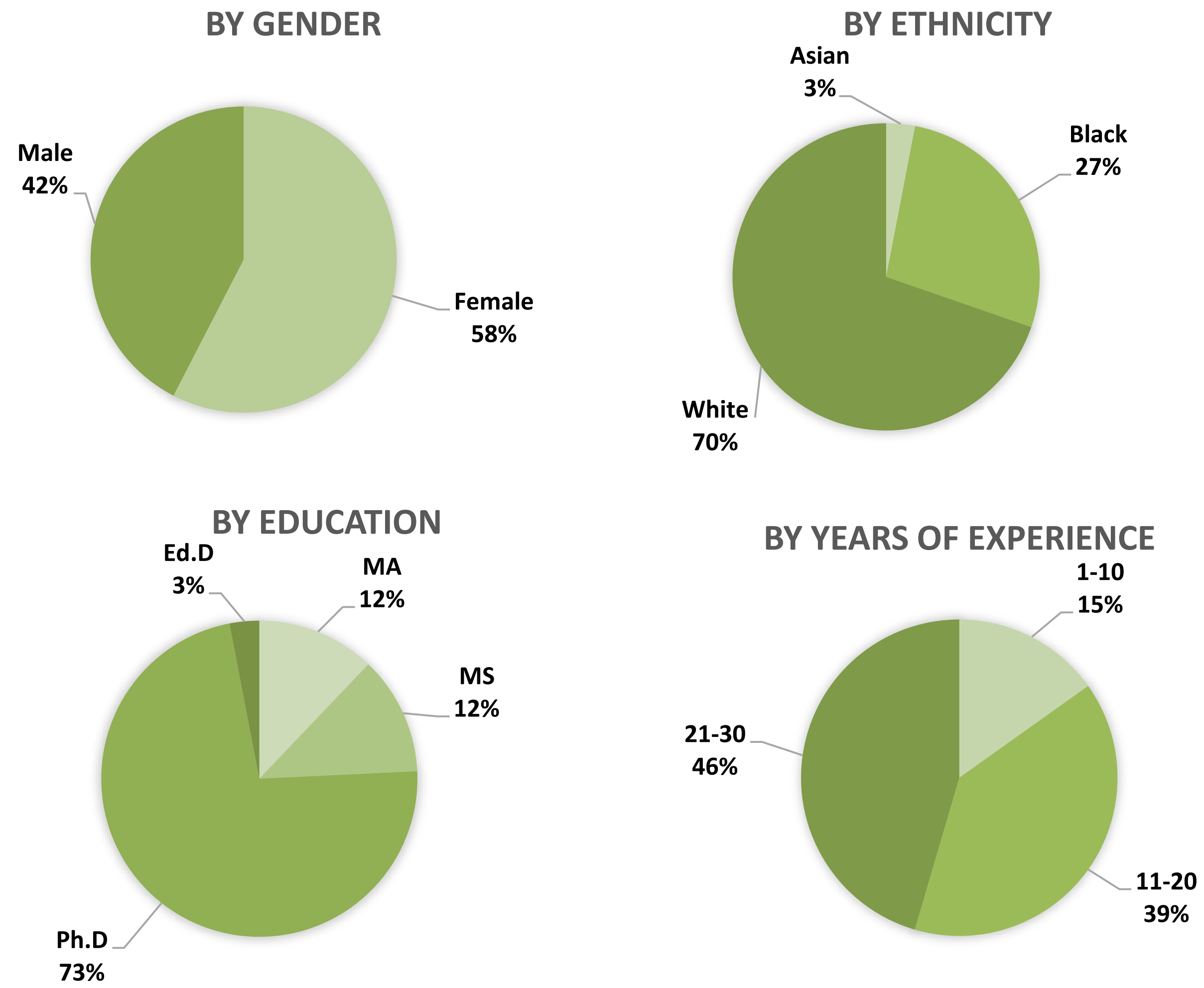
Items for each grade level were assembled into field test forms based on the assessment content specification and blueprint. Field test forms were organized according to the blueprints to be implemented for the operational assessment. Table 3-18 shows the number of forms, number of items in each form, and total number of items administered in the 2011 – 2012 stand-alone field test. Before the field test administration, outside content reviewers, following steps similar to operational form review, reviewed the assembled field test forms for clarity, correctness, potential bias or sensitivity, cuing of items, and curricular appropriateness.

The outside content reviewers were recruited by NCSU-TOPS from a pool of educators who have had no prior role with item writing or reviewing. In all, 33 outside content specialists from different subject areas (e.g. ELA/Reading, Math, and Science) have served as external form reviewers during this EOG and EOC test cycle. Descriptive summaries of their demographic and educational background are shown in the pie charts in *Figure 3-5*. These experts provided an independent outside evaluation of the forms. All the form reviews were done using the NCDPI's online test development system (TDS). All comments were recorded and reviewed and any issues addressed before the forms were administered.

Table 3-18 Number of Items Field Tested for Science EOG and EOC

Grade /Course	Administration(s)	Number of Forms	Number of Items per Form	Total Number of Items (Unique Items)
Grade 5	Spring 2012	8	60	480 (415)
Grade 8	Spring 2012	8	60	480 (425)
Biology	Spring 2012	10	75 (60OP+15FT)	60OP+150FT (400)

Figure 3-5 Demographic Information for Outside Form Reviewers



Chapter 4 Field-Test Administration and Operational Form Construction

This chapter describes the field test administration, including the sampling plan enacted to ensure that each form was administered to a representative sample of students. In addition, the chapter describes the psychometric analyses conducted on the field test data and the steps taken to construct the operational test.

4.1 Step 12–Field Test Sample and Administration[e]

Sampling for 2011–12 field testing of the North Carolina science assessments was accomplished using stratified random sampling at school level, with the goal being a selection of students within schools that were representative of the entire student population in North Carolina. The following stratifying variables were used to ensure the final sample was representative:

- Gender
- Ethnicity
- Region of the state
- Economically disadvantaged classification (based on free/reduced lunch program enrollment)
- Students with disabilities
- Students with limited English proficiency
- Previous year's test scores

Comparative descriptive statistics of the respective population and the field test sample across the various stratifying variables are shown in *Table 4-1* to comply with Standard 1.8 of the AERA/APA/NCME (2014) Standards, which states:

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

[e] NCDPI employs the same administration procedures for the field test and the operational assessment. Please see Chapter 5 for a detailed discussion of NC's administration procedures.

Table 4-1 Demographic Summary for Science Field Test 2012 Sample Participants

Category		Grade Level					
		5		8		Biology	
		Population	Sample	Population	Sample	Population	Sample
	N	117,975	21,377	112,668	20,991	120,496	21,765
Gender (%)	Female	49.4	49.8	49.3	49.6	49.5	49.1
	Male	50.5	50.2	50.6	50.4	50.1	50.9
Ethnicity (%)	Asian	2.6	2.5	2.5	2.3	3.0	8.2
	Black	26.1	25.6	26.6	25.4	28.4	14.5
	Hispanic	14.0	13.7	11.7	11.4	10.7	9.6
	White	52.1	53.4	53.8	55.8	52.7	60.4
	Other	5.2	4.8	5.3	5.1	4.8	7.3
Special Population (%)	ELL	5.8	5.9	4.5	4.6	3.3	1.5
	SWD	9.5	8.7	8.7	8.1	9.9	11.1
	EDS	56.6	52.8	52.8	47.6	46.5	29.7

ELL=English Language Learner, SWD=Student with disability, EDS=Economically Disadvantaged

Table 4-1 shows comparisons of the proportions of students selected for the field test sample against the total population. The desired sampling rate was set at 15% from each grade level. After attrition, the effective sample for grade 5 was 21,377, grade 8 was 20,911, and Biology was 21,765. Demographic proportions from the field test sample and population across the respective grades show a very similar distribution across the major demographic variables, except in Biology where the proportion of white students in the sample was about 7% more than in the population and that of black students was about 14% less in the sample. In terms of special population categories, the field test samples are representative of the population distribution for ELL, SWD, and Economically Disadvantaged students. Overall, the field test sample is representative of North Carolina students at the respective grade levels, and sample statistics can be generalized and interpreted to reflect population parameters within a reasonable amount of sampling error.

4.2 Step 13–Field Test Item Analyses

Field test data analyses provided statistical evidence used to determine whether items were retained for use on an operational North Carolina EOG or EOC form. Three main statistical methods were used to conduct item analysis from the field test: Classical Test Theory (CTT), Item Response Theory (IRT), and Differential Item Functioning (DIF) analyses. In addition, content experts conducted a qualitative review on all statistically flagged items. There are various qualitative and/or quantitative reasons items may be flagged, including multiple correct responses, no correct response, or statistical bias against certain student groups. Only those field test items demonstrating adequate statistical and content properties were considered for operational use.

4.2.1 Classical Analysis Summary of Field Test Items

Classical item analyses of the field test items were conducted in SAS and included evaluation of item p-value and item-to-total correlation (biserial) statistics to determine if items met NCDPI item quality criteria. Item p-value summarizes the proportion of examinees answering each item correctly and was used as an indicator of preliminary item difficulty. Valid ranges of p-values for multiple-choice items are between 0 and 1, where values close to 0 indicate extremely difficult items that very few students answer correctly and values close to 1 indicate very easy items that almost all students answer correctly. The general NCDPI rule is to keep items with a p-value range of 0.15 to 0.85.

The biserial correlation provides evidence of how well each item on a test form correlates with the form's total test score. It is a measure of item discrimination, or, in other words, a measure of how well an item differentiates high- and low-performing test takers. The general NCDPI rule is to keep items with a biserial value of 0.25 or higher. Any exception to this rule is made only for rare cases and with thorough vetting from the content experts and psychometricians. Items with negative biserials are not retained for use on the operational assessment. Table 4-2 shows descriptive statistics of p-values, biserials, and Omit rates from field test item pool.

Table 4-2 CTT Field Test 2012 Item Pool Descriptive Statistics for Science EOG and EOC

Grade	Item Types*	No. of Items	p-value Summary				Biserial Correlation				Omit
			Mean	SD	Min	Max	Mean	SD	Min	Max	%
5	MC	441	0.52	0.20	0.08	0.91	0.32	0.16	-0.26	0.68	0.11
	DD	19	0.40	0.22	0.10	0.87	0.38	0.16	-0.07	0.53	0.24
	TI	20	0.45	0.29	0.03	0.91	0.35	0.17	0.12	0.60	0.34
8	MC	441	0.55	0.19	0.10	0.97	0.37	0.16	-0.22	0.70	0.13
	DD	27	0.51	0.26	0.06	0.90	0.34	0.17	-0.04	0.66	2.25
	TI	12	0.17	0.13	0.04	0.49	0.32	0.11	0.06	0.44	0.57
Biology	MC	679	0.52	0.17	0.08	0.96	0.38	0.17	-0.31	0.68	0.15
	TE	71	0.39	0.23	0.00	0.78	0.41	0.20	-0.08	0.72	1.54

**MC=Multiple-Choice, DD=Drag & Drop, TI=Text Identify, TE=Technology Enhanced*

Results indicated that the mean p-values of the MC items are higher than the TE items and biserial correlations are reasonably high given the fact that all items in the pool, including items with negative biserial correlation, are included in the calculation. The Omit rate is low (2.25% or lower) for all grades and item types, with the higher Omit rates being for TE items. The criteria for inclusion in the operational forms are described in *Section 4.4.1*. Note that the items with p-value<0.10 and biserial correlation<0.15 were deleted from the operational-item selection pool.

4.2.2 Item Response Theory (IRT) Summary of Field Test Items

Item Response Theory (IRT) provided the main theoretical base for item calibration, form building, scoring, and scaling. NCDPI adopted the three-parameter logistic (3PL) unidimensional model to calibrate all multiple-choice items. Equation 4-1 presents the mathematical representation for the 3PL:

$$P_i(\theta) = c_i \frac{1-c_i}{1+\exp[-Da_i(\theta-b_i)]} \qquad (4\text{-}1)$$

where $P_i(\theta)$ is the probability that a randomly chosen examinee's given ability answers item *i* correctly (this is an S-shaped curve with values between 0 and 1 over the ability scale); a_i is the slope or the discrimination power of the item; b_i is the threshold or “difficulty parameter of an item; c_i is the lower asymptote or pseudo-chance level parameter; and D is a scaling factor of 1.7.

The IRT parameter estimates were calibrated using IRTPRO software (Cai, Thissen, & du Toit, 2011) with the Bayesian prior distributions for the item parameter calibration set to a~lognormal (0, 1) and c~Beta (5, 15). For TE items, the Bayesian prior distribution of c~Beta (A, B) was set by dividing the number of possible response combinations for TE items. The use of the Bayesian prior distribution ensured appropriate parameter estimates of chance-scores were accounted for during calibration. Table 4-3 shows summary descriptive IRT parameter statistics from the field test item pool. Results indicated that some of the items exhibited less than optimal item statistics. The items flagged for a<0.50, b>3, and g>0.45 were excluded from the operational item selection pool.

Table 4-3 IRT Field Test 2012 Item Pool Descriptive Statistics for EOG Science and EOC Biology

Grade	Item Type	No. of Items	Slope(a)				Threshold(b)				Asymptote(g)			
			Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
5	MC	418	1.25	0.63	-2.20	5.19	0.63	1.61	-3.44	13.46	0.22	0.07	0.06	0.51
	DD	19	0.76	0.84	-1.84	1.51	0.09	1.67	-3.95	2.07	0.08	0.07	0.00	0.21
	TI	20	1.22	0.49	0.55	2.37	0.82	2.01	-2.11	5.37	0.09	0.06	0.02	0.26
8	MC	414	1.34	0.60	-2.49	4.14	0.26	1.20	-3.18	4.19	0.22	0.07	0.05	0.51
	DD	21	1.01	0.55	0.34	2.46	-0.26	1.24	-2.32	1.99	0.20	0.20	0.01	0.55
	TI	9	1.13	0.44	0.62	1.92	2.15	0.66	0.67	2.96	0.07	0.06	0.00	0.15
Biology	MC	667	0.90	0.35	0.07	2.26	0.64	1.33	-2.55	9.46	0.23	0.07	0.08	0.50
	TE	66	1.49	3.92	0.13	28.33	0.98	1.61	-0.96	8.18	0.13	0.08	0.01	0.34

4.2.3 Differential Item Functioning

As the developer of the NC assessments, it is the responsibility of NCDPI to examine all assessment items for possible sources of bias. Standard 3.3 of the *Standards* (AERA, APA, &NCME, 2014) states, “Those responsible for test development should include relevant subgroups in validity, reliability/precision, and other preliminary studies used when constructing the test" (p. 64). Differential item functioning (DIF) measures statistical bias by examining the degree to which members of various groups (e.g., males versus females) perform differentially on an item. It is expected that groups of students with the same ability will have similar probability for answering items correctly, regardless of background characteristics. An item is considered as exhibiting DIF when students who are members of different subgroups but have approximately equal knowledge and skill on the overall construct being tested perform in substantially different ways (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014). It is important to remember that the presence or absence of true bias is a qualitative decision, based on the content of the item and the curriculum context within which it appears. NCDPI utilizes DIF statistics to quantitatively identify suspect items for further scrutiny.

NCDPI use the Mantel-Haenszel statistic and ETS Delta classification codes for flagging candidate DIF for multiple-choice items (Camilli & Sheppard, 1994). The Mantel-Haenszel (MH) chi-square statistic tests the alternative hypothesis that a linear association exists between the row variable (score on the item) and the column variable (group membership). The Mantel-Haenszel odds ratio is computed using the CMH option in PROC FREQ Procedure in SAS.

$$\boldsymbol{\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}} \qquad \textbf{(4-2)}$$

Where at each level of j (each item studied),

Group	Score on Studied Item		Total
	1	0	
Reference (R)	A_j	B_j	n_{Rj}
Focal (F)	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Transforming the odds ratio by the natural logarithm provides the DIF measure, such that:

$$\beta_{MH} = \log_e(\alpha_{MH}) \qquad \textbf{(4-3)}$$

The ETS classification scheme first requires rescaling the MH value by a factor of -2.35 providing the Delta (*D*) statistic as follows:

$$|D| = -2.35\beta_{MH} \qquad \textbf{(4-4)}$$

Items are then classified based on their Delta statistic into three categories:

- 'A' items are not significantly different from 0 using $|D| < 1.0$. No substantial difference between the two groups on item performance is found for items with A+ or A- classifications.
- 'B' items significant from 0 and either D not significantly greater than 1.0 or $|D| < 1.0$. An item with a B+ rating marginally favors the focal group (Females, African Americans, Hispanics, or Rural students). Item with a B- rating disfavors the focal group (favors Males, Whites, or Non-rural students,).
- 'C' items have D significantly greater than 1.0 and $|D| \geq 1.5$. An item with a C+ rating favors the focal group (females, African Americans, or Hispanics, Rural, EDS). Item with a C- rating disfavors the focal group (favors males, whites, rural, EDS).

Table *4-4* shows field test pool multiple-choice items by candidate DIF flag. During the initial construction of EOG and EOC assessments in 2011, the NCDPI investigated DIF for gender —male and female with male set as the reference group and female the focal group—and two ethnicity categories: "White" versus "Black" and "White" versus "Hispanic." In both ethnic categories, "White" was set as the reference group and "Black" and Hispanic" were the respective focal groups. For example, for EOG Science Grade 5, females performed somewhat better on 217 items compared to males of similar ability, and males performed somewhat better on 244 items compared to females of similar ability. Twelve items showed marginal DIF (B) in favor of females and six showed marginal DIF in favor of males. One item showed significant DIF, in favor of males. The rest of the table is interpreted in a similar fashion. NCDPI's rule is to remove all items with a DIF flag of "C" from the item bank and "B" items are sent for further review and only placed on an operational form upon a positive review from the bias panel, providing a replacement item is not readily available for that content domain. Across all grades the most "C" DIF items were flagged for the "White" versus "Hispanic" category.

Table 4-4 Mantel-Haenszel Delta DIF Summary for Science Field Test 2012

Grade	DIF Male/Female						DIF White/Black						DIF White/Hispanic					
	A+	A-	B+	B-	C+	C-	A+	A-	B+	B-	C+	C-	A+	A-	B+	B-	C+	C-
5	217	244	12	6		1	228	235	10	6		1	222	217	19	19	2	1
8	237	216	9	15	2	1	221	219	11	24	2	3	228	221	16	9		6
Biology*	355	389				1	357	379	5	3		1	362	370	9	3	1	

*5 Items were Technology Enhanced (DD and TI)

4.3 Step 14–Bias Review

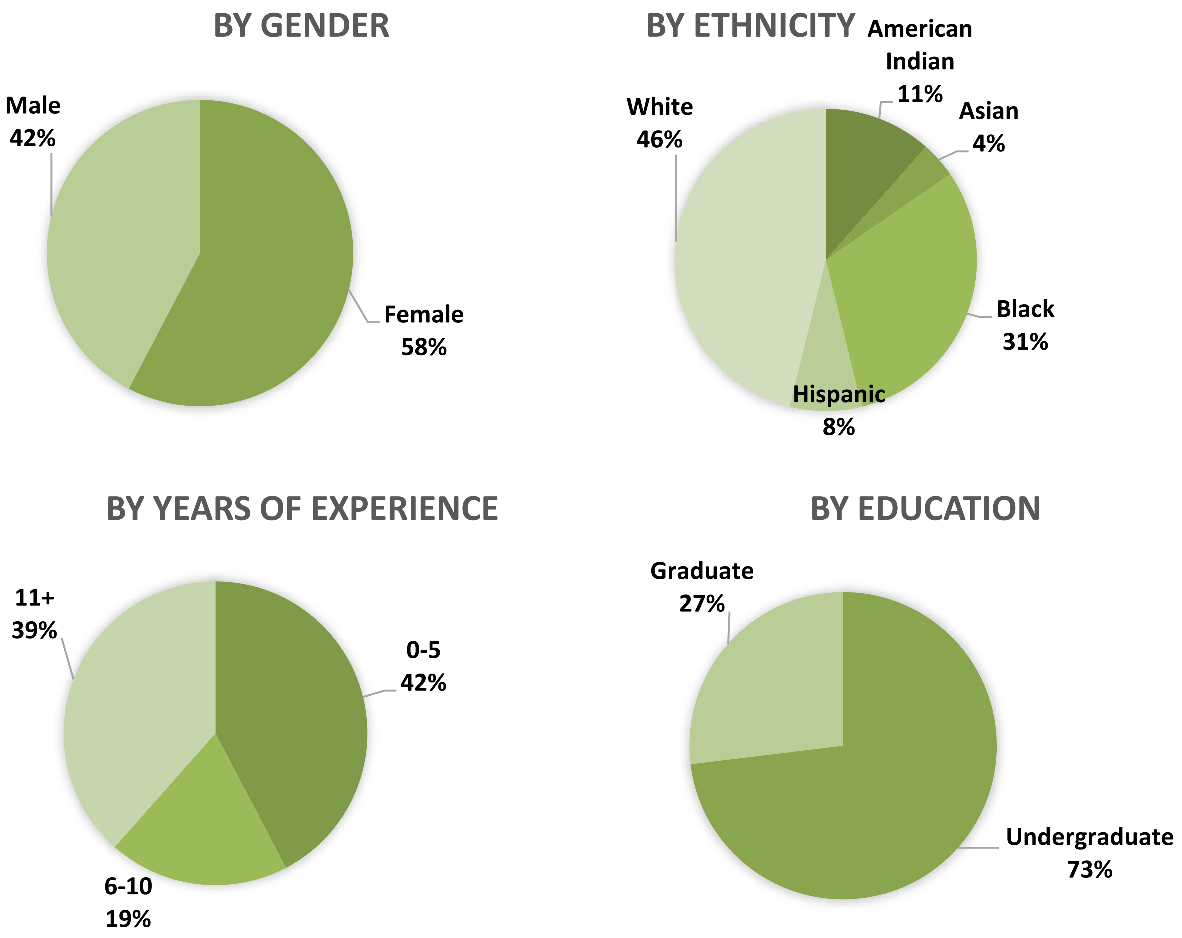
Fairness is an ongoing concern when administering and constructing a summative, statewide assessment. When constructing test forms, it is important to know the extent to which items perform differentially for various groups of students. The first step was flagging items for DIF. The second step was convening a bias review panel to examine all flagged items.

Standard 3.6 of the AERA/APA/NCME (2014) Standards states:

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

This standard puts responsibility on the test maker to examine all sources of possible construct-irrelevant variance. To meet this standard in terms of items flagged for DIF, NCDPI convenes Bias Review panels for each grade level. In this instance, the review panels were made up of 5 to 8 participants. Members were carefully selected based on their knowledge of the curriculum area and their diversity with respect to the student population. During the form building and review process for EOG and EOC in the 2011–2015 test development cycle, the NCDPI recruited a total of 26 reviewers to serve on the bias review panels. Their demographic information is illustrated in *Figure 4-1*.

Figure 4-1 Demographic Information for Bias Review Panels from 2011-2015



Before reviewing items, panelists had to complete an online, bias-review training process through the NC Review System see Appendix 4-A Bias and DIF Review Process for an overview of this process. Only "B"-flagged items were reviewed, all "C"-flagged items were removed from the item bank. For each item flagged as "B," panelists were asked to evaluate the item based on the following questions:

- Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
- Does the item contain any local references that are not a part of the statewide curriculum?

- Does the item portray anyone in a stereotypical manner? (This could include activities, occupations, or emotions.)
- Does the item contain any demeaning or offensive materials?
- Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
- Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
- Does the artwork adequately reflect the diversity of the student population?
- Are there other bias or sensitivity concerns?

The online review platform required that if there were any indication that the reviewer suspected an item was associated with a bias, sensitivity, or accessibility issue then he/she were to explicitly document their concern.

Following the review of all flagged items by the panels, a final determination had to be made whether to retain or delete any of these items from the operational item pool. Items that were flagged both for DIF category B and received an affirmative response to any of these questions asked during bias review or were commented on by the review panel were further reviewed and agreed upon by content specialists at the NCDPI and NCSU-TOPS. These experts included, at a minimum, the Test Measurement Specialist, Psychometrician, and Lead Content Specialist at NCSU-TOPS. These items were only included on an operational form if no other viable alternative was available in the item bank, all experts agreed the items measured content that was expected to be mastered by all students, and no obvious indication of specific construct-irrelevant variance is detected. The general rule was that all DIF C flagged items were exempted from the operational pool.

4.4 Timing Analyses from Field Test Administration

In keeping with the standards of fairness and to ensure standard administration so scores are comparable, the NCDPI conducted a timing analysis during the field test to set reasonable expectations of how long each assessment would take students to complete. The EOG and EOC assessments were not designed to be a timed test, but for practical reasons the NCDPI intended to use the data to set reasonable timing guidelines which would comply with standard 4.14—"For a test that has a time limit, test development research should examine the degree to which scores include a speed component and should evaluate the appropriateness of that component, given the domain the test is designed to measure" (p. 90).

During the field test, students' start and end time data were recorded. Summary data of how long it took students to complete each test is shown in Table 4-5. The table includes data for science EOG and EOC assessments administered under regular conditions; that is, with no accommodations of extended time, multiple test sessions, testing in a separate room, or special NCDPI-approved accommodations. Preliminary analysis showed that 6.5% of students in grade 5, 7.3% of students in grade 8, and 6.3% of students in Biology submitted their papers within 15 minutes. Similarly, 1.7% of students in grade 5, 2% of students in grade 8, and 1.3% of students in Biology took more than 300 minutes. These students were considered outliers for the timing study and were dropped from the analysis. The results indicated that the median times taken to complete the tests were 54, 41, and 43 minutes for grade levels 5, 8, and Biology respectively. Moreover, 95% of the grade 5 students completed their tests in 105 minutes, while grade 8 and Biology students completed in 74 minutes.

Based on these estimates and other practical considerations, the NCDPI recommended time allotted for the EOG Science be 180 minutes. The estimated time allotted for EOC Biology is 150 minutes. In keeping with standards of equity, the NCDPI requires all students participating in the assessments be allowed ample opportunity to complete the assessments as long as they are engaged and working and the maximum time allowed (i.e., 240 minutes) has not been reached. This is consistent with the *Standards* (2014, p. 51) which states, "although standardization has been a fundamental principle for assuring that all examinees have the same opportunity to demonstrate their standing on the construct that a test is intended to measure, sometimes flexibility is needed to provide essentially equivalent opportunities for some test takers." Given that the construct measured in EOG and EOC is not speeded, the NCDPI is allowing students up to four hours for EOG science and three hours for EOC Biology to complete the assessments in a single session. Students with approved accommodations can take even longer, as specified by their particular Individualized Education Program (IEP) or LEP plan.

Table 4-5 Science EOG and EOC Recorded Test Duration from Field Test 2012

EOG/EOC	N	Summary			Percentile				
		Number of Items	Avg.	SD	25th	Median	75th	95th	99th
Grade 5	17,945	60	57.89	24.95	41	54	70	105	141
Grade 8	17,803	60	43.61	16.37	32	41	53	74	96
Biology	20,866	75	45.17	16.98	34	43	53	74	97

4.5 Step 15–Operational Test Construction

The field testing plan was designed to generate enough items to construct four equivalent forms for EOG Science Grades 5 and 8 and EOC Biology. The use of multiple forms at each grade level ensures that a broader range of the content domain can be assessed at the breadth and depth required by the content standards. The justification for adopting multiple forms is that the adopted NC *Essential Standards* for Science are extremely rich; therefore, a single test form that fully addresses all competencies would be prohibitively long. Additionally, the use of multiple forms spiraled within a classroom reduces the incidence of test irregularities at the classroom level resulting from students copying. For the EOG at grades 5 and 8 and the Biology EOC, both computer-based and paper-based fixed forms were created. The paper-based fixed form is an exact replicate of the computer-based fixed form with the exception of the TE items. For each grade level, one form was selected and published as a released form on the NCDPI website. The released forms were available to teachers, students, and all interested stakeholders so they could be familiarized with the new assessment before the operational administration. Online versions were offered through the same platform students will use during the summative assessment.

4.5.1 Criteria for Item Inclusion in Operational Pool

Standard 3.2 of the *Standards* states that:

> *Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-*irrelevant *characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics (p. 64).*

Following the field test administration participating teachers completed an online item review of each item. The results for each item and comments were integrated in the NCDPI's online Test Development System. These feedback provided additional evaluative qualitative data for field test items. From a psychometric perspective, NCDPI carefully considers all items prior to their inclusion in the operational pool and the operational test form. All of the aforementioned item parameters were used to determine if items displayed sound psychometric properties to be used in operational forms. Field test items were classified into one of three category: "Keep," "Reserve," and "Delete" according to the following psychometric criteria:

- Items with these characteristics were flagged as "Delete" and removed from the item pool:
  - weak discrimination—the slope (*a* parameter) was less than 0.50
  - low correlation with total score—the item correlation (r-biserial) was less than 0.15
  - guessing—the asymptote (*c* parameter) was greater than 0.45
  - too difficult—the threshold (*b* parameter) was greater than 3.0 or the p-value was less than 0.10
  - DIF flag of C
- Items with these characteristics were used sparingly as "Reserve":
  - weak discrimination—the slope (*a* parameter) was between 0.50 and 0.70
  - low correlation with total score—the item correlation (r-biserial) was between 0.15 and 0.25
  - guessing—the asymptote (*c* parameter) was between 0.35 and 0.45
  - too difficult—the threshold (*b* parameter) was between 2.5 and 3.0 or the p-value was between 0.10 and 0.15
  - too easy—the threshold (*b* parameter) was between $^{-}2.5$ and $^{-}3.0$ or the p-value was between 0.85 and 0.90
- Items with these characteristics underwent additional reviews:
  - ethnic bias—the log odds ratio was greater than 1.50 or less than 0.67 (flagged "B")
  - gender bias—the log odds ratio was greater than 1.50 or less than 0.67 (flagged "B")

All other items not classified as "Delete" or "Reserve" were labeled as "Keep" and considered first choices during operational form construction. The number of items classified into the Delete, Reserve, and Keep categories are shown in Table 4-6. The table shows that over 60% of the items were classified as "Keep" or "Reserve," allowing a sufficient item pool for the construction of four parallel forms in Grades 5 and 8 EOGs and the Biology EOC assessments.

Table 4-6 Field Test 2012 Item Pool Summary for Science

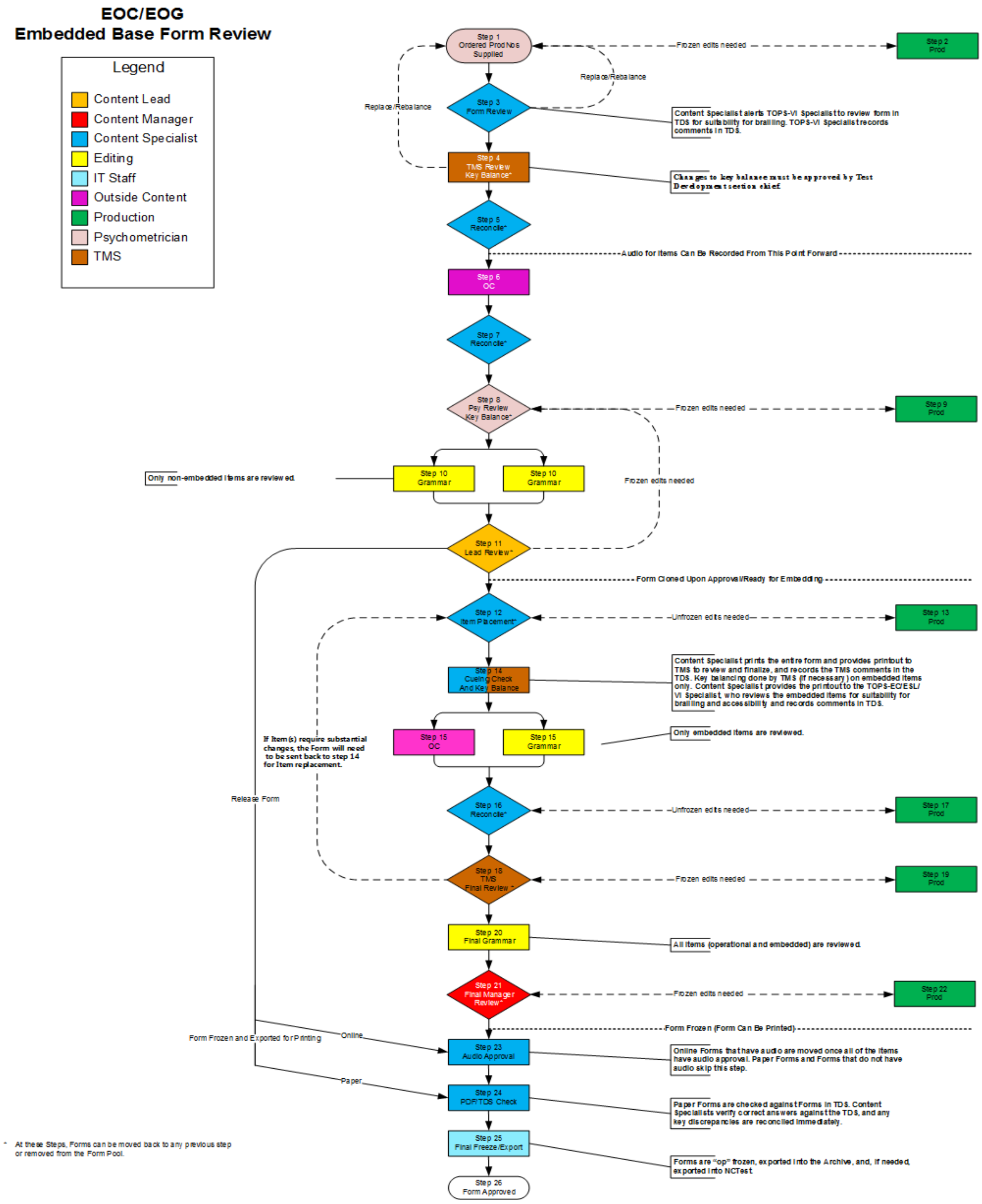
	Psychometric Evaluation Summary					
Grade	Keep		Reserve		Delete	
	N	%	N	%	N	%
5	141	29.4	162	33.8	177	36.9
8	206	42.9	127	26.5	147	30.6
Biology	379	50.5	207	27.6	164	21.9

4.5.2 Operational Form Assembly

Once the final item pool was reviewed and approved, psychometricians at NCDPI and test specialists at NCSU-TOPS began the iterative, operational-test construction process. NCDPI has instituted a 26-step iterative form building and review process (see Figure 4-2). For each grade level, operational forms are constructed to match the approved assessment blueprints described in *Section 3.2* and to match psychometric targets. An iterative process is used in order to optimally meet both considerations. The process begins with **Step 1**, in which **Psychometricians** build a base form from the item pool by selecting optimal items to match the content specification blueprint and statistical targets for the particular form. The form is sent to **Step 2, Production Edits** for revisions to artwork, graphs, or science selections. Then the form is sent to **Step 3, Content Specialist** for form review. At this step, the form is checked for content and cuing. If any issues are found the form is sent back to Step 1 for revision. Once the forms clear Step 3, it is sent to **Step 4, Test Measurement Specialist (TMS).** At this step, the TMS primarily checks items and form for alignment and key balance. Steps 1 through 4 are iterative until all areas are in agreement. Any item replacements recommended at any step are done at step 1; and if multiple items are replaced, the entire form review process is reset.

At Step 6, the form is sent to an outside content reviewer to offer general expert comments. Steps 8 through 11 involve grammar checks and key balance for multiple-choice items on the base form. Steps 12–18 involve cloning the base form with its operational items for the specified number of versions needed and then selecting field test items for review and addition into each of those form versions. Once all field test items have been approved, the form version is reviewed once more in its entirety by the TMS at Step 18, by Editing at Step 20, and the Content Manager at Step 21. If the TMS found no issues and Content Manager approved, the form is frozen and no future changes are usually allowed. Steps 23 through 26 are production steps in which computer-based versions are produced, audio is recorded for read-aloud, final PDFs are published and printed for paper-based forms and eventually large print, one test item per page, and braille forms creation as accommodations. Complete description of all the steps is available in Appendix 4-B Form Building & Test Development Process.

Figure 4-2 EOG/EOC Base Form and Review Steps



Content Specialist alerts TOPS-VI Specialist to review form in TDS for suitability for brailling. TOPS-VI Specialist records comments in TDS.

Changes to key balance must be approved by Test Development section chief.

Only non-embedded items are reviewed.

Content Specialist prints the entire form and provides printout to TMS to review and finalize, and records the TMS comments in the TDS. Key balancing done by TMS (if necessary) on embedded items only. Content Specialist provides the printout to the TOPS-EC/ESL/VI Specialist, who reviews the embedded items for suitability for brailling and accessibility and records comments in TDS.

If Item(s) require substantial changes, the Form will need to be sent back to step 14 for Item replacement.

Only embedded items are reviewed.

All items (operational and embedded) are reviewed.

Online Forms that have audio are moved once all of the items have audio approval. Paper Forms and Forms that do not have audio skip this step.

Paper Forms are checked against Forms in TDS. Content Specialists verify correct answers against the TDS, and any key discrepancies are reconciled immediately.

Forms are "op" frozen, exported into the Archive, and, if needed, exported into NCTest.

* At these Steps, Forms can be moved back to any previous step or removed from the Form Pool.

4.5.3 Psychometric Targets Based on Classical Test Theory

In setting expected form difficulty, NCDPI recognized that all item statistics were based on field tests in 2011 when the newly adopted *Essential Standards* for Science were still in their first year of implementation. Therefore, it was expected that field test statistics would be less stable during operational administration and as a result, expected form difficulty would have to be readjusted. As a reference point, the targeted expected p-value of each form was 0.625, which is the theoretical average of a student getting 100% correct on the test and a student scoring a chance performance (25% for a 4-foil multiple-choice test). That is (100 + 25)/2. The actual target was chosen by first looking at the distribution of the p-values for each grade-level item pool. While the goal was to set the target as close to 0.625 as possible, it was often the case that the target p-value was set between the ideal 0.625 and the average p-value of the item pool. Table 7-2 in *Section 7.4* shows mean p-value and biserial correlations for the field test and operational forms.

4.5.4 Psychometric Targets Based on IRT Parameters

Test Characteristic Curves (TCCs) generated from IRT parameters calibrated from the field tests were used in a pre-equated design to ensure that multiple parallel forms were developed at each grade level. Ideally, the expectation is that TCCs from alternate parallel forms will perfectly overlay each other. Furthermore, assuming that content and blueprint specifications are met, well-aligned TCCs ensure test forms are matched in difficulty and expected performance.

Once item parameters for items are calibrated, a probabilistic relationship between each item along the ability continuum of $-\infty$ to $+\infty$ can be represented with a nonlinear monotonically increasing function called an item characteristic curve (ICC) (Hambleton & Swaminathan, 1985). The ICCs represent a summary figure, which can be used to evaluate the statistical properties for each item. Conclusions about difficulty, discrimination, and pseudo-guessing score for each item can be inferred for examinees at different ability levels along the ability continuum. In form building, items are selected to match a particular target based on their ICC.

- **Test Characteristics Curve (TCC)**

In IRT, the Test Characteristics Curve (TCC) is essential for form assembly and scaling. TCCs are generally "S-shaped" figures with flatter ends that show the expected summed score as a function of theta (Θ_j) (Thissen, Nelson, Rosa, & Mcleod, 2001). Mathematically, the TCC function is the sum of ICCs for all items on the test see equation (4-5). During form assembly, items with known parameters are selected from the item bank based on a predetermined blueprint to match a target or base TCC. According to Thissen et. al. (2001, p.158), plotting TCCs for alternate forms on the same graph is an easy way to examine the relation of summed score with theta.

$$TCC = \sum_{k}^{I} \sum_{k=0}^{k-I} KT_{ik}\,(\theta) \qquad \textbf{(4-5)}$$

- **Test Information Function (TIF) and Conditional Standard Error (CSE)**

The concept of reliability (ρ) is central in CTT when evaluating the overall consistency of scores over replications, and it is generally reported in terms of standard error (SE), which is defined as $S_x\sqrt{1-\rho}$. Under the CTT framework, reliability and standard error are sample based; and regardless of where examinees are on the score scale, the amount of measurement error is uniform. Thissen and Orlando (2001, p. 117) highlighted that in IRT standard errors usually vary for different response patterns for the same test. Examinees with different response patterns or at different points on the theta scale will show variations in the amount of measurement precision. No single number characterizes the precision of the entire set for IRT scale score tests. Instead, the pattern of precision over the range of the test may be plotted as TIF and is defined as 1/SE^2. The concept of measurement precision as reported by TIF or CSE has been well documented in IRT literature and for more on this, see Hambleton & Swaminathan (1985), Thissen & Orlando (2001). Some features of TIF, as noted in Hambleton & Swaminathan (1985, p104):

- TIF is defined for a set of test items at each point on the ability scale.
- The amount of information is influenced by the quality and number of test items

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i\,(\theta)^2}{P_i\,(\theta)Q_i\,(\theta)} \qquad \textbf{(4-6)}$$

(I) The steeper the slope the greater the information

(II) The smaller the item variance the greater the information

- I(θ) does not depend upon the particular combination of test items. The contribution of each test item is independent of the other items in the test.
- The amount of information provided by a set of test items at an ability level is inversely related to the error associated with ability estimates at the ability level.

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Figure 4-3 through Figure 4-5 display TCCs for parallel operational forms assembled based on the field test item parameters for each grade level. The overlay of the TCCs indicates that the test forms are similar in psychometric characteristics. The estimated test information functions (TIFs) with associated conditional standard error of measurement (CSE) were also computed following IRT methodology. The TIFs and CSE plots are displayed in Appendix 4-C.

The TCCs show the theoretical expected score (vertical axis) for examinees by form across varying ability (horizontal axis) on the construct. Visual evidence of overlay TCCs in IRT is enough evidence to conclude that conditional on theta (ability) examinees are expected to have the same observed score across the different forms.

Figure 4-3 EOG Grade 5 Science TCCs Forms A, B, C, M, N, and O

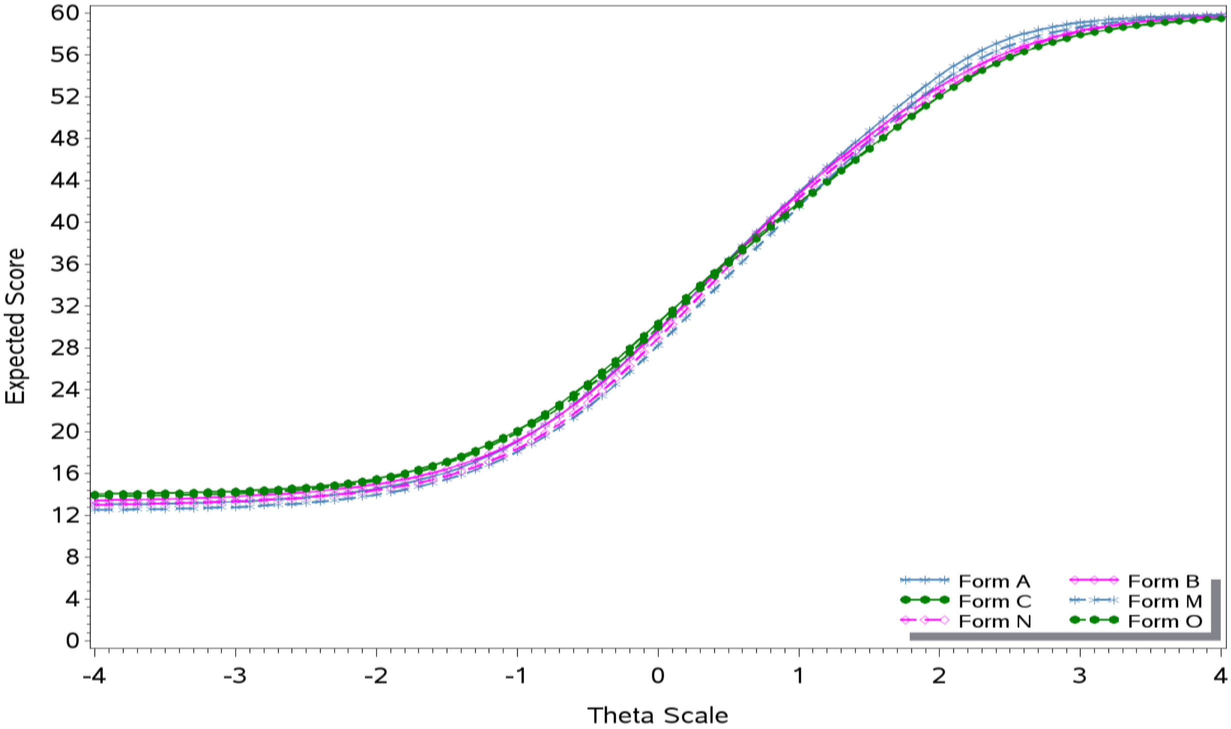


Figure 4-4 EOG Grade 8 Science TCCs Forms A, B, C, M, N, and O

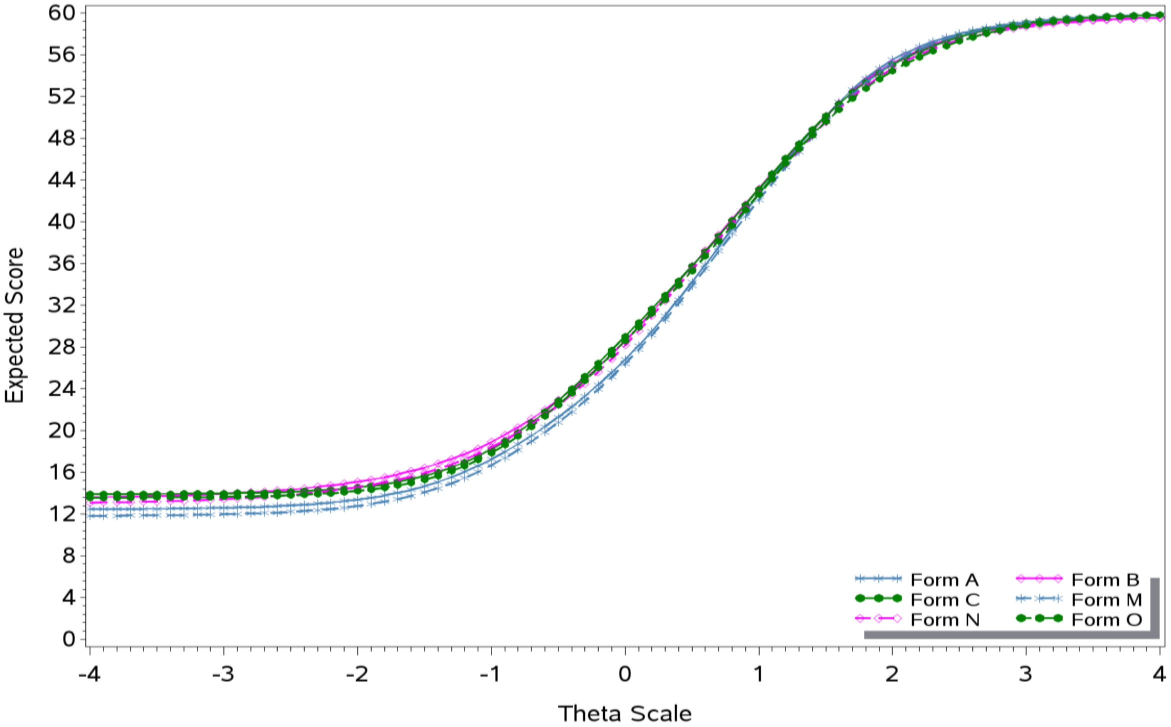
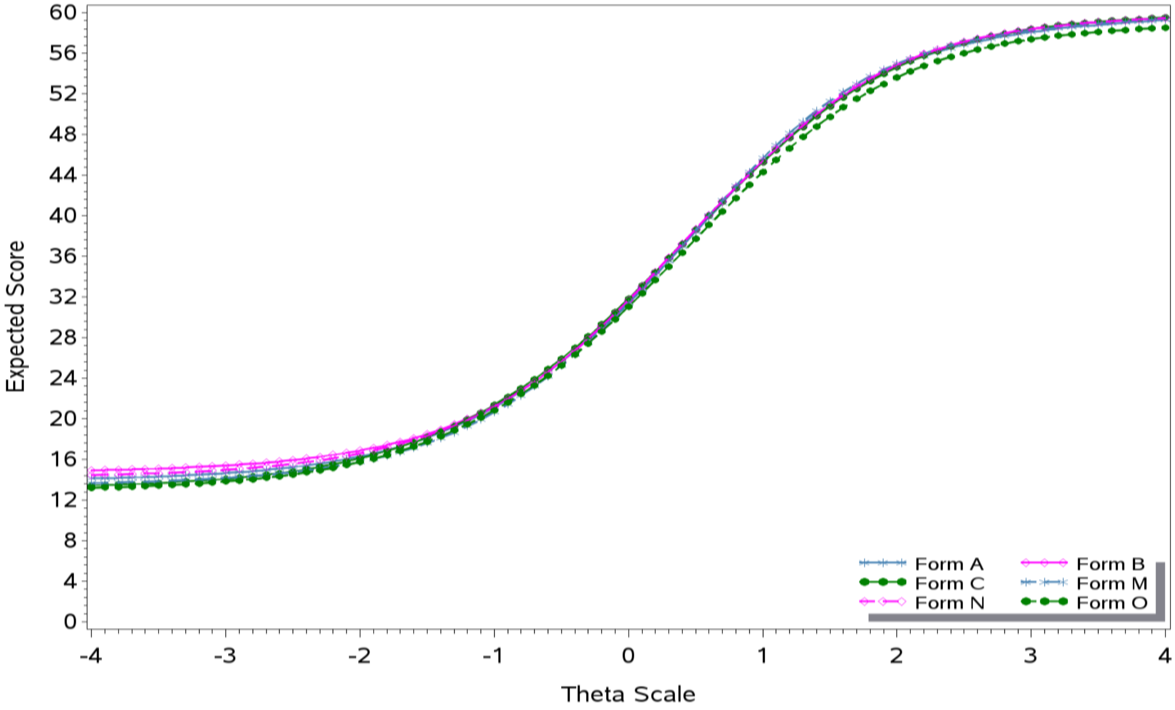


Figure 4-5 EOC Biology TCCs Forms A, B, C, M, N, and O



4.6 Step 16–Operational Test Forms Review

Once forms were assembled to meet content specifications, test blueprints, target p-values, and target-IRT item parameters, they were sent to outside content experts who provided an independent outside review of all assembled forms. Criteria for evaluating each test form included the following:

- Content of the test forms reflects the goals and objectives of the North Carolina *Standard Course of Study* for the subject (content validity).
- Content of the test forms reflects the goals and objectives as taught in North Carolina schools (instructional validity).
- Items are clearly and concisely written and the vocabulary appropriate to the target age level (item quality).
- Content of the test forms is balanced in relation to ethnicity, gender, socioeconomic status, and geographic district of the state (free from test/item bias).

- An item has one—and only one—best answer that is correct; the distractors should appear plausible for someone who has not achieved mastery of the representative objective (one best answer).

Reviewers were instructed to complete a mock administration of the tests (circling the correct responses in the booklet as well as recording their responses on a separate sheet) and to provide comments and feedback next to each item. After reviewing all items on a form, each reviewer independently completed a survey asking for their opinion as to how well the tests met the five criteria listed above. During the last part of the session, the reviewers discussed the tests and provided comments as a group. The reviewers' comments were recorded in Test Development System and were reviewed by the NCDPI and NCSU-TOPS content specialists. Items that were determined to be problematic at this point were replaced, and the forms rebalanced.

Apart from psychometric quality of item or content alignment concerns, items could also have been removed from a form because of cuing concerns, overemphasis of a particular subtopic (e.g., all area problems in one form were isosceles triangles), or for maintaining statistical equivalency. If a form had more than 10% of its items replaced as a result of this process, per NCDPI psychometric policy, the form went through the entire form review process again as it was no longer considered the same form that was reviewed previously. As a final review, test development staff members, with input from curriculum staff, content experts, and editors, conducted a final check on content and grammar for each test form.

4.7 Computer-Based Forms Review

After computer-based forms for Grades 5 and 8 EOG Science and Biology EOC are exported from the Test Development System (TDS) application into the NCTest platform, series of quality checks are performed to ensure all the specified interactions between items and the NCTest platform are fully functional across the different end users' approved devices. NSCU-TOPS and the NCDPI technology sections have instituted a five-phase quality check system which focuses on aspects ranging from technical and network comparability to accessibility by verifying that high contrast, large font, read-aloud features are working properly. Below is a summary description of the five-phase quality checks performed on all computer-based forms.

In Phase 1, forms are assigned to demo students who perform quality checks on each form for all the different presentation types (high contrast, large font, read-aloud) available during operational administrations. In Phase 2, NCSU-TOPS employees conduct quality checks to ensure the correctness of the forms and the items themselves. The Editing/Production group is notified if issues arise with respect to the content, whereas the NCTest group is notified if there are any issues with the apps or supporting resources. Phase 3 involves testing various features of the NCTest apps, such as highlighting, audio playback, or scrolling across the Chrome and iPad apps. On the NCTest chrome app, the features are checked at various resolutions to ensure the best experience for users. In Phase 4, forms are checked to ensure the data is being recorded accurately and the scoring keys for the items on each form are accurate. The NCDPI accountability IT group validates the data collected at this stage. In Phase 5, test measurement specialists at the NCDPI listen to all audio recordings and view all items with presentation settings (e.g. large font, high contrast). A complete final check is performed on desktops and iPads to ensure items interact with the user and display appropriately. Findings are then reported to NCSU-TOPS for any corrections, and all such corrections are monitored and verified as complete by the NCDPI.

Chapter 5 Test Administration

This chapter of the technical report describes the materials prepared and the activities engaged in by the NCDPI to assure a uniform administration of the test for all students across the state of North Carolina. If students take an assessment under different conditions, the comparability of the resulting test scores can be undermined. The chapter presents the efforts made to standardize test administration for the NC assessments to reduce construct-irrelevant variance that could thus undermine the comparability of test scores.

5.1 Test Administration Materials

NCDPI prepared materials prescribing the means for administering the NC EOG and EOC assessments. This section describes test administration materials prepared by the NCDPI and made available to test administrators to ensure standardized administration of the EOG and EOC assessments across the state as stated in standard 6.1 of the *Standards,* which states, "Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user" (p. 114).

An assessment guide is produced each for EOG science grades 5, science grade 8, and EOC Biology. However, there is only one proctor's guide used for all assessments, all grade levels, and courses.

- *Assessment Guide*: The assessment guide is the source document used for training all test administrators across the state. The guide provides comprehensive details on key features about each assessment. Key information provided includes a general overview of each assessment that covers the purpose of the assessment, eligible students, testing windows, and makeup testing options. The assessment guide also covers all preparations and steps that should be followed the day before testing, on test day and after testing. Samples of answer sheets are also provided in the assessment guide.
- *Proctor's Guide*: The proctor's guide serves as the source document with detailed guidelines for the selection of proctors, the definition of their roles, and training information. Key training topics covered in the proctor's guide include the defining of proctors' responsibilities as well as training on how to maintain test security, ensure appropriate testing conditions, maintain students' confidentiality, assist test administrators, monitor students, report test irregularities, and follow appropriate procedures for accommodations.

The NCDPI also provides the *Guidelines for Testing Students Identified as Limited English Proficient* document. This guide provides training on the following areas: ELL testing requirements, responsibilities of test coordinators, procedures for participation, available testing accommodations, and the monitoring of the accommodations.

Regarding the clarity of the test administration directions *Standard 4.15* of the Standards (AERA, APA, & NCME, 2014) states that "The directions for test administration should be presented with sufficient clarity so that it is possible for others to replicate the administration conditions under which the data on reliability, validity, and (where appropriate) norms were obtained. Allowable variations in administration procedures should be clearly described. The process for reviewing requests for additional testing variations should also be documented" (p. 90).

5.2 Test Administrators Training

The North Carolina Testing Program uses a train-the-trainer model to prepare test administrators to administer North Carolina tests. Regional Accountability Coordinators (RACs) receive training in test administration from the NCDPI Testing Policy and Operations staff at regularly scheduled monthly training sessions. Subsequently, the RACs provide training to Local Education Agency (LEA) test coordinators on the processes for proper test administration. LEA test coordinators provide this training to school test coordinators. The training includes information on the test administrators' responsibilities, proctors' responsibilities, preparing students for testing, eligibility for testing, policies for testing students with disabilities, ELL students, accommodated test administrations, test security (storing, inventorying, and returning test materials), and the *Testing Code of Ethics* (see Appendix 2-A).

5.3 Security Protocols Related to Test Administration

Test security is an ongoing concern in any testing program. When test security is compromised, it can undermine the validity of test scores. For this reason, the NCDPI has taken extensive steps to ensure the security of the assessments by establishing protocols for the school employees who administer tests, for handling and administering paper and pencil tests, and for administering computer-based tests.

5.3.1 Protocols for Test Administrators

Only school system employees are permitted to administer secure state tests. Those employees must participate in the training for test administrators described in *Section 5.2*. Test administrators may not modify, change, alter, or tamper with student responses on the answer sheets or test books. Test administrators must thoroughly read the *Assessment Guide* and the codified North Carolina *Testing Code of Ethics* before the actual test administration. Test administrators must also follow the instructions given in the *Assessment Guide* to ensure a standardized administration and must read aloud all directions and information to students as indicated in the manual. The school test coordinator is responsible for monitoring test administrations within the building and responding to situations that may arise during test administrations.

5.3.2 Protocols for Handling and Administering Paper Tests

When administering paper tests, school systems are mandated to provide a secure area for storing tests. The Administrative Procedures Act 16 NCAC 6D .0302 states, in part, that

> *LEAs shall (1) account to the department (NCDPI) for all tests received; (2) provide a locked storage area for all tests received; (3) prohibit the reproduction of all or any part of the tests; and (4) prohibit their employees from disclosing the content of, or specific items contained in, the test to persons other than authorize employees of the LEA.*

At the individual school, the principal is responsible for all test materials received. As established by SBE policy GCS-A-010, the *Testing Code of Ethics (*Appendix 2-A*),* the principal must ensure test security within the school building and store the test materials in a secure, locked facility, except when in use. The principal must establish a procedure to have test materials distributed immediately before each test administration. Every LEA and school must have a clearly defined system of check-out and check-in of test materials to ensure at each level of distribution and collection (LEA, school, and classroom) all secure materials are tracked and accounted for. LEA/charter school test coordinators must inventory test materials upon arrival from NCSU-TOPS and must inform NCSU-TOPS of any discrepancies in the shipment.

Before each test administration, the school test coordinator shall collect, count, and return all test materials to the secure, locked storage area. Any discrepancies are to be reported to the school system test coordinator immediately, and a report must be filed with the regional accountability coordinator.

At the end of each test administration cycle, all testing materials must be returned to the school test coordinator according to directions specified in the assessment guide. Immediately after each test administration, the school test coordinator shall collect, count, and return all test materials to the secure, locked facility. Any discrepancies must be reported immediately to the LEA test coordinator. Upon notification, the LEA test coordinator must report the discrepancies to the regional accountability coordinator and ensure all procedures in the Online Testing Irregularity Submission System (OTISS) are followed to document and report the testing irregularity. The procedures established by the school for tracking and accounting for test materials must be provided upon request to the LEA test coordinator and/or the NCDPI Division of Accountability Services/North Carolina Testing Program.

At the end of the testing window, NCDPI mandates that all assessment guides, used test booklets that do not contain valid student responses, unused test booklets, and unused answer sheets be securely destroyed immediately at the LEA. Secure test materials are to be retained by the LEA in a secure (locked) facility with access controlled and limited to one or two authorized school personnel only. After the required storage time (see Table 5-1) has elapsed, the LEA should securely destroy these materials.

Table 5-1 Test Materials Designated to Be Stored by the LEA in a Secure Location

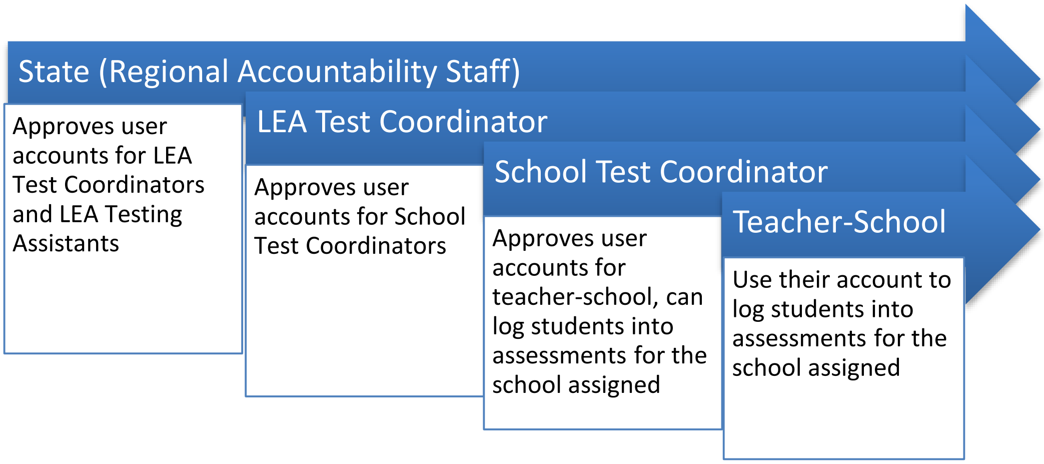
Test	Required Storage Time
All used answer sheets for operational tests (including scoring sheets for W-APT)	Six months after the return of students' test scores
Original responses recorded in a test book, including special print version test books (i.e., large print edition, one test item per page edition, Braille edition)	Six months after the return of students' test scores
Original Braille writer/slate and stylus responses	Six months after the return of students' test scores
Original responses to a scribe	Six months after the return of students' test scores
Original responses using a typewriter or word processor	Six months after the return of students' test scores
Answer sheets with misaligned answers (keep testing irregularities in a separate file)	Six months after the return of students' test scores
NC General Purpose Header Sheets	Store indefinitely
EOC or EOG Graph Paper EOC: Math I, Biology, and English II	Store indefinitely Retain unused test materials from fall for use in spring; retain unused test materials from spring for use in summer

5.3.3 Computer Mode Test Security Measures

Since the 2012–13 administration, Grades 5 and 8 EOG Science and Biology EOC operational assessments have been available in both computer and paper modes. The NCTest platform is used to administer computer-based and fixed-form assessments. The NC Education system manages student enrollments, monitors assessment start and stoppage times, and manages accommodation needs.

The NCDPI limits all LEA access to the computer-based assessment to specific testing days. The LEA test coordinator must enter test dates in NC Education for each assessment to be administered by computer. Assessments can only be accessed through NCTest on those specific dates. In addition, access is limited to users with a valid and verified NC Education username and password. Figure 5-1 shows the tiers of NCTest users along with the information about who assigns access.

Figure 5-1 NCTest *User Access Security Protocol*


State (Regional Accountability Staff)
Approves user accounts for LEA Test Coordinators and LEA Testing Assistants
LEA Test Coordinator
Approves user accounts for School Test Coordinators
School Test Coordinator
Approves user accounts for teacher-school, can log students into assessments for the school assigned
Teacher-School
Use their account to log students into assessments for the school assigned

The NCTest platform is accessed through a Hyper Text Transport Protocol Secure (HTTPS) Uniform Resource Locator (URL). Full HTTPS encryption is applied between the NCTest server located at NC State University and NCTest. The connection is encrypted using Transport Layer Security (TLS 1.2) and authenticated using AES_128_GCM with DHE_RSA as the exchange mechanism. At the time of log-in, the tests are sent securely from the NCTest server at NC State University to the local computer. Not all assessment content is sent at the time of login, only the text for all the test items is sent at that time. Graphics and audio files (for computer read-aloud accommodations) are sent as students move from item to item within the assessment.

After each item is answered, the students' responses are sent securely to the NCTest server at NC State University using the same full HTTPS encryption process. At the conclusion of the assessment, local users are instructed to clear all caches and cookies from local machines. After online student assessments are finalized, they are transferred nightly to the NCDPI and/or the scoring vendors. These transfers are done following the NCDPI Secure File Transfer Protocol (SFTP) encryption rules and logic. More information on these processes can be found in NCDPI's *Maintaining the Confidentiality and Security of Testing and Accountability Data* policy published annually in each Assessment Guide, *Testing Security: Protocol and Procedures for School Personnel* document and the *Test Coordinators' Policies and Procedures Handbook*. The NCDPI and NCTest systems operate within the same network and are hosted at NC State University.

5.4 Administration

5.4.1 Test Administration Window

In the 2012–13 administration, all eligible students enrolled in grades 5 and 8 were required to participate in the EOG assessments administered within the last fifteen (15) days of the school year. Based on the traditional school calendar, EOG assessments are administered in late spring on the school academic calendar.

The EOC has two administration windows: one in fall and another in spring. In the 2012–13 administration, students enrolled in semester schedules were required to take EOC assessments within the last fifteen (15) days of the semester. Students enrolled in a yearlong course schedule were administered the EOC assessment within the last twenty (20) days of the instructional period.

Beginning with the 2013–14 school year, the testing window was modified and changed so all students in grades 5 and 8 were administered the EOG assessment during the last ten (10) days of the school year; the EOC administration window was changed to the last five (5) days of the instructional period for the semester courses or the last ten (10) days of the instructional period for the yearlong courses. Districts can request a waiver to increase the testing window by five (5) days.

5.4.2 Timing Guidelines

The science EOG and EOC assessments are not power tests with strict time requirements. All examinees are given ample time to demonstrate their knowledge of the construct being assessed. The *Standards* (AERA, APA & NCME, 2014) states that "although standardization has been a fundamental principle for assuring that all examinees have the same opportunity to demonstrate their standing on the construct that a test is intended to measure, sometimes flexibility is needed to provide essentially equivalent opportunities for some test takers" (p.51). In keeping with the *Standards* (AERA, APA & NCME, 2014), the NCDPI requires all general students be allowed ample opportunity to complete the assessments as long as they are engaged and working and the maximum time allowed (i.e., four hours) has not elapsed.

Based on timing data collected during field tests and analyzed as described in *Section 4.4*, the NCDPI recommended time allotted for the EOG science be 180 minutes, with a maximum of 240 minutes. The estimated time allotted for EOC Biology is 150 minutes, with a maximum of 240 minutes. For both the EOG and EOC assessments, students with approved accommodations may take even longer, as specified by their particular IEP or LEP plan.

5.4.3 Testing Accommodations

State and federal law requires that all students, including SWD and students identified as ELL, participate in the statewide testing program. Students may participate in the state assessments on grade level (i.e., general, alternate) with or without testing accommodations. Eligible students participating in the EOG and EOC assessments are provided "test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs" (the *Standards*, p. 67). Testing accommodations are defined as "changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests" (Thurlow & Bolt, 2001, p. 3). Accommodations are provided to eligible students together with appropriate administrative procedures to assure that individual student needs are met and, at the same time, maintain sufficient uniformity of the test administration.

For any state-mandated test, the accommodation for an eligible student must (1) be documented in the student's current Individualized Education Program (IEP), Section 504 Plan, ELL documentation, or transitory impairment documentation and (2) the documentation must reflect routine use of the accommodation during instruction and similar classroom assessments that measure the same construct. When accommodations are provided in accordance with proper procedures as outlined by the state, results from these tests are deemed valid and fulfill the requirements for accountability.

According to *Standard 6.2, “*When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing” (p. 115). In compliance with this, NCDPI specifies the following accommodations in North Carolina EOG and EOC assessments in the *Assessment Guide*s. The accommodations can also be viewed in the document called “Review of Accommodations Used During Testing.”

- Test Administrator Reads Test Aloud in English
- Computer Reads Test Aloud—Student Controlled (computer-based assessments only
- Braille Writer/Slate and Stylus (Braille Paper)
- Large Print Edition
- One Test Item per Page Edition
- Braille Edition
- Assistive Technology Devices
- Cranmer Abacus
- Dictation to a Scribe
- Interpreter/Translator Signs/Cues Test
- Magnification Devices
- Word-to-Word Bilingual (English/Native Language) Dictionary/Electronic Translator (ELL only)
- Student Marks Answers in Test Book
- Student Reads Test Aloud to Self
- Hospital/Home Testing (eliminated effective 2013–14 school year)
- Multiple Testing Sessions
- Scheduled Extended Time
- Testing in a Separate Room

For information regarding appropriate testing procedures, test administrators who provide accommodations for students with disabilities must refer to the most recent publication of *Testing Students with Disabilities* and any published supplements or updates. The publication is available through the local school system or at http://www.ncpublicschools.org/accountability/policies/tswd/. In addition, test administrators must be trained in the use of the specified accommodations by the school system test coordinator or designee before the test administration.

According to the *Standards* (AERA, APA & NCME, 2014), an appropriate accommodation addresses student's specific characteristics but does not change the construct the test is measuring or the meaning of scores; however, when necessary, modifications that change the construct are provided to students to measure their standing on some intended construct.

5.4.4 English Language Learners

Per State Board policy GCS-C-021, students identified as (ELL)[f] must participate in the statewide testing program using the accommodated or non-accommodated standard test administration, with one exception: students identified as ELL who score below Level 4.0 Expanding on the WIDA-ACCESS Placement Test and are in their first year in U.S. schools are exempt from taking the ELA EOG assessment or the English II EOC assessment.

For both EOG and EOC assessments, ELL students are provided with ELL reading accommodations based on their scores on the WIDA-ACCESS Placement Test (W-APT™). State Board policy GCS-A-001 requires that students scoring below Level 5.0 Bridging on the reading subtest of the W-APT/ACCESS for ELLs receive state-approved ELL testing accommodations on all state tests (see Figure 5-2). Students scoring Level 5.0 Bridging or above on the reading subtest of the W-APT/ACCESS for ELLs® or exiting ELL identification must participate in all state tests without ELL accommodations. The state approved ELL testing accommodations for science include:

- Multiple Testing Sessions

[f] Once identified as ELL based solely on the results of the W-APT™, the student is required by state and federal law to be assessed annually with the state-identified English language proficiency test. The test currently used by North Carolina for annual assessment of English Language Learners (ELLs) is the Assessing Comprehension and Communication in English State-to-State for English Language Learners, or the ACCESS for ELLs®.

- Schedule Extended Time
- Testing in a Separate Room
- Student Reads to Self-Aloud
- English/Native Language Word-to-Word Bilingual Dictionary/Electronic Translator
- Test Administrator Reads Test Aloud in English
- Computer Reads Test Aloud

For information regarding appropriate testing procedures, test administrators who provide accommodations for students identified as limited English proficient must refer to the most recent publication of *Guidelines for testing Students Identified as Limited English Proficient* and any published supplements or updates. The publication is available through the local school system or at http://www.ncpublicschools.org/accountability/policies/slep/. In addition, test administrators must be trained in the use of the specified accommodations by the school system test coordinator or designee prior to the test administration.

Figure 5-2 ELL Proficiency Levels and Testing Accommodations

Subtest	1	2	3	4	5	6
	Entering	Emerging	Developing	Expanding	Bridging	Reaching
Reading	**Eligible to Receive State-Approved ELL Testing Accommodations for All State Tests**				Must Participate in General State Test Administration without ELL Testing Accommodations	

5.4.5 Mode of Test Administration

The EOG science and EOC biology assessments may be administered either as paper- or computer-based fixed forms. Districts could opt to use either a paper- or a computer-based form. The state’s goal is to gradually transition test administration for the EOG and EOC assessments to the computer-based mode as districts are able to build their resources and technology capacities. Beginning with the 2012–2013 administration, the Grades 5 and 8 Science EOG and Biology EOC assessments were available in both paper and computer modes.

For the 2012-13 administration, districts could opt to use paper-based forms in place of the computer-based form. Beginning with the spring 2016 administration, the state mandated the grade 8 EOG science assessment be administered as computer-based, fixed forms with the following exceptions:

1. Local education agencies (LEAs) or charter schools that do not have the technology capability to support administering computer forms
2. Individual students with disabilities who have documented accommodations that dictate a paper/pencil test format is necessary for accessibility

Table 5-2 shows the total number of students who took science EOG and EOC tests by mode during the 2013, 2014, and 2015 test administration windows. As shown in the table, a similar proportion of students (over 60%) were administered the computer-based forms across administrations in all grade levels. There is a decreasing, albeit minimal, trend in EOG tests (about 2% in grade 5 and 1% in grade 8) and an increasing trend (55% in 2013 to 64% in 2015) in Biology EOCs taken with the computer-based form.

Table 5-2 EOG and EOC Tests Administered by Mode

Grade and Year		Test Administration Mode			
		Computer		Paper	
		Total Test	*Percent*	*Total Test*	*Percent*
EOG Grade 5	*2013*	74,629	67%	36,907	33%
	2014	73,800	65%	39,159	35%
	2015	67,360	63%	40,218	37%
EOG Grade 8	*2013*	76,416	69%	33,876	31%
	2014	76,777	68%	36,809	32%
	2015	79,515	67%	38,531	33%
EOC Biology	*2013*	60,489	55%	49,495	45%
	2014	69,744	62%	42,012	38%
	2015	74,705	64%	42,157	36%

5.4.6 Student Participation

The Administrative Procedures Act 16 NCAC 6D. 0301 requires that all public school students enrolled in grades for which the North Carolina State Board of Education (NCSBE) adopts an assessment, including every child with disabilities, participate in the testing program unless excluded from testing. For EOG assessments, all students in grades 5 and 8 are required to participate in the end-of-grade science assessments or the corresponding alternate assessment, as indicated by the students' Individualized Education Programs (IEPs) or appropriate ELL documentation. For EOCs, all students enrolled in Biology must be administered the EOC test (16 NCAC 6G.0305 [g]). Students who are repeating the course for credit must also be administered the EOC assessment.

According to State Board policy GCS-A-001, school systems shall, at the beginning of the school year, provide information to students and parents or guardians advising them of the districtwide and state-mandated assessments that students are required to take during the school year. In addition, school systems must provide information to students and parents or guardians to advise them of the dates the tests will be administered and how the results from each assessment will be used. Information provided to parents about the tests must include whether the NCSBE or local board of education requires the test. School systems must report test scores and interpretative guidance from districtwide and/or state-mandated tests to students and parents or guardians within thirty (30) days of the generation of the score at the school system level or receipt of the score and interpretive documentation from the NCDPI.

5.4.7 Medical Exclusions

There may be rare circumstances in which a student with a significant medical emergency and/or condition may be excused from the required state tests. For requests that involve significant medical emergencies and/or conditions, the LEA superintendent or charter school director must submit a written request to the NCDPI. The request must include detailed justification explaining why the student's medical emergency and/or condition prevent participation in the respective test administration during the testing window and the subsequent makeup period. Most of what is submitted for the medical exception is housed at the school level (IEP, dates of the scheduled test administration(s) and makeup dates, number of days of instruction missed because of the emergency/condition, expected duration/recovery period, explanation of the condition and how it affects the student on a daily basis, etc.) The student's records remain confidential and any written material containing identifiable student information is not disseminated or otherwise made available to the public. For more information on the process for requesting special exceptions based on significant medical emergencies and/or conditions, please review the annual memo (*Request for Testing Exceptions Based on Significant Medical Emergencies and/or Conditions*) located at http://www.ncpublicschools.org/accountability/generalinfo.

Chapter 6 Scoring and Scaling

This chapter describes the processes used for scoring items and the procedure adopted to create final reportable scale scores. The first two sections of this chapter summarize the automated scoring procedures that transform student responses into a number correct score for MC items. Sections three and four describe the procedures used to transform raw scores into a reportable scale across different grades. The final section describes the data certification processes used by the NCDPI to ensure the quality of student data. The information in this chapter is intended to comply with AERA/APA/NCME (2014) *Standard 4.18,* which states:

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

Information in the chapter is presented with enough detail to meet Standard 4.18, but not so much as to compromise the integrity of the test items.

6.1 Automated Scoring of Fixed Response Items

NCDPI's WinScan software program is used for scoring all EOG and EOC student responses. WinScan is a specialized scoring and reporting software program created and managed by the NCDPI/Accountability Services Division. At the beginning of each testing window, a new release of WinScan is updated and distributed to all LEAs. Each version is programmed using the score keys and raw-to-scale score conversion tables for all approved operational test forms. WinScan is then used at each LEA to score and report test results as soon as student response materials are sent to the LEA test coordinator office from schools.

For paper-based forms, the school system's test coordinator establishes the schedule for receiving, scanning, and scoring EOG tests at the LEA level. The school system's test coordinator, upon receipt of student response sheets, (1) scans the answer documents, (2) provides the results (reports) from the test administrations soon after scanning/scoring is completed, and (3) stores all answer sheets in a secure (locked) facility for six months following the release of test scores. After six months, all student answer sheets are recycled or destroyed in a secure manner in accordance with NCDPI procedures as described in the assessment guide. The regional accountability coordinator (RAC) has the responsibility of scanning and scoring tests for charter schools and for providing long-term storage for specific test materials such as used answer sheets and used test books (only available for the *Student Marks Answers in Test Book* accommodation).

Computer-mode forms are scored electronically via a centrally hosted server at NCDPI using WinScan software. Once WinScan assigns scores for each item, data are merged with student-level records then made electronically available to test coordinators. Once the data are available, school system test coordinators can generate school rosters, class rosters, and individual reports. Initial district or school-level reporting occurs at the LEA level.

6.2 Scale Scores

After scoring is completed, raw scores for EOG and EOC assessments are transformed and reported on a scale score metric based on IRT-summed score procedures described in this section. Advantages of reporting scale scores:

- A standard metric is provided to report scores when multiple test forms are used.
- Scale scores can be used to compare the results of tests that measure the same content area but are composed of items presented in different formats.
- Scale scores can be used to minimize differences among various forms of the tests.

For practical reasons, the NCDPI uses summed-score and IRT expected a posteriori (EAP) theta estimates to establish raw-to-scale conversions for the North Carolina EOG and EOC tests. According Standard 5.2: “The procedures for constructing scales used for reporting scores and the rationale for these procedures should be described clearly” (the *Standards*, p.102). This section presents a summary of the procedures used to transform raw scores into scale scores. For in-depth review of the procedure see Thissen and Orlando (2001, p. 119). For any IRT model with item scores indexed ($u_i = 0,1$), the likelihood for any summed scores $x = \sum u_i$ is:

$$\boldsymbol{L_x(\theta) = \sum_{\sum u_i = x} L(u/\theta)} \qquad \textbf{(6-1)}$$

Where $L\left({}^{u}/_{\theta}\right) = \prod_i T(u_i/\theta)$ and $T(u_i/\theta)$ is the traceline for response u to item i. The first summation is over all such response patterns that the summed score equals x. The probability of each score is

$$\boldsymbol{P_x = \int L_x(\theta) g(\theta)} \qquad \textbf{(6-2)}$$

And the expected θ associated with each summed score is

$$\boldsymbol{E(\theta/x) = \frac{\int \theta L_x(\theta) g(\theta)}{P_x}} \qquad \textbf{(6-3)}$$

With posterior standard deviation (PSD) given by

$$\boldsymbol{PSD\left(\theta/x = \sum u_i\right) = \left\{\frac{\int [\theta - E(\theta/x)]^2 L_x(\theta) g(\theta)}{P_x}\right\}^{1/2}} \qquad \textbf{(6-4)}$$

Scoring was done in IRTPRO using calibrated item parameters to estimate EAP theta scores. To ensure all theta are on the same scale, the population mean and standard deviation of the current year is used during scaling to create summed score-to-scale conversion tables for all EOG forms. By creating separate raw-to-scale tables for each form, any minor statistical form differences are accounted for and equated. Thus it makes no difference to students which form is administered.

6.3 Data Certification

Before the release of test scores for official reporting, the NCDPI performs data certification to ensure all items were correctly scored using correct keys. The NCDPI rule is to perform data certification analyses once 10% of the expected population has tested during the current cycle. The certification process requires the completion of three main quality control steps: (1) content review of flagged items; (2) independent scoring of student responses and (3) computing CTT statistics and comparing to the field test.

During the first step, the NCDPI test measurement content specialist completes each flagged item without the answer key. Item statistics are reviewed and comments are documented.

In step 2, the NCDPI independently scores student response strings and checks for agreement with scores reported from the WinScan system. The standard is to have a 100% agreement rate between scores from WinScan and the independent scoring.

In step 3 of the certification process, CTT item statistics are computed and checked against field test statistics to make sure items performed as expected. During this step, any item that showed significant variation from the field test statistics is further investigated to make sure the scoring is correct. If any issues are found either because of a wrong scoring key or an improper rendering of any sort, the item is dropped from the form as an operational item and a new raw-to-scale table is generated for that form and updated in WinScan.

Upon completion of certification analyses, the test data generated are certified as accurate, provided that all NCDPI-directed test administration guidelines, rules, procedures, and policies have been followed at the district and school levels in conducting proper test administrations and in generating the student response data. Finally, the NCDPI issues an official communiqué affirming forms have been certified and scale scores are approved for official reporting.

Chapter 7 Analyses of Operational Data

This chapter describes the analyses of operational data after the first operational administration of EOG and EOC assessments in 2012–13. The chapter begins with a description of the random spiraling process used to administer three parallel forms across North Carolina. The chapter goes on to summarize item analysis results from the operational administration in 2012–13, which includes CTT- (p-value, point-biserial, Cronbach alpha) and IRT-based analysis (item calibration and scoring, test characteristics curves, test information functions, and conditional standard errors).

7.1 Pre-Equated Testing Model

NCDPI's testing program uses a pre-equated model based on IRT to score test forms and compute raw-to-scale tables for each form before operational administration. This model allows the department to satisfy NCSBE policy GCS-A-001 (h): "School systems shall report scores resulting from the administration of districtwide and state-mandated tests to students and parents or guardians along with available score interpretation information within thirty (30) days from the generation of the score at the LEA level or receipt of the score and interpretive documentation from the NCDPI."

For the first administration of the North Carolina EOG and EOC assessments in 2012–13, test results were delayed so post-item analysis could be conducted on items administered in an operational setting. The reasons for the delay were twofold:

- First, the three operational parallel forms were constructed using data from stand-alone online and embedded paper and pencil field test administrations for grades 5 and 8 EOG science and EOC Biology. Field test data are usually considered unstable, and it is common to experience drift in item parameters between a field test and an operational administration. In North Carolina's case, the items were field tested when districts and schools were still transitioning to the new standards and students had not had ample opportunity to learn under these new standards. Also, student motivation is generally expected to differ between the field test and operational administration.
- Second, the NCDPI needed time to reanalyze all forms based on operational data to ensure stable base-year item parameters as well as scale scores were used for standard setting.

7.2 Spiraled Form Administration

Six parallel forms in Grades 5 and 8 EOG Science and EOC Biology (Paper: A, B, C; Computer: M, N, O) were administered operationally for the first time in the 2012–13 school year. Forms M, N, and O confirm paper forms A, B, and C in terms of test blueprint except TE item types. At every grade level, all forms were administered to randomly equivalent groups of examinees for whichever mode of administration was used. Within each grade, the forms were spiraled within classrooms. Spiraling forms ensures that item parameters calibrated from random samples of students who were administered different test forms are put on the same IRT scale and can be compared directly without need for equating.

Table 7-1 shows descriptive summary of demographic variables for students who were administered science EOG and EOC assessments in 2012–13. The student counts listed in these tables are the number of valid tests administered, not the actual official enrollment records. The actual difference between the total student population and sample included in item analysis is trivial and given the very large sample sizes at every grade, such differences are not expected to impact final item and test statistics reported. On average, over 100,000 students per grade level at grades 5 and 8 and in high school were administered the EOG science or EOC Biology assessments. Notice that more students were administered computer-based forms over paper-and-pencil for all science assessments except grade 5 form N. The reason for a low n-count for form N was due to a display issue of one TE item. The form was pulled from rotation and the display issue corrected; however, form N was not returned to rotation. The results further indicated that the gender distribution across forms were very similar. However, ethnicity-wise, there were some discrepancies. It is important to mention here that the schools' assignment to mode of administration was not random. A further analyses revealed that some school districts considered to be high-performing chose paper over online modes of administration.

Following completion of the 2012–13 operational administration, data for each form from all students who participated in the general EOG and EOC operational administration were reanalyzed, first using CTT, then by IRT calibrations.

Table 7-1 Demographic Summary for Science EOG and EOC Operational Tests 2012–13

Grade and Form		N	Gender (%)		Ethnicity (%)						
			Female	Male	Asian	Black	Hispanic	Amer. Indian	Multi-Racial	Native Hawaiian/Pacific Islander	White
Science Grade 5	A	12,450	49.04	50.80	4.22	29.45	15.01	2.45	3.91	0.07	44.73
	B	12,220	48.70	51.16	4.38	28.68	15.04	2.42	3.84	0.10	45.40
	C	12,237	48.93	50.89	4.16	29.31	14.11	2.61	3.99	0.13	45.49
	M	36,537	49.99	49.93	2.02	23.01	15.26	0.80	3.69	0.09	55.05
	N	1,711	51.14	48.63	1.69	29.28	10.93	1.81	3.57	0.18	52.31
	O	36,381	49.18	50.77	2.16	23.04	14.67	0.74	3.81	0.10	55.42
	All	**111,536**	**49.38**	**50.52**	**2.80**	**25.15**	**14.82**	**1.36**	**3.80**	**0.10**	**51.87**
Science Grade 8	A	11,455	48.88	50.76	3.88	29.72	13.00	0.72	3.33	0.08	48.90
	B	11,288	49.68	49.91	3.82	29.51	12.83	0.74	3.45	0.08	49.17
	C	11,133	50.10	49.50	4.18	28.98	13.26	0.58	3.13	0.11	49.37
	M	25,539	50.14	49.82	1.96	24.97	11.85	1.71	3.71	0.09	55.60
	N	24,893	49.20	50.78	1.87	24.77	12.48	1.67	3.42	0.10	55.60
	O	25,984	49.80	50.09	1.99	24.79	12.07	1.80	3.51	0.14	55.51
	All	**110,292**	**49.67**	**50.17**	**2.56**	**26.24**	**12.41**	**1.41**	**3.47**	**0.10**	**53.60**
Biology	A	16,671	49.99	49.87	3.39	31.58	10.99	1.39	3.52	0.08	48.91
	B	16,421	50.59	49.25	3.48	31.84	11.41	1.46	3.32	0.05	48.29
	C	16,403	49.52	50.31	3.43	31.32	11.35	1.41	3.24	0.10	48.98
	M	19,953	49.80	50.14	2.16	23.42	10.68	1.31	3.56	0.10	58.73
	N	20,060	49.97	49.93	2.24	23.12	11.12	1.28	3.59	0.08	58.46
	O	20,476	50.26	49.63	2.06	23.19	11.22	1.43	3.55	0.12	58.33
	All	**109,984**	**50.02**	**49.86**	**2.73**	**26.99**	**11.11**	**1.37**	**3.47**	**0.09**	**54.10**

7.3 Operational Forms Item Analyses

At the conclusion of testing during the 2012–13 administration window, NCDPI reanalyzed data for all operational forms. The purpose of these post-administration analyses was to establish final item parameters, create official raw-to-scale scoring tables, and provide item statistics and student-level data for standard setting. This section presents summary results of the post-administration item analyses conducted after the 2012–13 window and evidence of item statistics drift between field test and operational administration. First, for each form all operational items were reanalyzed following the CTT and IRT procedures described in *Section 4.2*. For IRT analyses, single-group calibrations were performed for each form. IRT item parameters together with basic CTT statistics were compared to similar statistics used during form building from field test data.

7.3.1. EOG and EOC IRT Calibration Across Modes

All operational items in the six parallel forms (A and M, B and N and C and O) created from field test data were reviewed using the psychometric criteria presented in *Section 4.5.1* Following these analyses, no items performed differentially between the paper and computer modes. Therefore, no need of scaled around and no items were removed from the final operational forms for science.

The process for identifying differential performing items and Scaled Around included DIF sweep procedures in IRTPRO concurrent calibration. The DIF sweep option in IRTPRO (Cai, Thissen, & du Toit) allows a two-step calibration process in which items administered in two different modes (paper and computer) are first evaluated for evidence of differential functioning. During the first step, separate parameter estimates were calibrated across modes for each item. The purpose of the DIF sweep calibration is to classify items into two categories: 1) anchor items and 2) candidate DIF items. Anchor items display no mode effects, while candidate DIF items display some degree of mode effects. Mode effects can be visualized by superimposing the ICCs of two items onto the same graph. Items that display mode effects will display separate lines that differ substantially from one another. For instance, if an item is more difficult when administered on computer, the ICC for the computer administered item will be shifted to the right compared to the ICC from the paper administered item.

Effect size measures were calculated to quantify the magnitude of the observed difference both on the threshold and slope parameters of the item. Items that displayed mode effects were classified as candidate DIF items during the second step; items that did not show any mode effects were set as anchor items.

In the second step, for items labeled as candidate DIF, separate parameters were estimated across mode conditioned on group ability using the anchor set. In this manner, any mode effects was captured within the IRT parameters. During form assembly, effort was taken to avoid using any items showing a mode effect. If any items with mode effects were used, these differences in difficulty or discrimination were then accounted for in the raw-to-scale score conversion tables generated for each form. Through these procedures item parameters from all forms and modes are said to be on the same IRT scale, and by generating separate raw-to-scale tables, any form and mode effects present across alternate forms are accounted for and scale scores are directly comparable independent of form administered.

7.3.2. Parallel Forms Test Characteristic Curves (TCC)

Figure 7-1 through Figure 7-3 show TCCs computed from operational administration item parameters for parallel forms. The TCC plots show the expected score for each form plotted over a theoretical ability range from -4 to 4. The goal during form building was to have identical TCCs for alternate forms across the entire ability range. TCCs for alternate forms across grades showed small variations at different sections along the ability scale during operational administration. Small variations in TCCs of alternate forms were tolerated and accounted for in the raw-to-scale score tables. Also, students' experiences were not noticeably different, and no artificial restriction of range was imposed by taking a form that was differentially too easy or hard. These TCCs for parallel forms follow the same general pattern as those constructed from field test data in Figure 4-3 through Figure 4-5 except grade 5, in which Forms C/O were slightly easier compared to the other forms. Major differences between the TCCs from operational and field test administration were the gradient of the operational TCCs was slightly lower and the steepest section of the TCCs from the operational analysis was slightly shifted to the left of the ability scale, indicating the forms became easier.

When comparing the alternate form TCCs based on the operational item parameters, a couple of observations can be made. First, the TCCs between the modes of administration, say A and M, overlaid to each other suggesting similar Psychometric characteristics of the forms and equivalent samples taking the test. Second, the TCCs across forms were also overlaid for most ability ranges, except for grade 5 Form C/O where the TCCs were toward the left from the other forms, indicating the Form C/O was easier than the other two forms. Following the 2012-13 administration, Forms C and Form O were retired from operational rotation and two new forms (Paper: Form D and Computer based: Form P) were created and introduced in the operational administration in 2013-14 (see Section 7.5). Because NCDPI uses a pre-equated model that ensures parameters from all parallel forms are located on the same IRT scale, any difference across forms are corrected by calibrating each form separately based on equivalent group samples and by creating a separate raw-to scale score table.

Figure 7-1 TCCs for Grade 5 Science Operational Forms A, B, C, M, N and O

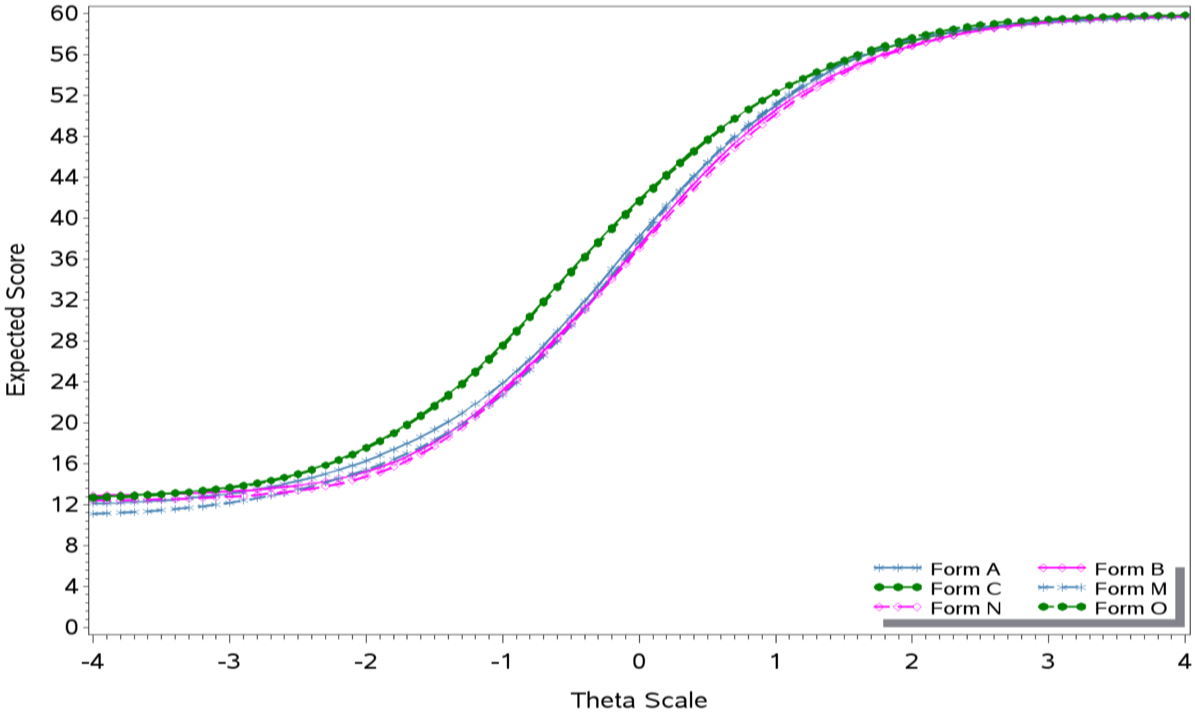


Figure 7-2 TCCs for Grade 8 Science Operational Forms A, B, C, M, N and O

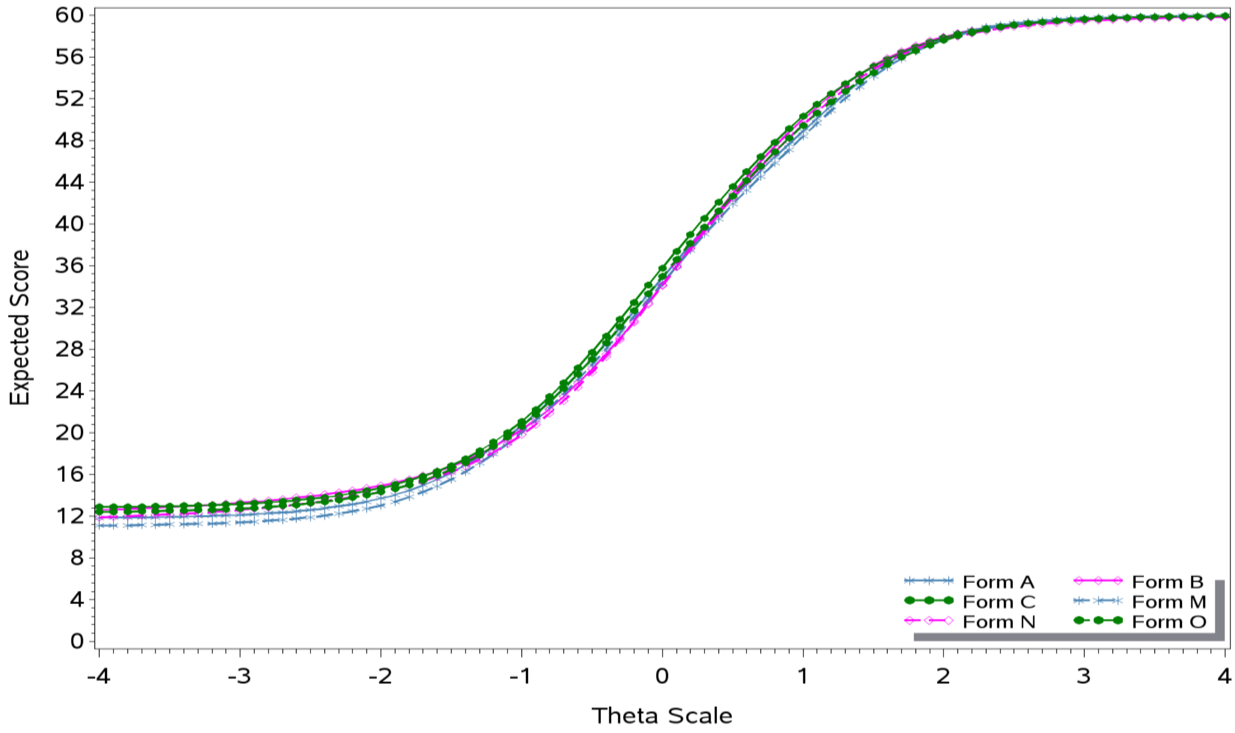
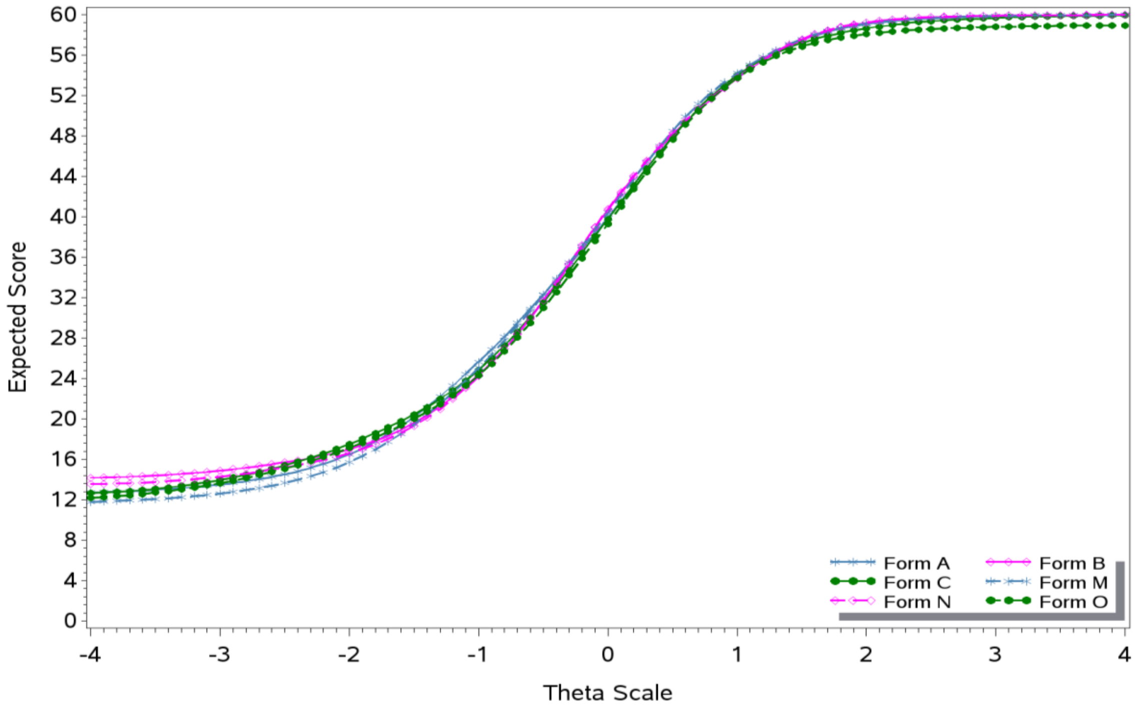


Figure 7-3 TCCs for Biology Operational Forms A, B, C, M, N and O



7.3.3. Measurement Precision-Test Information Function and Conditional Standard Error

In CTT, the concept of reliability is at the center of evaluating test form. Test reliability as defined under CTT has two important drawbacks which have also received considerable attention (Hambleton & Swaminathan, 1985):

- The reliability coefficient is group dependent and, hence, has limited generalizability
- The standard error of measurement is a function of the reliability coefficient and assumes equal error across the entire scale.

The IRT test information function (TIF) offers a viable alternative to the CTT concepts of reliability and standard error. In IRT, measurement precision is defined independently of examinee samples and can be defined at specific levels of the scale. The relative contribution of each item to the overall test precision can be directly evaluated. The general rule is that the test should be most informative around crucial decision points along the scale, such as proficiency cut scores. Figure 7-4 to Figure 7-6 show TIFs by forms with their associated standard error of measurement. Because the NCDPI used TCCs as targets for building alternate forms, the goal was to select items that minimize the differences between TCCs of alternate forms. As a result the displayed TIFs for alternate forms are not as closely uniform as the TCCs. The implication is that relative information of alternate forms vary slightly. But overall, the forms provide similar information in the middle of the ability ranges.

The standard error is inversely related to TIF. As indicated in the figures, the standard errors are the lowest in the middle of the distribution and are larger towards the extreme (Figure 7-4 through Figure 7-6).

Figure 7-4 Science Grade 5 Test Information Functions and Standard Errors for Operational Forms

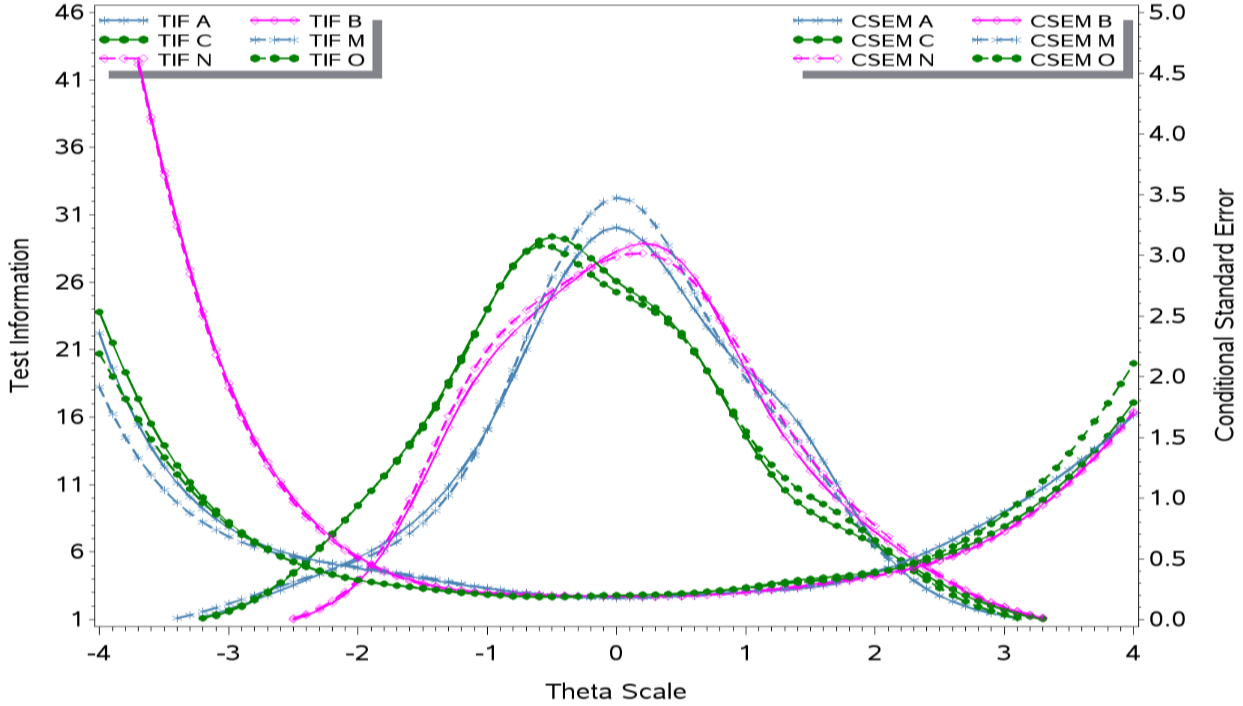


Figure 7-5 Science Grade 8 Test Information Functions and Standard Errors for Operational Forms

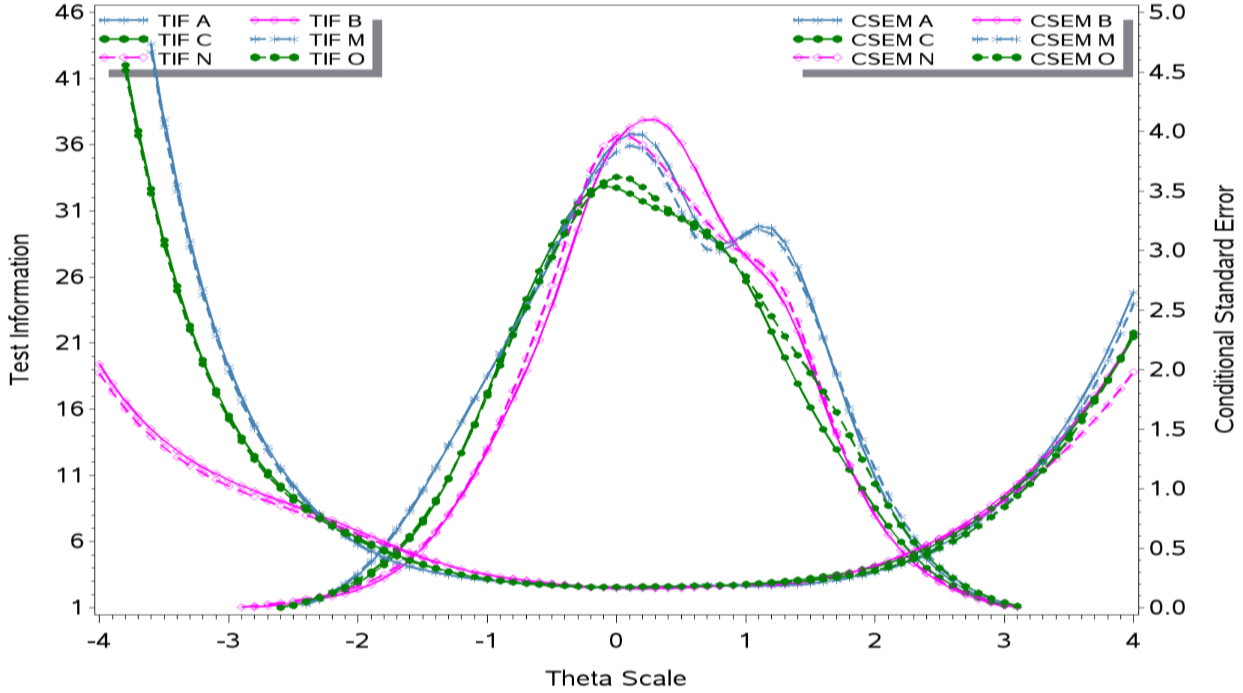
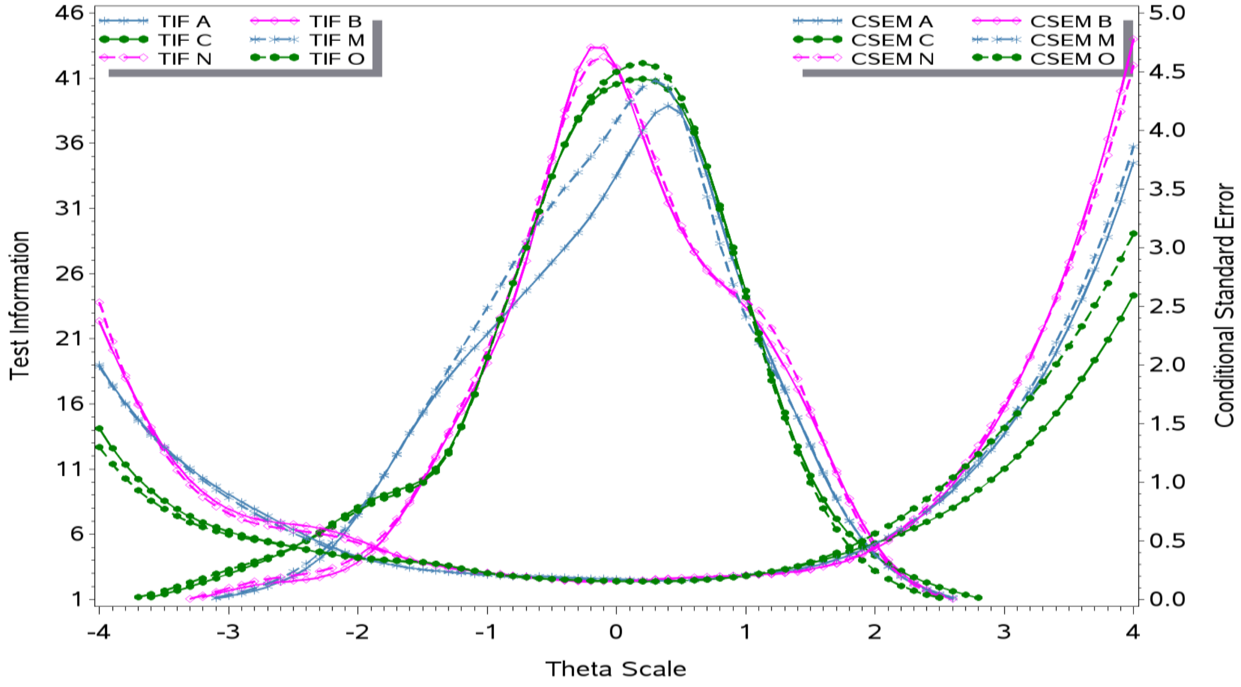


Figure 7-6 Biology Test Information Functions and Standard Errors for Operational Forms



7.4 Item Parameter Drift Between Field Test and Operational Administration

The rationale for delaying scores from the first operational administration was the hypothesis that item parameters may drift from field test administration to operational administration. The NCDPI conducted statistical analysis to justify using operational item parameters during standard setting instead of field test data. The reason was that operational parameters and scale scores would provide stable data for setting a baseline. Results from these studies provided evidence in support of the hypothesis of parameter drift and the NCDPI decision to use operational data in conducting the standard setting study. Table 7-2 presents comparison of form-level average CTT summary statistics (p-values and biserials) from the field test and operational administration.

Table 7-2 Average CTT Statistics for Science EOG and EOC 2012–2013

Grade/Form		No. of Items	Field-Test CTT Summary		Operational CTT Summary		
			P-value	Biserial Corr.	P-value	Biserial Corr.	Reliability (Alpha)
5	A	60	0.53	0.38	0.65	0.39	0.90
	B	60	0.53	0.37	0.67	0.38	0.90
	C	60	0.53	0.37	0.68	0.39	0.90
	M	60	0.51	0.38	0.62	0.39	0.91
	N	60	0.52	0.37	0.61	0.39	0.90
	O	60	0.53	0.36	0.66	0.39	0.90
8	A	60	0.50	0.42	0.61	0.42	0.91
	B	60	0.53	0.40	0.60	0.42	0.91
	C	60	0.53	0.40	0.62	0.41	0.91
	M	60	0.50	0.42	0.57	0.41	0.92
	N	60	0.52	0.40	0.58	0.40	0.92
	O	60	0.52	0.41	0.59	0.40	0.92
Biology	A	60	0.55	0.45	0.66	0.43	0.92
	B	60	0.55	0.44	0.67	0.43	0.92
	C	60	0.55	0.46	0.66	0.42	0.92
	M	60	0.54	0.47	0.65	0.42	0.92
	N	60	0.55	0.44	0.65	0.42	0.92
	O	60	0.54	0.45	0.66	0.42	0.92

The general trend was that the average p-value increased from field test to operational administration ranging from 0.06 to 0.12 across all EOG and EOC science tests. This indicated that students' performance on test items on average was higher than estimated from the field test data, sometimes significantly. The reliability of the operational forms ranged from 0.90 to 0.92, which is reasonable for tests of this length.

IRT parameters calibrated using field test data and again after the operational administration are presented in Table 7-3.

Table 7-3 Average IRT Statistics for Science EOG 2012–2013

Grade/Form		No. of Items	Field-Test IRT Summary			Operational IRT Summary		
			a	b	c/g	a	b	c/g
5	A	60	1.38	0.49	0.22	1.34	-0.17	0.20
	B	60	1.26	0.53	0.22	1.36	-0.13	0.20
	C	60	1.30	0.54	0.23	1.40	-0.33	0.20
	M	60	1.36	0.56	0.21	1.31	-0.19	0.18
	N	60	1.27	0.57	0.21	1.36	-0.12	0.20
	O	60	1.28	0.55	0.22	1.38	-0.35	0.20
8	A	60	1.51	0.55	0.21	1.62	0.09	0.20
	B	60	1.40	0.51	0.22	1.49	0.06	0.21
	C	60	1.46	0.55	0.23	1.49	0.06	0.21
	M	60	1.50	0.57	0.20	1.56	0.10	0.18
	N	60	1.38	0.52	0.22	1.45	0.07	0.19
	O	60	1.45	0.57	0.23	1.51	0.08	0.21
Biology	A	60	0.99	0.37	0.23	1.58	-0.29	0.21
	B	60	0.98	0.41	0.24	1.66	-0.20	0.23
	C	60	0.97	0.31	0.22	1.68	-0.30	0.20
	M	60	1.02	0.35	0.22	1.60	-0.33	0.19
	N	60	0.96	0.38	0.24	1.66	-0.21	0.22
	O	60	0.97	0.30	0.22	1.67	-0.34	0.20

A similar trend as noted in the p-values in Table 7-2 was confirmed by the IRT b-parameter (Table 7-3). The ICCs from the post administration calibration on average shifted to the left, indicating that the items were perceived as less difficult for students during the operational administration. A complete distributional summary of the difference in IRT difficulty parameters (b-parameters) between operational and field test administration is shown using boxplots in Figure 7-7 through Figure 7-9. The middle 50% (25th to 75th percentile) of the differences across all forms by grades is shifted to the left of 0, indicating that the b-parameter for most items was smaller from the field test to the operational administration. This further suggests that students performed higher on the test during operational administration.

Figure 7-7 Grade 5 Science b-parameter Difference Operational and Field Test

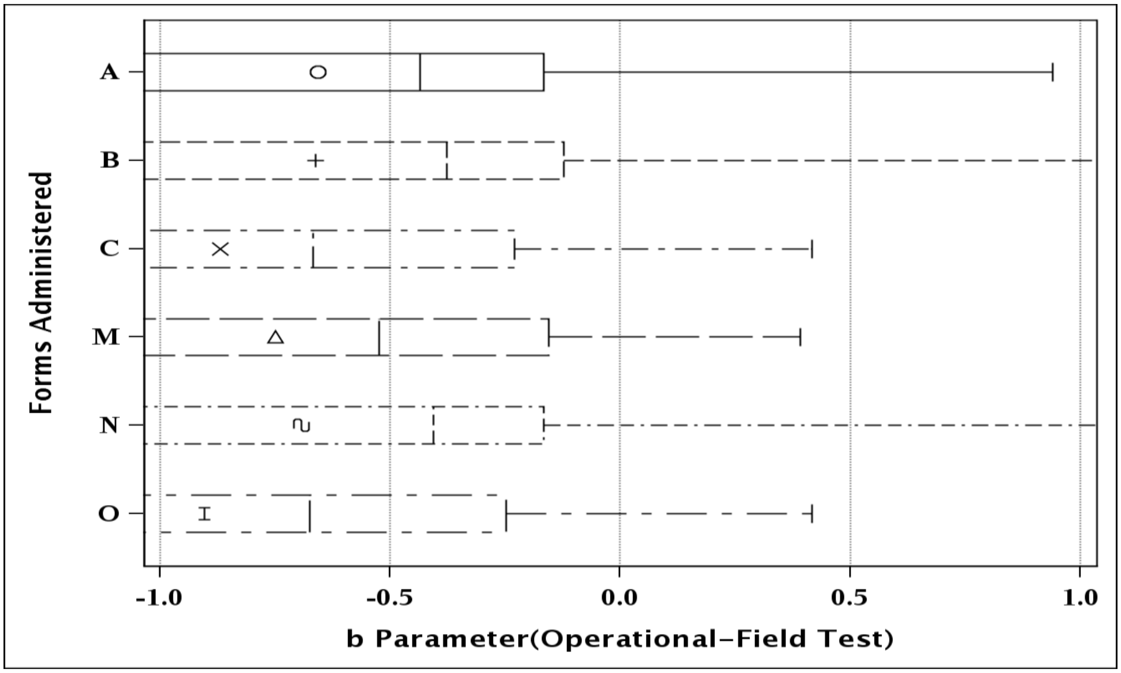


Figure 7-8 *Grade 8 Science b-parameter Difference Operational and Field Test*

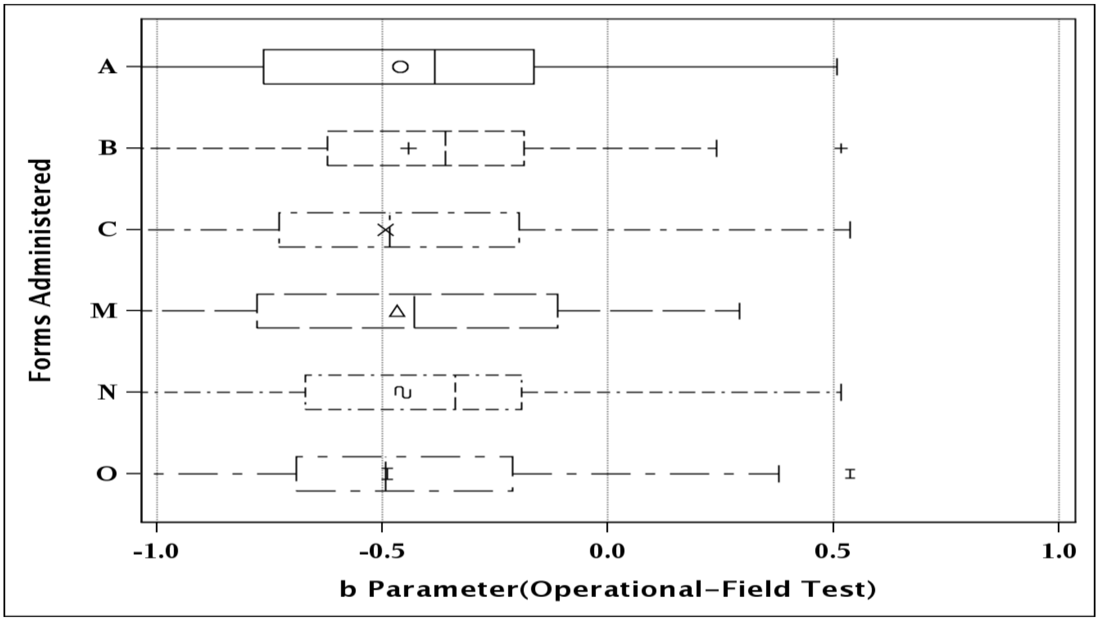
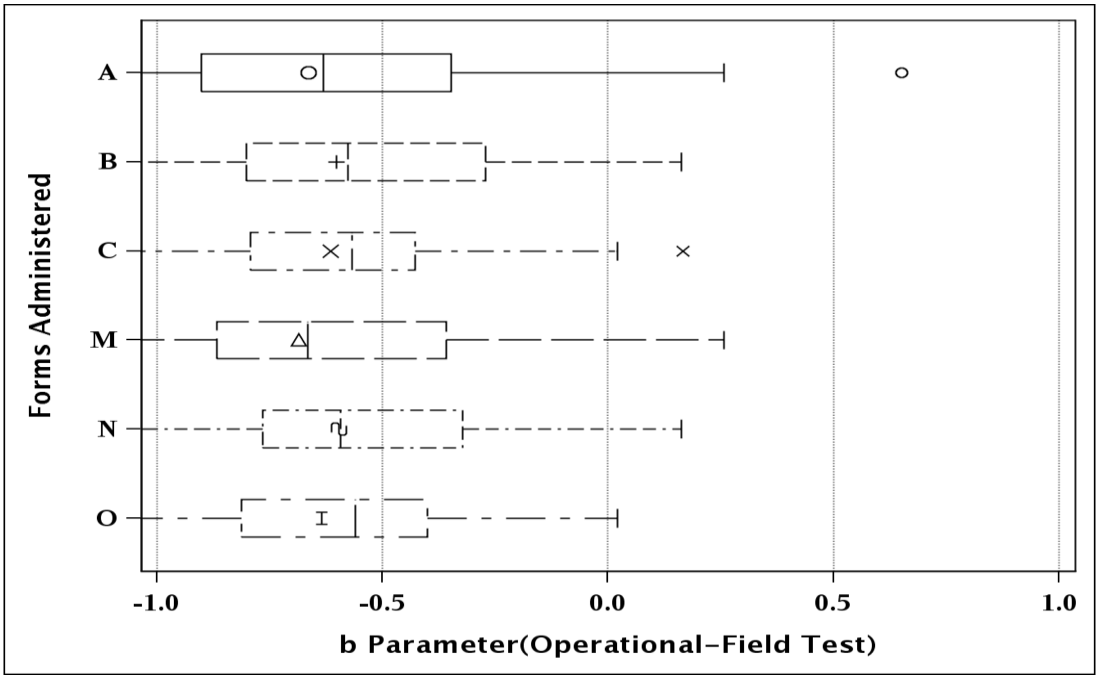


Figure 7-9 Biology b-parameter Difference Operational and Field Test



To summarize the exact magnitude of the differences in parameter drift, the standardized mean differences of the p-values and b parameter were computed using a variation of the effect size statistics.

$$effect\ \ size = \frac{\overline{\chi}_{op} - \overline{\chi}_{ft}}{((sd_{op} + sd_{ft})/2)} \tag{7-1}$$

- where $\overline{\chi}_{op}$ and sd_{op} are mean and standard deviation from post operational item parameter
- and $\overline{\chi}_{ft}$ and sd_{ft} are mean and standard deviation from field test item parameter

Table 7-4 shows the effect size summary computed for CTT p-value and IRT b-parameter between field test and operational statistics. Using Cohen (1998) classification, most of the effect sizes for p-values ranged from 0.37 to 0.88, and b-parameter ranged from -0.49 to as large as -0.81, indicating on average a medium-to-large effect from field test to operational parameters estimation.

Table 7-4 Science Effect Size Summary of Operational and Field Test Statistics

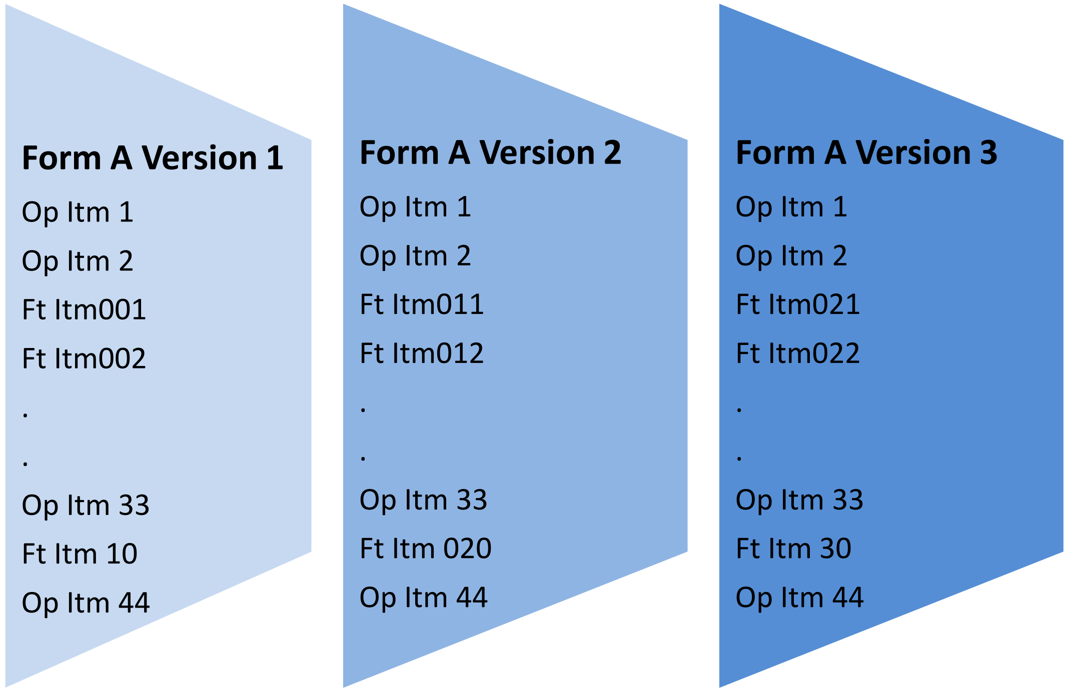
Grade/Form		Operational Items	P-value Standardized Mean Difference	Threshold Standardized Mean Difference
5	A	60	0.72	-0.67
	B	60	0.88	-0.55
	C	60	0.87	-0.77
	M	60	0.64	-0.75
	N	60	0.57	-0.58
	O	60	0.78	-0.81
8	A	60	0.61	-0.52
	B	60	0.49	-0.49
	C	60	0.60	-0.59
	M	60	0.46	-0.52
	N	60	0.37	-0.50
	O	60	0.41	-0.57
Biology	A	60	0.74	-0.74
	B	60	0.88	-0.76
	C	60	0.70	-0.65
	M	60	0.67	-0.77
	N	60	0.75	-0.75
	O	59*	0.68	-0.69

**One item was dropped from the test form*

7.5 Ongoing Form Maintenance and Item Development

As indicated in chapter 1 and 7 of this report the NCDPI relies on a continuous embedded field testing plan for ongoing item development. During operational administration field test items are embedded with operational items and administered to students. For both EOG and EOC science, a total of 15 field test items are embedded in each operational version. For each operational test form, distinct versions are created following a predefined embedding plan See Figure 7-10 for a schematic example.

Figure 7-10 Item Field Test Embedding Plan



The figure shows field test items (Ft ItmNo.) embedded within operational items (Op ItmNo.). Each version of Form A is differentiated from the next version by the distinct set of field test items embedded. The number of versions created for each form depend on future form building needs and overall number of students expected to be administered the EOG or EOC test. During operational administration, versions and forms are spiraled randomly within each classroom across the state. This ensures field test items are administered to random subset of students and subsequent item parameters are generalizable to the entire state population for the given grade level.

7.6 Development of Forms D and P for Grade 5 Science

As indicated earlier, grade 5 forms C and O showed a sign of scale drift from the field test to operational administration and from the other operational test forms. The TCCs of forms C/O were toward the left from forms A, B, M, and N. In order to use two paper and two computer-based forms alternately in each subsequent administration, the NCDPI decided to create new forms (paper: Form D, and computer-based: Form P). These forms conformed closely to the grade 5 science test specifications. Table 7-5 shows content standards distributions of forms D and P and Table 7-6 shows item types for Form P which closely matched with computer-based forms M and N.

Table 7-5 Content Standards and Weight Distribution of Form D/P, Grade 5 Science

Domain	Blue Print (%)	Form D		Form P	
		No. of Items	%	No. of Items	%
Forces and Motion (5.P.1)	13–15	8	13.3	8	13.6
Matter: Properties and Change (5.P.2)	12–14	8	13.3	8	13.6
Energy: Conservation and Transfer (5.P.3)	11–13	6	10.0	6	10.2
Earth Systems, Structures and Processes (5.E.1)	15–17	10	16.7	10	16.9
Structures and Functions of Living Organisms (5.L.1)	14–16	10	16.7	10	16.9
Ecosystems (5.L.2)	14–16	10	16.7	9	15.3
Evolution and Genetics (5.L.3)	13–15	8	13.3	8	13.6
Total	100	60	100	59*	100

**One TE item was dropped from the form P*

Table 7-6 Online Form P-Content Standards by Item Type, Grade 5 Science

Domain	Form P			
	DD	MC	TI	Total
Forces and Motion (5.P.1)	0	8	0	8
Matter: Properties and Change (5.P.2)	0	5	0	5
Energy: Conservation and Transfer (5.P.3)	0	8	1	9
Earth Systems, Structures and Processes (5.E.1)	1	9	0	10
Structures and Functions of Living Organisms (5.L.1)	0	10	0	10
Ecosystems (5.L.2)	0	9	0	9
Evolution and Genetics (5.L.3)	0	8	0	8
Total	1	57	1	59

Forms D and P were used operationally for the first time in the 2014–15 administration. The classical statistics (mean p-values and biserial correlation), as well as the reliability of the forms (Cronbach alpha) based on 2014–15 operational student responses are shown in Table 7-7. The mean p-values and biserial correlations are in the same range as the other grade 5 science operational forms. Moreover, Cronbach alpha of forms D and P are 0.92 and 91 respectively, indicating reasonably high reliability of the new forms.

Table 7-7 Average P-value and Reliability Statistics for Grade 5 Science Forms A, B, M, N, D and P

Grade/Form		No. of Items	Operational CTT Summary		
			P-value	Biserial Corr.	Reliability (Cronbach Alpha)
5	A	60	0.65	0.39	0.90
	B	60	0.67	0.38	0.90
	D	60	0.67	0.38	0.92
	M	60	0.62	0.39	0.91
	N	60	0.61	0.39	0.90
	P	59*	0.64	0.38	0.91

**One TE item was dropped from the form P*

During the item calibration, one TE item in Form P showed a sign of mode effect. Expert review of the item suggested that the item had a display issue. NCDPI decided to drop the item from the form, therefore, Form P contains only 59 items. The average IRT statistics (a, b, and c) of the forms are shown in Table 7-8, with Forms D and P highlighted. The mean parameter values are similar to forms A, B, M, and N. The TCCs of Forms D and P plotted together with forms A, B, M, and N are shown in Figure 7-10, and TIFs and SEs are shown in Figure 7-11. The TCCs are closely overlapped, indicating that the new forms are psychometrically similar with other grade 5 science operational forms in terms of difficulty of the tests across the ability ranges. The TIFs and CSEMs indicated that the new forms (D and P) provided higher information to wider ability ranges, with slightly lower information and higher SEs in the middle of the distribution compared to the forms A, B, M, and N.

Table 7-8 Average IRT Statistics for Grade 5 Science Forms A, B, D, M, N and P

Grade/Form		No. of Items	Operational IRT Summary		
			a	b	c/g
5	A	60	1.34	-0.17	0.20
	B	60	1.36	-0.13	0.20
	D	60	1.37	-0.13	0.20
	M	60	1.31	-0.19	0.18
	N	60	1.36	-0.12	0.20
	P	59*	1.36	-0.14	0.20

**One TE item from form P was dropped*

Figure 7-11 TCCs for Grade 5 Operational Forms A, B, D, M, N and P

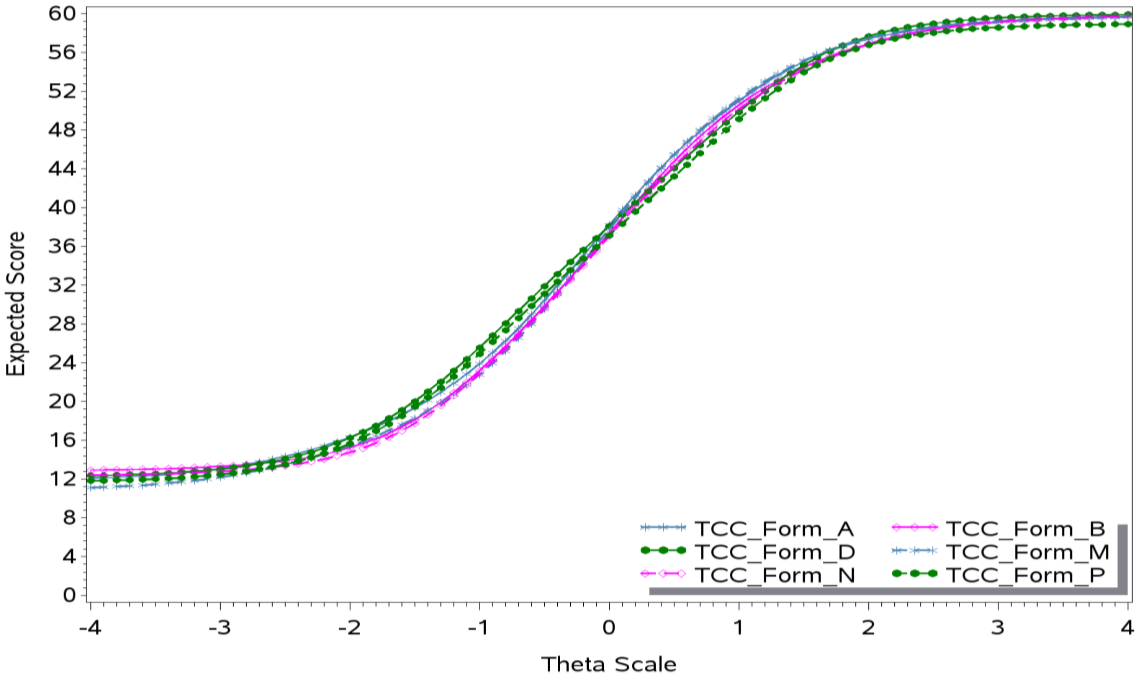
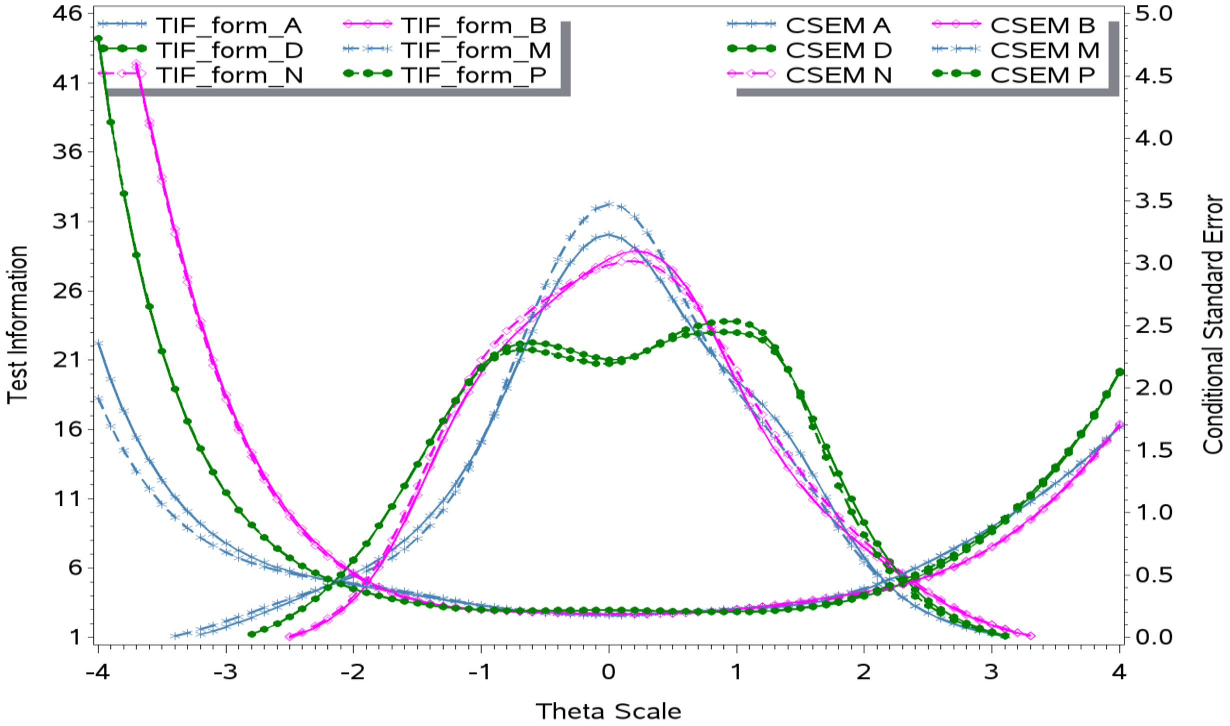


Figure 7-12 TIFs and SEMs for Grade 5 Operational Forms A, B, D, M, N and P



Chapter 8 Standard Setting

Standard 5.21 of the *Standards* (AERA, APA, & NCME, 2014) states that "*when proposed score interpretation involves one or more cut scores, the rational and procedures used for establishing cut score should be documented".* Standard setting is a process used to define achievement or proficiency levels and the cut scores corresponding to those levels with associated proficiency level descriptors (PLDs). A cut score is simply the score that serves to classify students whose score is below the cut score into one level and those whose scores are at or above the cut score into the next and higher level.

8.1 Standard Setting Overview

Standard setting is a process used to define achievement or proficiency levels. Standard setting is recommended whenever an assessment system undergoes major revisions or changes to the underlying standards, as was the case in 2010 with the adoption of the new *North Carolina Essential Standards* for Science and the development of The READY accountability assessment system to measure students' College-and-Career readiness. In July 2013 after the first operational administration of EOG and EOC, the NCDPI contracted with Pearson Education to conduct a standard setting workshop in order to recommend cut scores and achievement levels for the newly developed Science EOG and EOC assessments.

Three panels (Grade 5 Science, Grade 8 Science, and Biology) of North Carolina Science educators convened to make cut score recommendations for the EOG and EOC assessments. A total of 53 (16 for grade 5, 17 for grade 8, and 20 for Biology) North Carolina Science educators and postsecondary educators convened in Chapel Hill, North Carolina between July 22 and July 26, 2013, using the item mapping method to make content-oriented recommendations for cut scores. Science teachers with exceptional children or ELL experience were recruited. The item mapping procedure (Lewis, Green, Mitzel, Baum, & Patz, 1998; Mitzel, Lewis, Patz, & Green, 2001) based on ordered item booklets prepared by the NCDPI staff was used by panelists in a series of rounds to recommend cut scores. All training during the standard setting workshop was facilitated by Pearson Education staff. The executive summary of the standard setting report is available in Appendix 8-A Standard Setting Report, and the full report can be found at the following link:

http://www.ncpublicschools.org/docs/accountability/testing/technotes/sstechreport1213.pdf

At the conclusion of the standard setting workshop, three recommended cut scores with four achievement levels were presented to the North Carolina State Board of Education for adoption. An abbreviated version of the final standard setting study prepared by Pearson[g] for the NCDPI is presented in the ensuing sections.

8.1.1 Panelists Background

All panelists were asked to provide voluntary demographic information. A brief summary of panelist characteristics and major demographic variables are presented in Table 8-1 through Table 8-4. Complete panelist demographics are provided in the full standard setting technical report.

The panelists' years of experience as educators are summarized in Table 8-1. As illustrated by the table, the educational experience of the 53 panelists ranged from less than 5 years to above 21 years, resulting in a very diverse group of educators for the standard setting.

Table 8-1 Panelist Experience as Educators

Panel	N	Years in Current Position					
		1–5	6–10	11–15	16–20	21+	NR
Science 5	16	1	5	5	5	0	0
Science 8	17	3	6	5	1	2	0
Biology	20	2	5	6	4	3	0

Note: NR = no response.

The panelists' professional backgrounds are summarized in Table 8-2. Teachers reported as teaching on or off grade are reported in the context of their committee. For example, panelists who primarily teach a grade level outside of the panel's range (e.g., a Grade 8 teacher who participated in the science 5 panel) are listed in the off-grade column. Finally, other groups of educators are summarized in the remaining columns of these tables. As shown in the table, all grade levels were represented by panels, including a variety of professional backgrounds.

[g] Copyright © 2013, Pearson and North Carolina Department of Public Instruction

Table 8-2 Panelist Professional Background: Single-Grade Panels

Panel	ON	OFF	SED	SPE	COA	HED	OTH	RET	NR
Science 5	7	4	0	1	0	0	4	0	0
Science 8	11	4	1	0	0	0	0	0	1
Biology	17	0	0	1	0	1	0	1	0

Note: ON = on-grade, OFF = off-grade, SED = special education, SPE = specialist, COA = coach, HED = higher education, OTH = other, RET = retired, NR = no response.

In addition to reporting their own demographic characteristics (Table 8-3), panelists were asked to report their district geographic location within the state (Table 8-4), as well as district size and community setting (Table 8.5). As demonstrated by the information provided in these tables, panelists making up the standard setting committees showed representative diversity of geographic regions, district sizes, and community settings across North Carolina.

Table 8-3 Panelist Gender and Ethnicity

Panel	Gender			Ethnicity						
	F	M	NR	AA	AS	HI	NA	WH	MU	NR
Science 5	16	0	0	4	0	0	0	12	0	0
Science 8	13	4	0	0	1	1	1	13	1	0
Biology	17	3	0	1	0	1	0	18	0	0

Note: F = female, M = male, NR = no response, AA = African American, AS = Asian, HI = Hispanic, NA =Native American, WH = white, MU = multiple responses, NR = no response.

Table 8-4 Panelist Geographic Region

Panel	C	NC	NE	NW	SC	SE	SW	W	MU	NR
Science 5	4	2	0	0	2	1	5	2	0	0
Science 8	5	1	1	1	4	2	2	1	0	0
Biology	3	4	1	3	1	2	5	0	0	1

Note: C = central, NC = north central, NE = northeastern, NW = northwestern, SC = south central, SE = southeastern, SW = southwestern, W = western, NR = no response.

Table 8-5 Panelist District Characteristics

	District Size				Community Setting			
Panel	NR	SM	MD	LG	NR	RU	SU	W
Science 5	0	2	7	7	0	7	6	3
Science 8	0	3	8	6	0	8	4	5
Biology	1	4	6	9	1	6	8	5

Note: NR = no response, SM = small, MD = medium, LG = large, RU = rural, SU = suburban, UR = urban

8.1.2 Vertical Articulation Committee

Each standard setting breakout session room, which contained between 16 and 20 total panelists, was arranged to include three tables. At various points throughout the process, panelists within a committee broke up and worked together in groups of between 5 and 7 individuals at each table. Each of the three tables had at least one designated table leader who was selected by the NCDPI and trained by the lead facilitator. At the conclusion of the standard setting activities, table leaders were asked to stay for one additional task: participating in the vertical articulation committee. Demographic characteristics of the vertical articulation committee were collected by way of survey (see Appendix E in the Standard Setting Report).

8.1.3 Method and Procedure

A total of nine panels set standards for the 17 grades and subjects (ELA: grades 3–8 and English II, math: grades 3–8 and Math I, science: grades 5 and 8, and biology). For the single-grade science committees (Science 5, Science 8, and Biology), panelists recommended standards for a single grade/subject. The single-grade panels convened between July 24 and 25, 2013. For a full agenda of the various panels refer to *Appendix E* (Standard Setting Report).

8.1.4 Table Leader Training

For the single-grade panels, table leader training was held during the morning of Wednesday, July 24. During this training session, table leaders were introduced to the standard setting facilitators, trained on their role in the standard setting process, and received a general introduction and instruction on the item mapping process. Following table leader training, representatives of the NCDPI and Pearson Education presented an opening session to all panelists. The single-grade opening session occurred on July 24.

8.1.5 Opening Session and Introductions

After the conclusion of the opening session, panelists dispersed to their breakout session meeting rooms. Each panel convened in a separate breakout session room to complete the required standard setting activities. Each panelist was provided a folder containing secure materials to be used throughout the meeting. Panelists were asked to mark all materials they received with their unique assigned panelist identification number. Prior to beginning the standard setting activities, panelists signed security agreements and completed a demographic information survey. Concurrent with this activity, panelists introduced themselves to their colleagues within their breakout session meeting room.

8.1.6 Achievement Level Descriptors

Following committee introductions, the single-grade panels spent a portion of July 24 to write achievement level descriptors (ALDs) for their single assigned assessment, and then the panels moved on to other standard setting activities that day. Breakout session facilitators provided panelists with ALD training that covered the purpose of ALDs, and facilitators shared several real-world examples demonstrating characteristics of effective ALDs. Panelists were trained on strategies to link ALDs to the test blueprint and curriculum standards, both of which were made available to panelists. Panelists were provided draft ALDs from NCDPI, which included general, policy-oriented statements about student achievement across levels. Panelists were tasked with adding content-oriented statements to the draft ALDs to further define student achievement in the context of the assessment. The panels' final drafted ALDs, which were turned over to NCDPI for review and future revisions, as deemed necessary, are provided in Appendix D of the Standard Setting Report.

8.1.7 Setting Standards

“Just Barely” Level Descriptors

Following ALD writing activities, panelists performed tasks to set standards for their assigned subject areas and grades. Panelists began by drafting and discussing “just barely” level descriptors: statements describing performance expectations for students who are *just barely* at the three cut points separating the four achievement levels. The “just barely” level descriptors are critical to standard setting for two reasons. First, discussing characteristics of students who are just barely at a particular cut point dividing two adjacent achievement levels aids panelists in developing a strong understanding of the differences in observed student performance across achievement levels. Second, in subsequent steps occurring during the standard setting process, panelists referred to the "just barely" level descriptors to anchor their judgments to a common understanding of achievement expectations.

Ordered Item Book Review

Next, panelists completed a "test-taking" activity to familiarize themselves with the assessment's test items, which was accomplished by reviewing the ordered item book (OIB). NCDPI staff produced the OIBs, which contained items used during the spring 2013 administration. Each page of the OIB contained one item, and items were ordered in ascending empirical difficulty as estimated from actual student performance such that the first page of the OIB included the least difficult item, and the last page of the OIB contained the most difficult item. Panelists were instructed to review and answer the items in the OIB. Each ordered item book was accompanied by an item map, which contained useful item-level information such as OIB page number, key, reading selection ID (for tests with reading selections only), and linked content standard. After completing the OIB review, panelists were given an opportunity to share their thoughts on and reactions to the test's content with their colleagues in the breakout session.

8.1.8 Standard Setting Training and Practice Round

Following the completion of the ordered item book review, the breakout session facilitator provided panelists with training on the standard setting process. The item mapping procedure (Lewis, Green, Mitzel, Baum, & Patz, 1998; Mitzel, Lewis, Patz, & Green, 2001) is the judgmental process that was used in this standard setting. According to this procedure, panelists are asked to identify the item in the ordered item book that is the last item that a student who is just barely at a given achievement level should be able to answer correctly more often than not. The locations for the items in the ordered item book were established using a guess-adjusted response probability of two-thirds (or 2/3), representing the point on the item characteristic curve at which the probability of a correct response is two-thirds of the way between the curve's lower asymptote and 1.0.

Following item mapping methodology training, panelists completed a practice round of judgment. Using a shortened ordered item book and item map, each of which were comprised of 10 items spanning the empirical difficulty range observed in the full OIB, panelists practiced the item mapping methodology by reading the items in the practice OIB and placing a single cut for Achievement Level 3 only. The purpose of the practice round was to reinforce panelists' understanding of the item mapping process by allowing them to apply the concepts covered during the standard setting training. Following the practice round, the breakout session facilitator led a short committee-wide discussion to gather panelists' thoughts on and reactions to the item mapping procedure, as well as to respond to any lingering questions or misunderstandings.

Round 1 Standard Setting

Once all questions from the practice round were addressed, panelists began the standard setting process. For each assessment, panelists set three recommended cut scores, which separate test scores into four distinct achievement level categories. Prior to beginning the standard setting activity, panelists were instructed to complete a short readiness survey on which panelists affirm that they understand the process and feel prepared to begin (see Appendix F of the Standard Setting Report). Panelists were encouraged to seek clarification from the breakout session facilitator on any remaining questions or concerns, should they have any, prior to beginning the first round of judgment. Upon unanimous positive affirmation of readiness to proceed, committees began the standard setting process. The standard setting process consisted of three rounds of judgment. Panelists completed readiness surveys affirming their understanding of the process and willingness to proceed prior to beginning each of the three rounds. The committees were instructed to set their cuts in order starting at Level 2, then at Level 3, and finally at Level 4.

Panelists worked independently to place their bookmarks across all three rounds of judgment. For each round, panelists were instructed to place three bookmarks within the ordered item booklet corresponding to their cut score recommendations: one for Level 2, one for Level 3, and one for Level 4. Panelists wrote the page numbers corresponding to their three recommended cut scores on the recording sheet (see Appendix G of the Standard Setting Report). The breakout session facilitator collected all of the committees' recording sheets at the conclusion of each round of judgment and handed them over to the data analysts for data entry and processing.

Behavioral Descriptors

Panelists were provided with feedback data after each round of judgment; however, due to the processing time requirements, panelists engaged in other activities while awaiting feedback data in order to avoid long periods of downtime for panelists between rounds of judgment. For single-grade committees, panelists developed behavioral descriptors between rounds 2 and 3; for the three-grade committees, panelists completed this activity between rounds 1 and 2. Panelists wrote brief phrases or sentences that described observable, content-oriented behavioral characteristics of students across the score scale. The breakout session facilitator managed the discussion on this topic and recorded the panel's behavioral descriptions. Although not a primary output of emphasis of the standard setting meeting, these behavioral descriptors created by North Carolina educators were collected by the NCDPI for a longer-term goal of eventually being incorporated into an integrated feedback system designed to offer stakeholders more concrete feedback on student performance beyond scores and achievement level outcomes.

To help guide panelists' discussions while they created behavioral descriptions, panelists were provided with content domain item maps. The content domain item map was similar to the OIB item map in that it provided panelists with useful information on the items in the ordered item booklet, but the content domain item map differed from the OIB item map in several important ways. Whereas the OIB item map presented items in the same order as they appeared in the ordered item booklet, the content domain item map organized items on the page vertically by empirical difficulty (reported on a temporary score scale metric constructed solely for the purposes of this standard setting) and grouped them horizontally into columns by their content domains.

Round 1 Feedback and Discussion and Round 2 Standard Setting

After each round of judgment, panelists were provided with feedback data to consider and discuss. Following round 1, panelists received table-level and panel-level feedback. They were provided the cut scores for each panelist at their table based on the round 1 ratings, in addition to the minimum, maximum, mean, and median cut score at each cut point for that table. In reviewing the judgment agreement data with the other committee members seated at their table, panelists were asked to consider and discuss the following:

- How similar their cut scores were to that of the rest of the table (i.e., is a given panelist more lenient or stringent than the other panelists?)
- If a panelist had cut scores dissimilar to the table, why?
- Do panelists have different conceptualizations of "just barely" level students?

Panelists were instructed by the breakout session facilitator that reaching consensus was not the goal of these discussions, but panelists should share their perspectives to get a feel for why observed cut score judgment differences might exist. The table leaders, with assistance from the breakout session facilitator, helped guide this discussion so that all panelists at their table had an opportunity to share thoughts and perspectives with the other panelists at the table. Panelists compared bookmarks and discussed the differences between these bookmarks. Using data provided in the feedback handouts, panelists discussed their judgments related to items in the range between the highest and lowest bookmarks for each achievement level. An example of the rating agreement feedback data provided to each table of panelists is provided in Table 8-6 **.**

Table 8-6 Example Table-Level Rating Agreement Feedback Data

Judge	Level 2 Cuts	Level 3 Cuts	Level 4 Cuts
A1	41	72	82
A2	30	63	80
A3	23	55	75
A4	22	62	78
A5	43	70	82
A6	37	73	82
Mean	33	66	80
Median	34	67	81
Minimum	22	55	75
Maximum	43	73	82

Following table-level discussions, panelists were provided committee-wide feedback data and engaged in a similar conversation, moderated by the breakout session facilitator, at the committee level. As a large group, panelists shared highlights of discussions they held at their tables, and they discussed observed cut score differences across the tables. An example of the committee-level rating agreement feedback data is provided in Table 8-7 .

Table 8-7 Example Committee-Level Rating Agreement Feedback Data

Table	Judge	Level 2	Level 3	Level 4
1	A1	41	72	82
	A2	30	63	80
	A3	23	55	75
	A4	22	62	78
	A5	43	70	82
	A6	37	73	82
2	B7	23	50	66
	B8	22	50	70
	B9	22	49	72
	B10	25	60	72
	B11	25	63	82
	B12	35	68	81
3	C13	22	53	68
	C14	14	42	60
	C15	23	43	68
	C16	23	54	73
	C17	23	55	66
	C18	26	55	72
Overall	Mean	27	58	74
	Median	23	55	73
	Minimum	14	42	60
	Maximum	43	73	82

In addition to the round 1 cut score agreement data, panelists were shown external data to further inform their judgments in subsequent rounds of judgment. Panelists were provided with empirical item difficulty data showing the proportion of all test-takers from the spring 2013 administration who correctly answered each item (i.e., item p-values). The breakout session facilitator also shared with panelists the ACT Explore® cut score, which was linked to the North Carolina assessment by the NCDPI, representing the score point at which students are on track to be College-and-Career Readiness. Finally, the facilitator shared with panelists the expected cut scores obtained by the NCDPI from a recent survey of North Carolina educators.

The cut scores shared with panelists were translated into page numbers in the ordered item book to help facilitate comparisons between the external data and panelists' own cut score judgments (see Table 8-8). For some assessments, the cut score from the teacher survey for Level 2 was lower than the estimated empirical difficulty level associated with the first page of the ordered item booklet. In these instances, the cut was set to page 1.

Table 8-8 Linked Page Cuts from the Teacher Survey and ACT Explore

Assessment	Level 2	Level 3	Level 4	Explore®
Science 5	4	25	57	52
Science 8	4	21	57	67
Biology	9	26	63	*

*Note: No linked ACT Explore® cut scores were provided for the EOC panels.

Following discussion of round 1 cut scores and the providing feedback data, panelists proceeded to the second round of judgment. Following discussion of external feedback data, panelists once again completed readiness surveys and began round 2, using the same procedure that was previously outlined in the description of round 1.

Round 2 Feedback and Discussion and Round 3 Standard Setting

Following round 2, panelists received updated cut score agreement feedback data and engaged in discussions at the table level as well as across the committee. Additionally, panelists were shown a graphical display of student impact data. The impact data displayed the percentages of spring 2013 test takers who would be classified into the four achievement levels based on the panel's median cut score recommendation. Impact was shown for the overall North Carolina test-taking population, and impact was also broken down into gender and ethnicity subgroups. Panelists were given an opportunity to discuss the appropriateness of their cut scores given the current impact data. Following discussion of the round 2 feedback data, panelists completed readiness surveys and proceeded to the third and final round of judgment.

Round 3 Feedback and Discussion

Following round 3, panelists were shown their final recommended cut scores, which were based on the committee's median cut score judgments from this final round of judgment. Panelists were shown impact data, which again included overall impact as well as impact broken down into gender and ethnicity.

8.1.9 Standard Setting Evaluations

After reviewing and discussing the round 3 impact data, panelists completed an evaluation survey capturing their reactions to the final cut score recommendations and associated impact data. The standard setting workshop activities concluded at this point for the single-grade committees. Following the conclusion of standard setting activities, all panelists were dismissed with the exception of table leaders, who attended the vertical articulation session on Friday, July 26.

8.2 Vertical Articulation

Table leaders from each committee convened in a single room to participate in the vertical articulation session. During this session, impact data were compared across grade levels within subject areas (e.g., Grades 5 and 8 science) and also across subjects. Panelists were asked to evaluate and discuss, from a policy perspective, the reasonableness of the committees' content-oriented cut score recommendations and the impact of imposing these achievement expectations on student test scores. Panelists were guided through a process whereby they evaluated the reasonableness of impact for particular grades/subjects, both in isolation and in contrast to other grades and subject areas. Table leaders from each committee were present in the vertical articulation meeting, which allowed them an opportunity to share with the entire group their reflections on the execution of the standard setting procedure as well as the discussions that occurred within their committees.

Following group discussions of the cuts and impact data, the lead facilitator asked the vertical articulation committee if they felt any cut score changes may be appropriate, given the observed patterns of impact data. The lead facilitator projected a spreadsheet with cut scores and impact data, and panelists were permitted to suggest potential revised cut scores to see real-time changes to impact data based on these potential revisions. Following NCDPI's instructions, the lead facilitator did not limit the range of potential cut score changes available to the vertical articulation committee. The lead facilitator provided verbal notice to the panel at any point at which their recommended cut scores (discussed in terms of page numbers) deviated more than +/- 1 standard error of the original median page cut, where the standard error of the median was computed as:

$$SE_{Median} = \frac{\sigma}{\sqrt{N}} \tag{8-1}$$

In addition to the standard error of the median, the lead facilitator also considered the range of the original panel's cut score judgments when engaging the vertical articulation committee in discussion of potential changes to the cut scores. In instances where the vertical articulation committee expressed a desire to explore possible cut scores outside the observed range of content-oriented cut scores recommended by the original panel, the lead facilitator notified the vertical articulation panel of this fact.

Each participant on the vertical articulation panel considered the original recommended cut scores and their impact data as well as other potential cut scores and the changes in impact data associated with these potential cuts. Each member of the vertical articulation committee provided a unique, independent recommendation to either keep or change the cut scores. Consistent with the previous phase of the standard setting meeting, members of the vertical articulation committee completed readiness surveys and unanimously affirmed their understanding of the process and willingness to proceed prior to rendering their final recommendations. The lead facilitator impressed upon the vertical articulation panel that their holistic, policy-oriented cut score recommendations would supplement, not overwrite, the content-oriented cut recommendations provided by the standard setting panels and would provide the North Carolina State Board of Education with additional information to consider when deciding which cut scores to adopt. Each member of the vertical articulation committee provided an independent recommendation to either keep or adjust the cut scores for every grade and subject. Panelists recorded their judgments on provided forms (see Appendix M of the Standard Setting Report) and returned them to the lead facilitator for processing. After completing the vertical articulation process for all grades and subjects, panelists completed an evaluation survey of the vertical articulation process (see Appendix N of the Standard Setting Report).

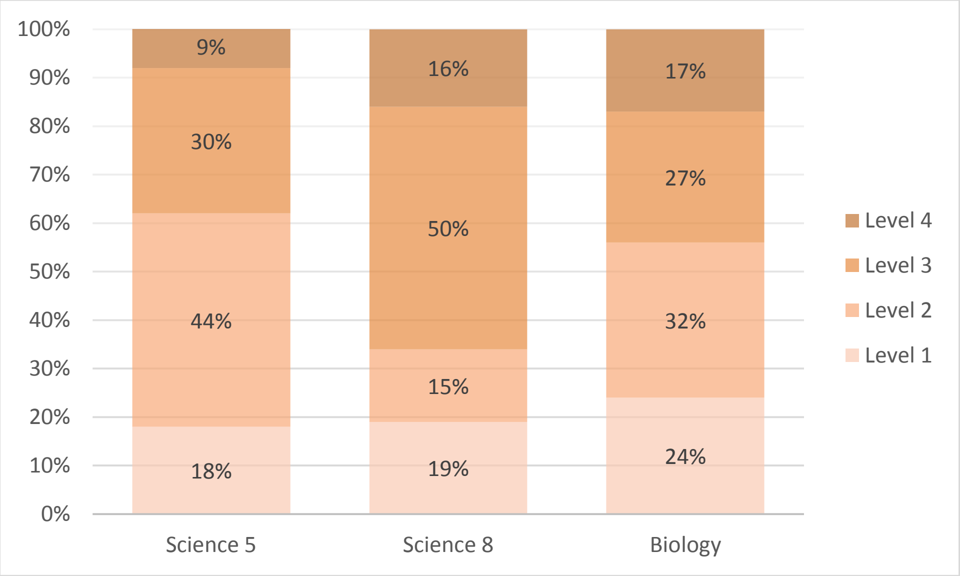
8.3 Standard Setting Results

The standard setting panels' final recommended cut scores, obtained prior to the vertical articulation session, are presented in *Table 8-9*. The reader should note that these cut scores are reported as page numbers within the ordered item book, not raw scores. The NCDPI will translate these page cuts into the final reporting scale in a future study, which will be documented separately from this standard setting technical report. ***Figure 8-1*** displays impact data for the Science EOG and Biology EOC assessments, respectively, based upon these cut score recommendations. Tables and figures showing individual panelists' page cuts across rounds are provided in Appendix I of the full report.

Table 8-9 Pre-Vertical Articulation Page Cuts

Assessment	Level 2	Level 3	Level 4
Science 5	12	45	69
Science 6	6	20	64
Biology	20	47	68

Figure 8-1 Pre-Vertical Articulation Impact Data

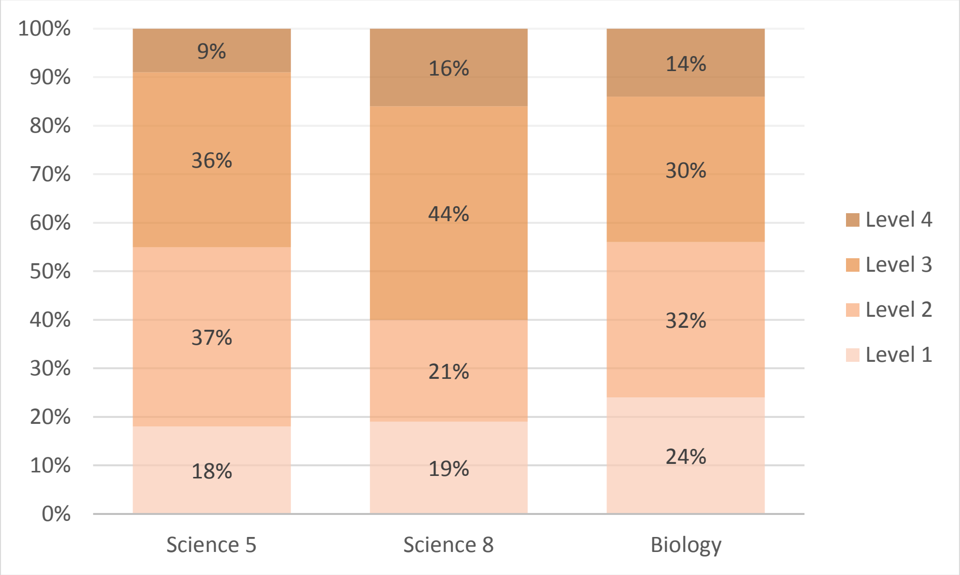


Cut scores obtained following the vertical articulation session are shown in *Table 8-10*, and impact data associated with these recommended cut scores are displayed in *Figure 8-2*.

Table 8-10 Post-Vertical Articulation Page Cuts

Assessment	Level 2	Level 3	Level 4
Science 5	12	40	69
Science 6	6	25	64
Biology	20	47	71

Figure 8-2 Post -Vertical Articulation Impact Data



After the standard setting, NCDPI translated these page cuts into the scale scores cuts shown in Table 8-11 . The scale score cuts represent the lower cuts for the adjacent achievement level. For example the Science 5 “Level 2” cut of 242 is interpreted as students with a scale score of 241 or lower are placed in “Achievement Level 1,” and students who score at or between 242 and 251 are considered to be performing at “Achievement Level 2.”

Table 8-11 Scale Scores Cuts Based on Four Achievement Levels

Assessment	Level 2	Level 3	Level 4
Science 5	242	252	263
Science 8	241	248	260
Biology	243	252	261

8.4 Validity of the Standard Setting

At the completion of the standard-setting meeting, an internal evaluation of the overall standard setting process was conducted. This evaluation was facilitated using Kane's (2001) framework, which calls for the evaluation of sources of procedural, internal, and external validity evidence. According to Kane, evidence is needed to support the quality of the design and implementation of the standard setting procedure. Procedural validity was supported by evidence that the steps conducted and procedures followed are supported by national experts and research (e.g., Cizek, 2001; Lewis, Green, Mitzel, Baum, & Patz, 1998; Mitzel, Lewis, Patz, & Green, 2001) and from survey responses by the panelists. This final report summarizes the procedural evidence by detailing the process followed from the description of data collection procedures, implementation of the item-mapping method, final results, and committees' reports (formative and summative) of the process. Formative evaluations, such as readiness surveys, indicated that all standard-setting committee members understood and were adequately prepared to complete the task(s). In addition, as bolstered by the standard setting evaluation survey presented in the results section, standard setting committees generally were confident that the cut scores they recommended aligned well with the achievement level descriptors. A second source of evidence, internal validity evidence, includes evidence of the reliability of the classifications. The standard error of the median cut scores obtained from this sample of panelists was low, with all but two of the indices less than or equal to three pages of the ordered item book, one value of four, and one value of five. As a consequence, even with a different set of raters, the cut scores would likely fall within plus-or-minus three pages of the current recommendations at all grades, subjects, and cut points with the possible exception of two, which may show slightly higher variability. In summary, the validity evidence suggests that the standard setting for the North Carolina EOC and EOG assessments was well-designed and appropriately implemented.

8.5 Standards Adoption and Revision

In October 2013, the North Carolina State Board of Education adopted College-and-Career Readiness Academic Achievement Standards and Academic Achievement descriptors for the EOG and EOC assessments. After considering much input on the importance of having more definitive discrimination for student achievement in the reported levels, the NC*SBE* adopted at its March 2014 meeting a methodology to add a new achievement level. With this additional achievement level, beginning in 2013– 14 student performance on EOG and EOC assessments will be reported based on five achievement levels as described in Table 8-12 and Table 8-13.

Table 8-12 Revised 5 Achievement Levels

Revised Achievement Level	Meets On-Grade-Level Proficiency Standard	Meets College-and-Career Readiness Standard
Level 5 denotes **Superior Command** of knowledge and skills.	Yes	Yes
Level 4 denotes **Solid Command** of knowledge and skills.	Yes	Yes
Level 3 denotes **Sufficient Command** of knowledge and skills.	Yes	No
Level 2 denotes **Partial Command** of knowledge and skills.	No	No
Level 1 denotes **Limited Command** of knowledge and skills.	No	No

Table 8-13 Science Scale Score Cuts Based on Five Achievement Levels 2014 and Beyond

Assessment	Level 2	Level 3	Level 4	Level 5
	Partial	Sufficient	Solid	Superior
5	242	249	252	263
8	241	245	248	260
Biology	243	250	252	261

The level 4 became the new Level 5 “Superior Command,” and students who scored at this level are considered to have met the grade level proficiency standard and are also considered to have met the college-and-career readiness standard. The old Level 3 became the new Level 4 “Solid Command,” and students who scored at this level are considered to have met the grade level proficiency standard and are also considered to have met the college-and-career readiness standard.

The new Achievement Level 3 “Sufficient Command” identifies students who met the grade level proficiency standard but do not meet the college-and-career readiness standard. This distinction assists schools in the delivery of differentiated instruction that best meets the needs of the individual student. The new Level 3 minimum scale score was created by subtracting one conditional standard error of measurement (CSEM) from the original Level 3 scale score. Level 1 “Limited Command” and level 2 “Partial Command” remained unchanged and describe students who have neither met on grade level proficiency standard nor have met college-and-career readiness standards.

Chapter 9 Test Results and Reports

This chapter is divided into two main sections and presents test-level summary statistics for science EOG and EOC assessments based on reported scale scores and achievement levels from 2012–13 and 2014– 15 operational administrations. Section one highlights descriptive summary results of scale scores and reported achievement levels for EOG and EOC forms across major demographic variables. The second section of this chapter presents samples and summary descriptions of the various standardized reports created by the NCDPI which are available to LEA to share assessments results with stakeholders.

9.1 Scale Score Summary

9.1.1 Scale Score Distribution

The scale scores distribution from the first operational administration of the EOG and EOC assessments in 2012– 13 are displayed in the bar charts in Figure 9-1 through Figure 9-3. The descriptive statistics of scale scores are shown in the upper left corner of the chart. Overall, the distributions of scale scores across all grade levels are close to normal with mean ~250 and standard deviation ~10.

Figure 9-1 EOG: Grade 5 Scale Score Distribution 2012–13

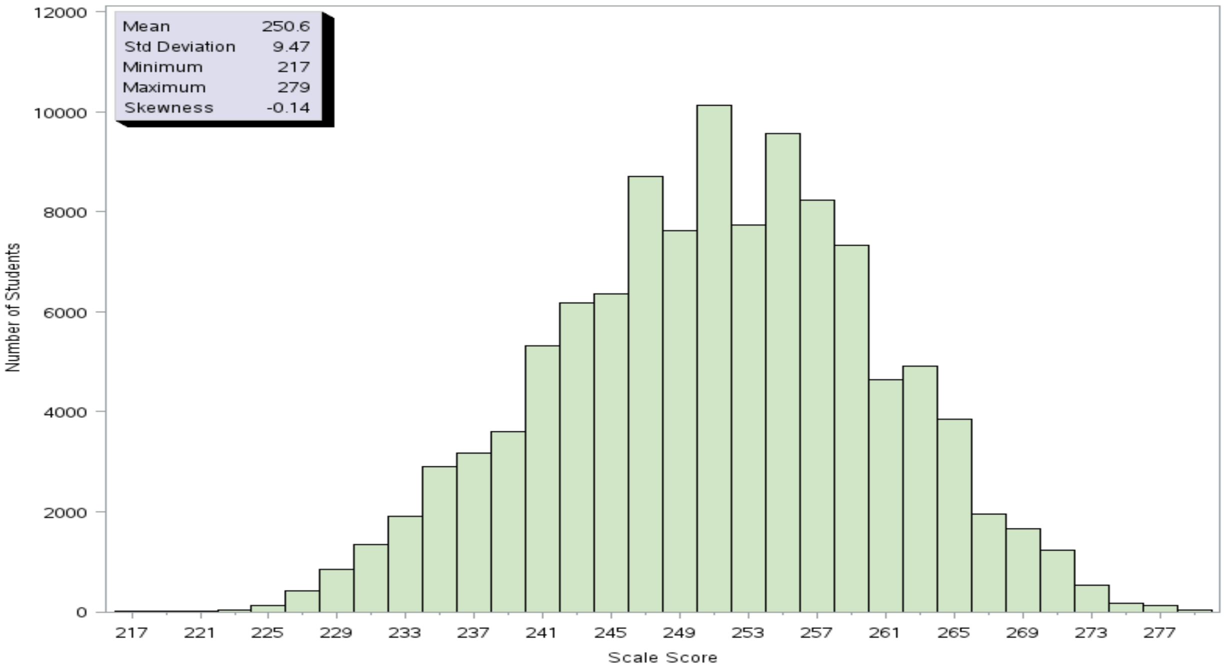


Figure 9-2 EOG: Grade 8 Scale Score Distribution 2012–13

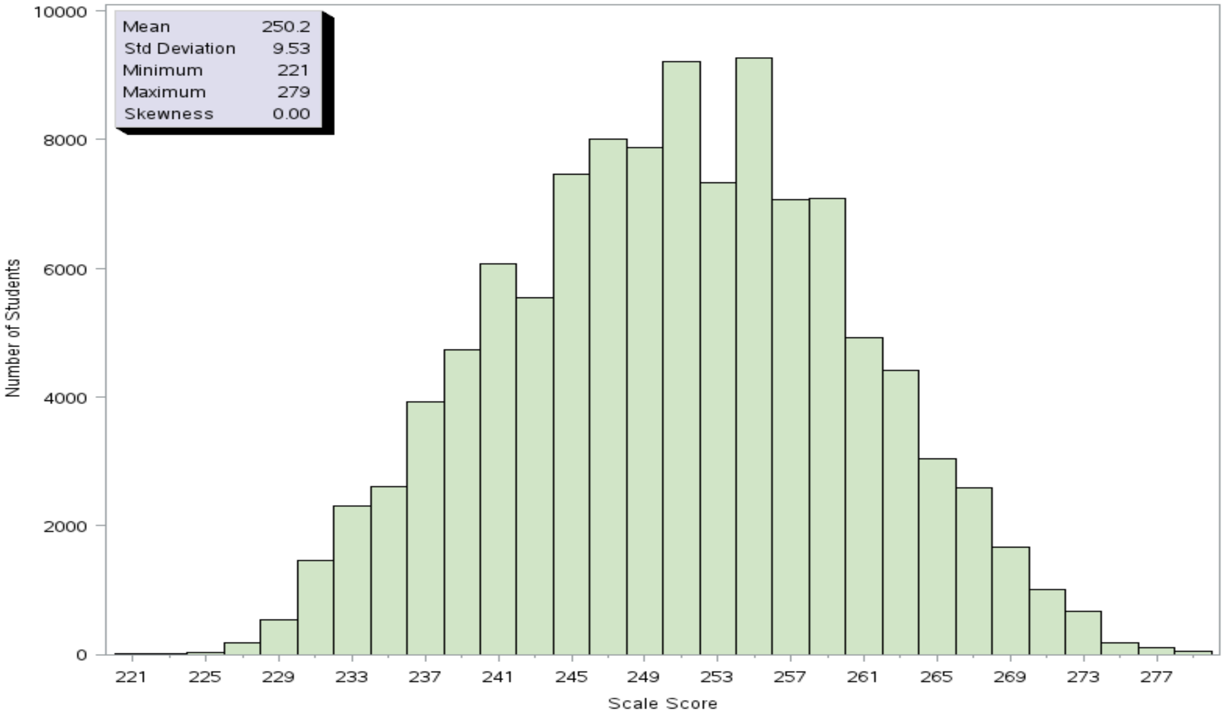


Figure 9-3 EOC: Biology Scale Score Distribution 2012-13

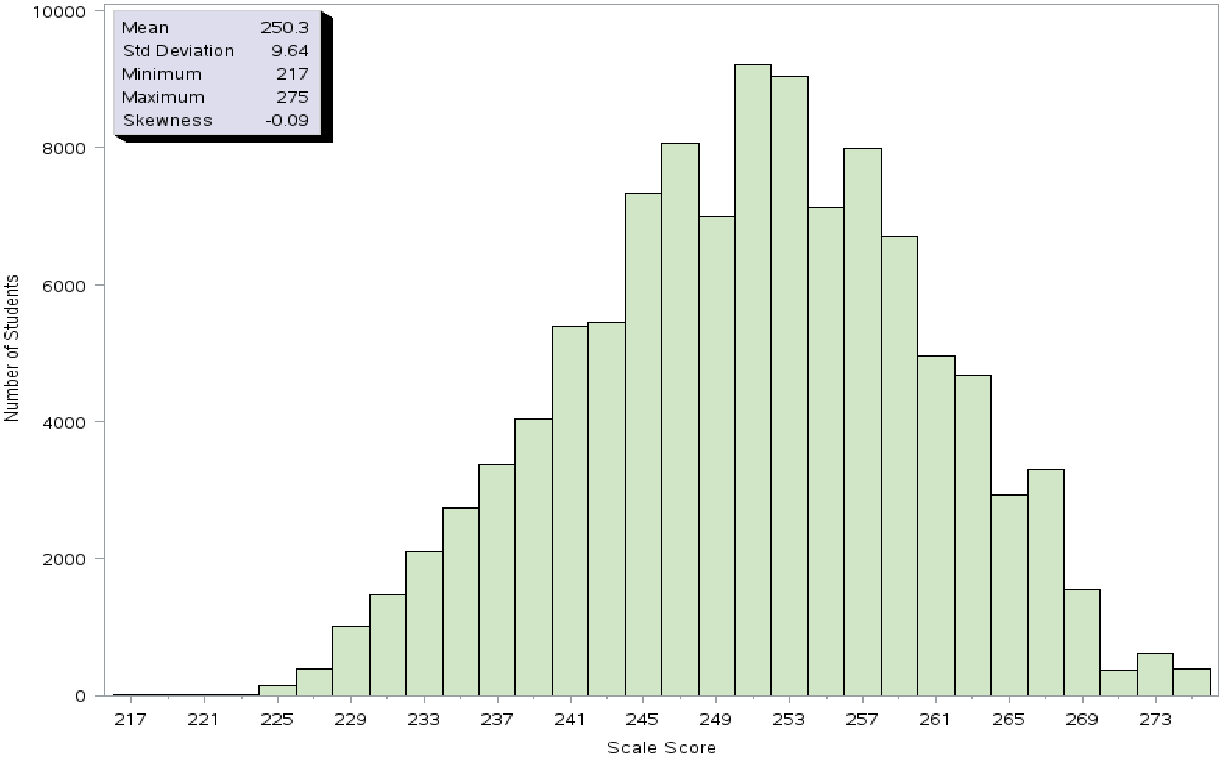


Table *9-1* presents a longitudinal overview of science EOG and EOC scale scores descriptive statistics for the past three administrations (2012–13, 2013–14 and 2014– 15). The number of students taking EOG and EOC assessments across the state has steadily increased across the years, with the exception of grade 5 science from spring 2014 to spring 2015 when the total number of students who took the science EOG assessment decreased from the previous administration. Descriptive summary statistics from Table 9-1 indicate the mean scale scores have been consistent across the past three years. The mean scale scores for science grade 5 for the past three years have increased from 250.6 to 252.0 from 2013 to 2014 and decreased to 251.9 in 2015. For grade 8, scores are trending upward, albeit minimally. For Biology, the scale score trend is nearly flat.

Table 9-1 Descriptive Statistics of Scale Scores by Grade across Administrations, Population

	2013			2014			2015		
Grade	N	Mean	SD	N	Mean	SD	N	Mean	SD
5	110289	250.6	9.4	111659	252.0	9.8	106607	251.9	10.2
8	108981	250.3	9.5	112108	250.8	9.6	116642	250.9	10.0
Biology	104373	250.5	9.6	106639	250.5	9.6	111316	250.1	10.1

The standard deviation (SD) for the EOG and EOC assessments has remained ~10 or has increased slightly across grades from 2012-13 to 2014-15 administrations.

9.1.2 Scale Score Distribution by Gender

Scale score summary by gender for EOG and EOC assessments across three administrations show similar trends observed in the population distribution. Across all grades, the distribution between male and female students is almost even, with male students having a slight majority. In terms of performance, male students on average scored about 0.4 to 1.4 scale score points higher than female students except in Biology in 2015, where the trend is reversed and female students outperformed males by about 0.4 scale score point (see Table 9-2**)**. The SD of scale scores was very similar in both gender groups and followed a similar pattern with a slightly increasing trend across years.

Table 9-2 Scale Scores by Grade and Gender, Population

Grade	Gender	2013			2014			2015		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
5	F	54,632	250.1	9.2	55,088	251.5	9.5	51,937	251.6	9.8
	M	55,657	251.1	9.7	56,571	252.5	10.0	54,670	252.1	10.5
8	F	54,301	249.6	9.1	55,427	250.3	9.2	57,133	250.6	9.5
	M	54,680	251.0	9.8	56,681	251.4	10.0	59,509	251.2	10.4
Biology	F	52,509	250.3	9.3	52,698	250.5	9.4	54,937	250.3	9.8
	M	51,864	250.7	9.8	53,941	250.5	9.8	56,379	249.9	10.4

9.1.3 Achievement Levels

The achievement level classifications for the overall population across grades and administrations are displayed in Table 9-3 and by gender in Table 9-4. Note that the cut scores for the base administration (2012–13) were different from 2013–14 administration and beyond, and as a result students were classified into four achievement levels in 2012– 13 and five achievement levels in 2013– 14 and subsequent administrations. Therefore, the proportion of students in different achievement levels for 2012–13 cannot be directly compared with those from subsequent administrations. For 2013– 14 and beyond, level 3 "Sufficient Command" was added, and levels 3 and 4 became levels 4 and 5 respectively. For 2012– 13, level 3 is missing from Table 9-3 in order to accommodate the display of the proportion of students across years on the same table. The short-term trend of students classified as college-and-career ready (levels 4 and 5) between 2013–14 and 2014–15 on average shows a slight increase in grades 5 and ,8 and about 1% decrease in Biology.

The achievement level classifications by gender across grades and administrations (*Table 9-4*) should be interpreted the same way as the overall population with regards to the achievement levels for 2012– 13. A similar trend as the total population can be observed for each gender group. The results across administrations and grades further indicated that there are higher proportions of male students over female students who scored level 4 or above (college-and-career readiness) for grades 5 and 8. The gaps between the male and female students, however, are closed over administrations. The same trend is true for Biology except for the fact that the proportion of female students exceeded male students in the college-and-career ready category in 2014–15 administration.

Table 9-3 Achievement level classifications by Grade and Year

Grade	Year	N	*% Achievement Level*				
			1) Limited Command, Not CCR	*2) Partial Command, Not CCR*	*3) Sufficient Command, Not CCR*	*4) Solid Command, CCR*	*5) Superior Command, CCR*
5	2012–13*	110,289	17.6	35.3		36.8	10.3
	2013–14	111,659	15.2	18.9	11.9	39.9	14.1
	2014–15	106,607	16.7	18.6	10.5	39.6	14.6
8	2012–13*	108,981	16.6	22.4		43.9	17.1
	2013–14	112,108	15.7	11.1	9.5	44.7	18.9
	2014–15	116,642	16.6	10.7	8.9	43.5	20.4
Biology	2012–13*	104,373	20.8	32.1		32.0	15.0
	2013–14	106,639	20.8	23.8	8.9	31.2	15.3
	2014–15	111,316	23.7	22.1	8.7	30.5	15.0

** There were four achievement levels in 2012-13 hence the results are not comparable with 2013-14 and 2014-15.*
Note: CCR=College-and-Career Ready

Table 9-4 EOG Achievement level classifications by Gender

				Achievement Level				
Grade	Year	Gender	N	*1) Limited Command, Not CCR*	*2) Partial Command, Not CCR*	*3) Sufficient Command, Not CCR*	*4) Solid Command, CCR*	*5) Superior Command, CCR*
5	2012–13*	Female	54,632	18.2	37.4		35.6	8.8
		Male	55,657	17.0	33.2		38.0	11.8
	2013–14	Female	55,088	15.7	20.2	12.4	39.3	12.4
		Male	56,571	14.8	17.7	11.4	40.4	15.7
	2014–15	Female	51,937	16.3	19.7	11.1	39.9	12.9
		Male	54,670	17.1	17.6	9.8	39.3	16.1
8	2012–13*	Female	54,301	17.0	24.6		44.3	14.1
		Male	54,680	16.1	20.3		43.5	20.1
	2013–14	Female	55,427	15.6	12.1	10.5	45.9	16.0
		Male	56,681	15.9	10.2	8.7	43.6	21.7
	2014–15	Female	57,133	15.8	11.3	9.7	45.3	17.9
		Male	59,509	17.3	10.1	8.0	41.8	22.8
Biology	2012–13*	Female	52,509	21.0	33.3		31.8	13.9
		Male	51,864	20.7	31.0		32.2	16.1
	2013–14	Female	52,698	20.4	24.5	9.2	31.1	14.9
		Male	53,941	21.1	23.1	8.6	31.4	15.8
	2014–15	Female	54,937	22.6	22.5	9.1	31.2	14.6
		Male	56,379	24.8	21.6	8.4	29.8	15.4

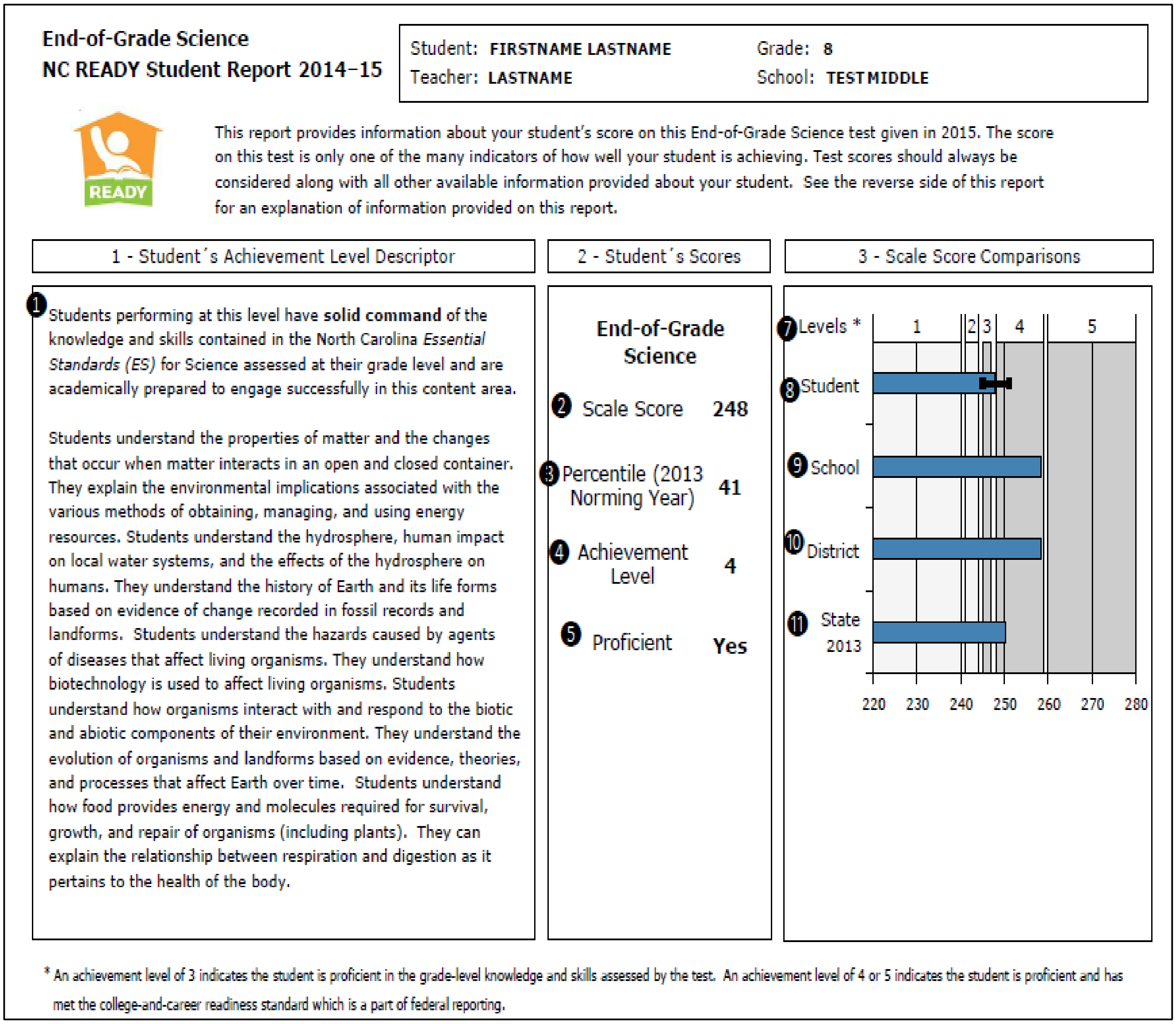
** There were four achievement levels in 2012-13 hence the results are not comparable with 2013-14 and 2014-15.*
Note: CCR=College-and-Career Ready

9.2 Sample Reports

9.2.1 Individual Student Report (ISRs)

A sample ISR report for Grade 8 science is shown in Table 9-4. Key features are labeled and explained in the *Index of Terms by Label Number* section in the ISR. The ISR provides information concerning individual student performance on the science EOG and EOC.

Figure 9-4 Sample Individual Student Report for Grade 5 EOG Science Assessment



End-of-Grade Science
NC READY Student Report 2014–15

Student: **FIRSTNAME LASTNAME**	Grade: **8**
Teacher: **LASTNAME**	School: **TESTMIDDLE**

READY

This report provides information about your student's score on this End-of-Grade Science test given in 2015. The score on this test is only one of the many indicators of how well your student is achieving. Test scores should always be considered along with all other available information provided about your student. See the reverse side of this report for an explanation of information provided on this report.

1 - Student´s Achievement Level Descriptor

1 Students performing at this level have **solid command** of the knowledge and skills contained in the North Carolina *Essential Standards (ES)* for Science assessed at their grade level and are academically prepared to engage successfully in this content area.

Students understand the properties of matter and the changes that occur when matter interacts in an open and closed container. They explain the environmental implications associated with the various methods of obtaining, managing, and using energy resources. Students understand the hydrosphere, human impact on local water systems, and the effects of the hydrosphere on humans. They understand the history of Earth and its life forms based on evidence of change recorded in fossil records and landforms. Students understand the hazards caused by agents of diseases that affect living organisms. They understand how biotechnology is used to affect living organisms. Students understand how organisms interact with and respond to the biotic and abiotic components of their environment. They understand the evolution of organisms and landforms based on evidence, theories, and processes that affect Earth over time. Students understand how food provides energy and molecules required for survival, growth, and repair of organisms (including plants). They can explain the relationship between respiration and digestion as it pertains to the health of the body.

2 - Student´s Scores

End-of-Grade Science

2 Scale Score	**248**
3 Percentile (2013 Norming Year)	**41**
4 Achievement Level	**4**
5 Proficient	**Yes**

3 - Scale Score Comparisons

* An achievement level of 3 indicates the student is proficient in the grade-level knowledge and skills assessed by the test. An achievement level of 4 or 5 indicates the student is proficient and has met the college-and-career readiness standard which is a part of federal reporting.

The “Student Achievement Level Descriptor” section (label 1) describes the level of achievement that the student is expected to have mastered given his or her assessment score. The achievement level descriptors can be viewed at http://www.ncpublicschools.org/accountability/testing//shared/achievelevel.

The “Scale Score” (label 2) presents a scale score that is converted from a raw score. A raw score is the number of assessment questions the student answered correctly. The scale score depicts the growth in achievement from year to year. The Percentile (2013 Norming Year) (label 3) compares a student’s performance on the current year assessment to that of all North Carolina students who took the assessment in the norming year (2012–13). The norming year for an assessment is generally the first year the assessment was administered. The percentile shows a student performed at a level equal to or better than the stated percentage of students who took the assessment during the norming year. For example, the student scores 248 in Grade 8 science, performing as well as or better than 41% of the students who took the assessment in the norming year. The student is at the 41th percentile.

The “Achievement Level” (label 4) shows the level at which a student performed on the assessment. Achievement levels are predetermined performance standards from standard setting that allow a student’s performance to be compared to grade-level expectations. Five achievement levels (i.e., Levels 1, 2, 3, 4, and 5) are reported. Achievement levels of 3, 4, and 5 indicate grade-level Proficiency (label 5). Achievement levels of 4 and 5 indicate college-and-career readiness. The achievement level descriptors can be viewed at http://www.ncpublicschools.org/accountability/testing/shared/achievelevel/.

The “Levels” (label 7) refers to “Achievement Levels,” which allow a student’s performance to be compared to grade-level expectations. The scale score of a “Student” (label 8) is represented by a blue bar. A small black bar on top of the student’s scale score is the confidence interval. The confidence interval indicates the range of scores that would likely result if the same student completed multiple similar tests. For example, if a student takes the same test a second time, the scale score would very likely fall around level 3 or 4. The average “School” score (label 9) and “District” score (label 10) are also represented each by a blue bar. The average scale score for the school and district are based on the fall or spring test administration for the given school year of the report. The average “State” score for 2013 (label 11) is represented by the fourth blue bar. The state average is based on the scores of all North Carolina students who took the test in the norming year (2013).

9.2.2 Class Roster Reports

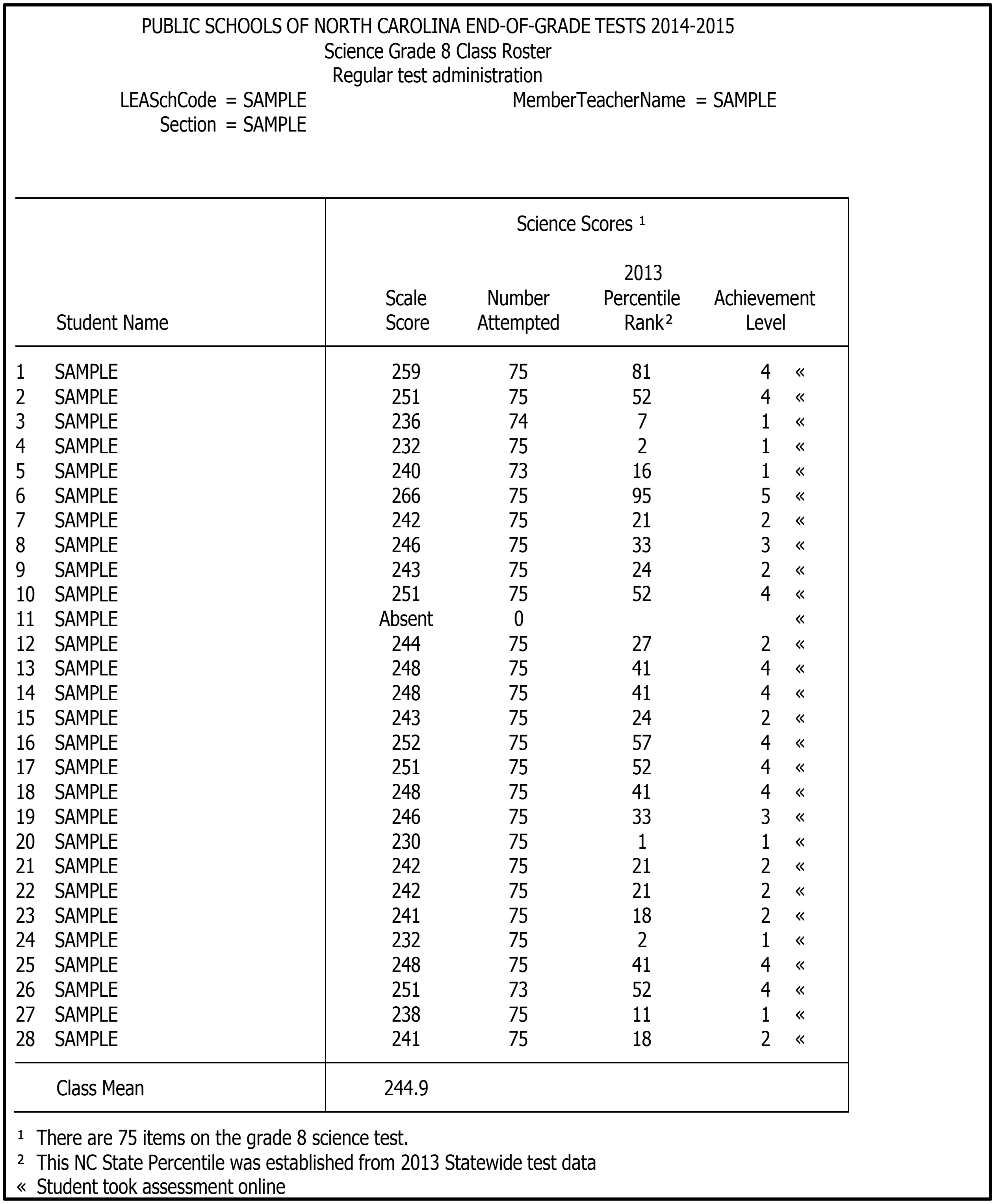
The Class Roster Report takes on many different combinations. A class roster report can contain grade-specific student scores for each content area independently, or it can contain grade-specific student scores for combinations of content areas. **Error! Reference source not found.** displays a sample grade 8 science EOG class roster report. This report is often produced both at the class level and the school level. The report's features and layout do not differ across levels.

In the Class Roster Report, "LEASchCode" refers to the Local Education Agency (LEA) school code, "MemberTeacherName" refers to the instructor's name, and "Section" refers to the class period. The "Scale Score" column presents a score that is converted from a raw score. A raw score is the number of assessment questions the student answers correctly. The scale score depicts growth in achievement from year to year.

The 2013 Percentile Rank column refers to the science percentiles that were established from 2013 statewide assessment data. The "Achievement Level" column shows the level at which a student performed on the assessment. Achievement levels are predetermined performance standards that allow a student's performance to be compared to grade-level expectations. Five achievement levels (i.e., Levels 1, 2, 3, 4, and 5) are reported. Achievement levels of 3, 4, and 5 indicate grade-level proficiency. Achievement levels of 4 and 5 indicate college-and-career readiness.

The achievement level descriptors can be viewed at http://www.ncpublicschools.org/accountability/testing/shared/achievelevel/. The "Class Mean" is the average of the class scores. It is the sum of all scores in the class divided by the number of students in the class. For example, the class in the report received an averaged scale score of 244.9 in grade 8 science.

Figure 9-5 Sample Class Roster Report for EOG Grade 5 (diff. font in table)



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2014-2015
Science Grade 8 Class Roster
Regular test administration

LEASchCode = SAMPLE   MemberTeacherName = SAMPLE
Section = SAMPLE

	Student Name	Science Scores [1] Scale Score	Number Attempted	2013 Percentile Rank [2]	Achievement Level	
1	SAMPLE	259	75	81	4	«
2	SAMPLE	251	75	52	4	«
3	SAMPLE	236	74	7	1	«
4	SAMPLE	232	75	2	1	«
5	SAMPLE	240	73	16	1	«
6	SAMPLE	266	75	95	5	«
7	SAMPLE	242	75	21	2	«
8	SAMPLE	246	75	33	3	«
9	SAMPLE	243	75	24	2	«
10	SAMPLE	251	75	52	4	«
11	SAMPLE	Absent	0			«
12	SAMPLE	244	75	27	2	«
13	SAMPLE	248	75	41	4	«
14	SAMPLE	248	75	41	4	«
15	SAMPLE	243	75	24	2	«
16	SAMPLE	252	75	57	4	«
17	SAMPLE	251	75	52	4	«
18	SAMPLE	248	75	41	4	«
19	SAMPLE	246	75	33	3	«
20	SAMPLE	230	75	1	1	«
21	SAMPLE	242	75	21	2	«
22	SAMPLE	242	75	21	2	«
23	SAMPLE	241	75	18	2	«
24	SAMPLE	232	75	2	1	«
25	SAMPLE	248	75	41	4	«
26	SAMPLE	251	73	52	4	«
27	SAMPLE	238	75	11	1	«
28	SAMPLE	241	75	18	2	«
	Class Mean	244.9				

[1] There are 75 items on the grade 8 science test.

[2] This NC State Percentile was established from 2013 Statewide test data

« Student took assessment online

9.2.3 Scale Score Frequency Reports

The Scale Score Frequency Reports available in WinScan are used to summarize scale score information at the class, school, district, and state levels. The reports present the frequency, percent, cumulative frequency, and cumulative percent of each scale score at a specific grade. These reports can be created for each EOG and EOC assessment. Figure 9-6 presents a sample Score Frequency Report for the grade 8 EOG science assessment.

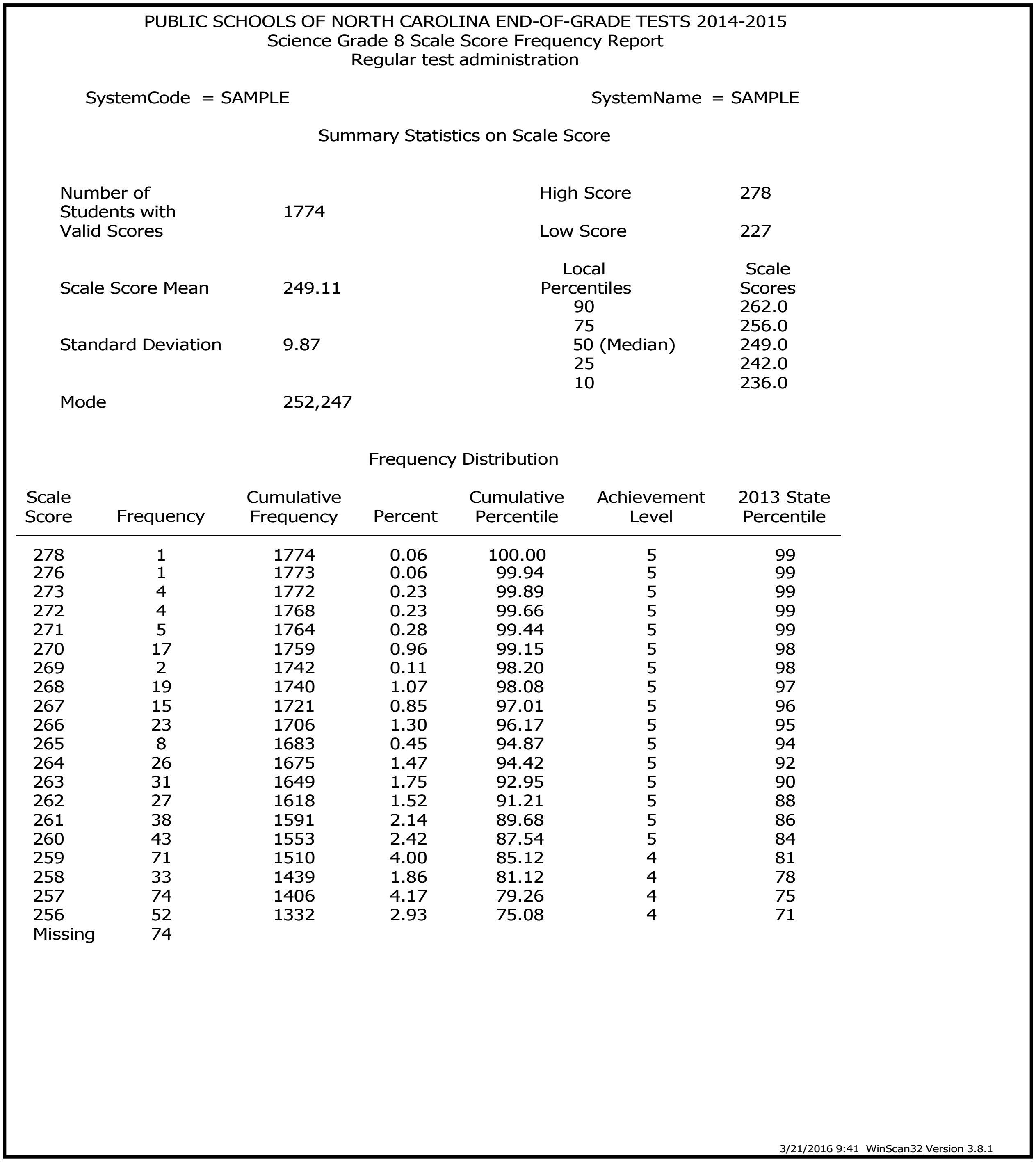
The Score Frequency Report consists of three sections: the header, a summary statistics on scale score, and frequency distribution. The first line of the sample Score Frequency Report header describes the type of assessment (EOG or EOC) and the school year (2014-15). The second line displays the specific type of assessment, the grade, the subject area, and the type of report. The SystemCode indicates the LEA school code and the SystemName indicates the LEA's name.

The summary statistics on scale score table indicates that 1774 students in this report had valid scores. The highest score was 278 and the lowest score was 227. The arithmetic mean of the scale score was 249.11, the standard deviation was 9.87, and there were multiple modes (252 and 247). The percentile scores are listed at the far right of the table. The scale scores are listed for the 10th, 25th, 50th, 75th, and 90th percentiles. In this sample, a scale score of 256 corresponds to a 75th percentile. This means that 75% of the 1774 students earned a score of 256 or less.

In the Frequency Distribution table, the Scale Score column presents every score earned by the 1774 students. The Frequency column on the report presents the number of students that earned each scale score. For example, one students earned a scale score of 278. A "Missing" label would indicate that 74 students did not receive a score. The Cumulative Frequency column shows that 1332 students earned up to and including a scale score of 256.

The Percent column presents the percent of students that earned a given scale score (number of students that earned the score divided by total number of observations) indicating 2.93% of the students earned a score of 256. The Cumulative Percent column shows 75.08% of the students earned up to and including a scale score of 256. The Achievement Level column displays the achievement level associated with each scale score. In this example, a scale score of 256 corresponds to an achievement level of 4.

Figure 9-6 Sample Score Frequency Report for EOG Grade 8 Science



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2014-2015
Science Grade 8 Scale Score Frequency Report
Regular test administration

SystemCode = SAMPLE

SystemName = SAMPLE

Summary Statistics on Scale Score

Number of Students with Valid Scores	1774	High Score	278
		Low Score	227
		Local Percentiles	Scale Scores
Scale Score Mean	249.11	90	262.0
		75	256.0
Standard Deviation	9.87	50 (Median)	249.0
		25	242.0
		10	236.0
Mode	252,247		

Frequency Distribution

Scale Score	Frequency	Cumulative Frequency	Percent	Cumulative Percentile	Achievement Level	2013 State Percentile
278	1	1774	0.06	100.00	5	99
276	1	1773	0.06	99.94	5	99
273	4	1772	0.23	99.89	5	99
272	4	1768	0.23	99.66	5	99
271	5	1764	0.28	99.44	5	99
270	17	1759	0.96	99.15	5	98
269	2	1742	0.11	98.20	5	98
268	19	1740	1.07	98.08	5	97
267	15	1721	0.85	97.01	5	96
266	23	1706	1.30	96.17	5	95
265	8	1683	0.45	94.87	5	94
264	26	1675	1.47	94.42	5	92
263	31	1649	1.75	92.95	5	90
262	27	1618	1.52	91.21	5	88
261	38	1591	2.14	89.68	5	86
260	43	1553	2.42	87.54	5	84
259	71	1510	4.00	85.12	4	81
258	33	1439	1.86	81.12	4	78
257	74	1406	4.17	79.26	4	75
256	52	1332	2.93	75.08	4	71
Missing	74					

3/21/2016 9:41 WinScan32 Version 3.8.1

The 2013 State Percentile column displays to the science percentiles that were established from 2013 statewide assessment data. This column shows that a scale score of 256 was in the 71st percentile in 2013.

9.2.4 Achievement Level Frequency Reports

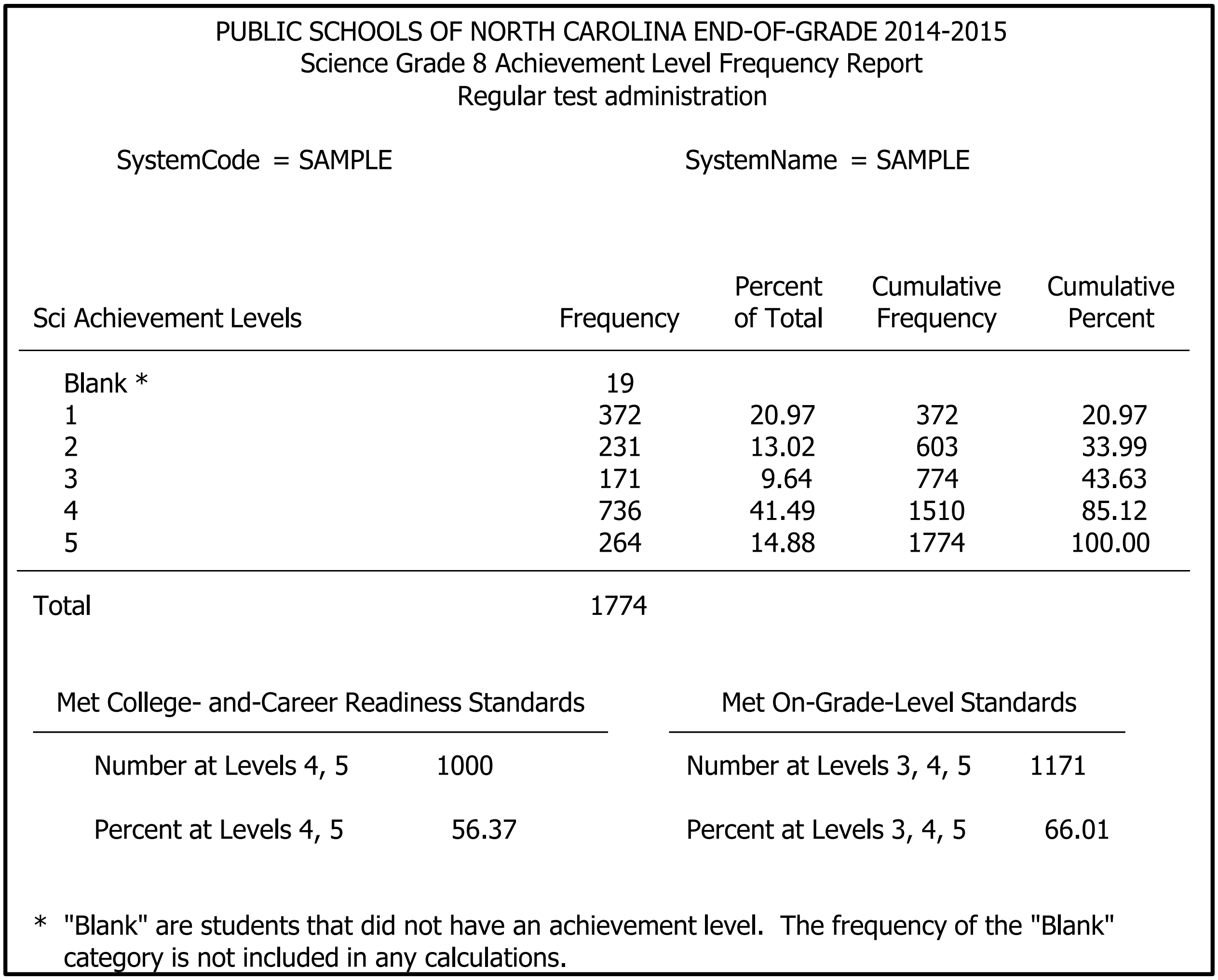
Figure 9-7 displays a sample Achievement Level Frequency Report for grade 8 EOG science assessment. The first line of the header indicates the report is for the 2014-2015 school year. The second line indicates the subject area, grade, and report type. In this sample report, the exam was a regular administration. SystemCode indicates the LEA school code and SystemName indicates the LEA School name. The values under "Sci Achievement Levels" indicate five proficiency levels (1=Limited Command, 2=Partial Command, 3=Sufficient Command, 4=Solid Command, and 5=Superior Command). The corresponding frequencies indicating what proportion of the total students were classified into different proficiency levels. Students who do not have an achievement level are classified as "blank". In the report there were 19 students who did not get a valid score.

The Frequency column presents the number of students that earned each achievement level. The total count of students excludes blank scores. The sample shows 736 students earned an achievement level of 4 in science that corresponds to 41.49% of the total students (number of students that earned the achievement level divided by total number of students with valid scores).

The Cumulative Frequency column presents the total number of students who earned up to and including an achievement level in a given row. This column shows 1510 students earned up to and including an achievement level of 4 in science. The Cumulative Percent column displays the percent of students that earned up to and including an achievement level in a given row. In the sample shown, 85.12% of the students earned up to and including an achievement level of 4 in science.

The report also provides number and percent of students who were college-and-career ready (Levels 4 and 5) and met on-grade-level standards (Levels 3, 4, and 5). The summary statistics just below the frequency table show 1000 of 1774 students were classified as level 4 or 5 and 1171 of the 1774 were classified as level 3, 4 or 5 in science. This corresponds to 66.01% of the students at grade-level proficient (levels 3 and above) and 56.37% at college-and-career ready (levels 4 and above) in grade 8 science.

Figure 9-7 Sample Achievement Level Frequency Report for EOG Grade 8 Science



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE 2014-2015
Science Grade 8 Achievement Level Frequency Report
Regular test administration

SystemCode = SAMPLE SystemName = SAMPLE

Sci Achievement Levels	Frequency	Percent of Total	Cumulative Frequency	Cumulative Percent
Blank *	19			
1	372	20.97	372	20.97
2	231	13.02	603	33.99
3	171	9.64	774	43.63
4	736	41.49	1510	85.12
5	264	14.88	1774	100.00
Total	1774			

Met College- and-Career Readiness Standards		Met On-Grade-Level Standards	
Number at Levels 4, 5	1000	Number at Levels 3, 4, 5	1171
Percent at Levels 4, 5	56.37	Percent at Levels 3, 4, 5	66.01

* "Blank" are students that did not have an achievement level. The frequency of the "Blank" category is not included in any calculations.

9.2.5 Goal Summary Reports

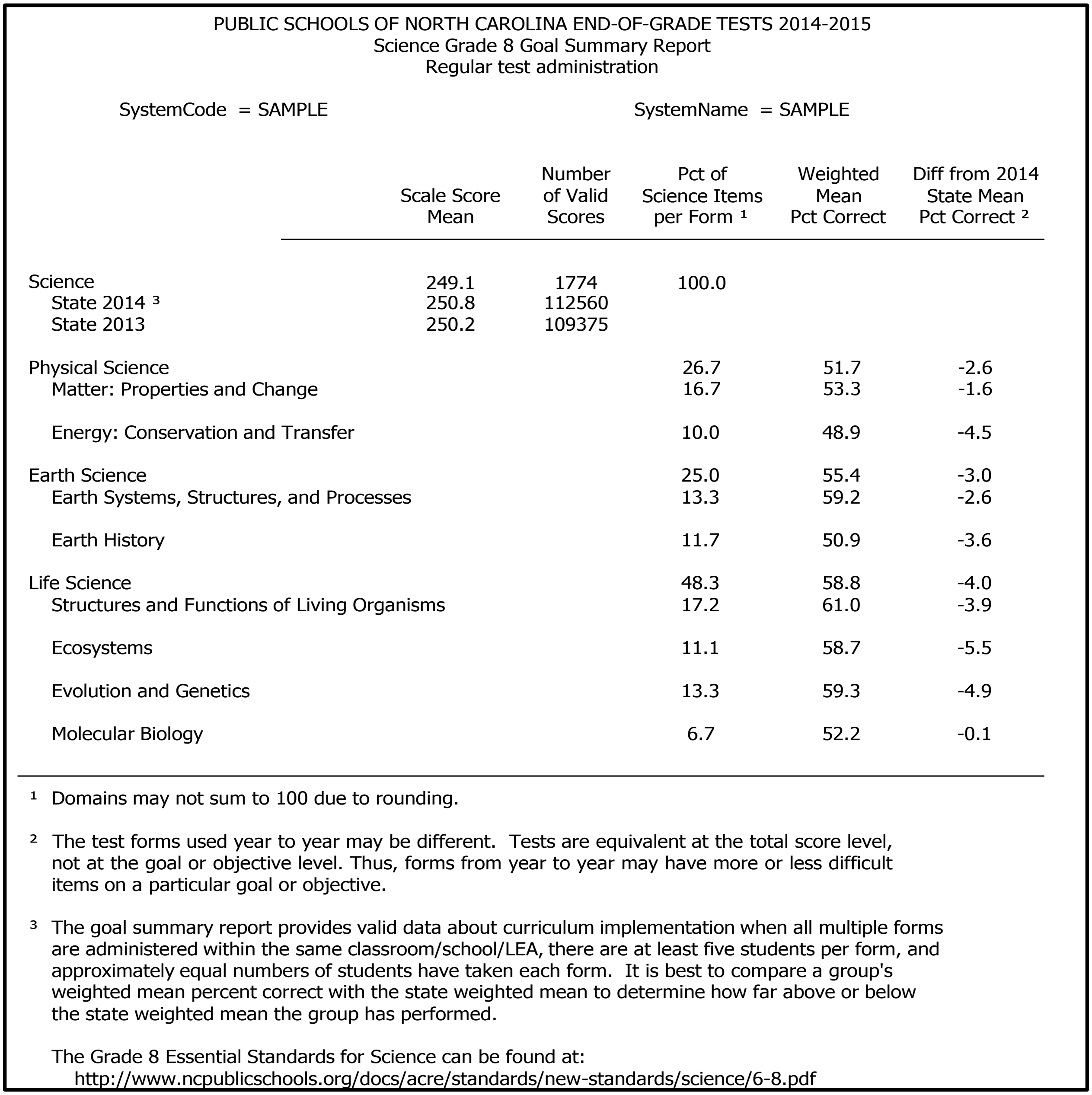
The Goal Summary Report is a grade-specific report that summarizes student performance for each learning goal or essential standard. The Goal Summary Report can group students at the school, district, or state level. Typically, the Goal Summary Report reflects scores at the goal or domain level, for example, Physical Science, Earth Science, and Life Science. Figure 9-8 displays a sample goal summary report. The standard protocol for reporting subscale scores requires that any domain with fewer than five items does not produce a level of reliability sufficient for score reporting. The goal summary report provides valid data about curriculum implementation only when 1) all forms are administered within the same classroom, school, or LEA; 2) there are at least five students per form; and 3) approximately equal numbers of students have taken each form. It is best to compare a group's weighted mean percent correct with the state's weighted mean to determine how far above or below the state weighted mean the group has performed.

In this sample report, SystemCode indicates the LEA school code and SystemName refers to LEA name. The Scale Score Mean column presents the average scale scores at the state level during the administration. As the name suggested, Number of Valid Scores column presents the number of students whose scores were validly reported. For example, EOG Grade 8 science administrated in 2013 has 109375 valid scores in North Carolina with a mean at 250.2.

The "Pct of Science Items per Form" column presents the percent of the items per form that align with each content domain. In the grade 8 science in 2014-15 administration, 26.7% items in each form came from Physical Science content domain. The "Weighted Mean Pct Correct" column provides averaged scores for each content domain from different forms. If the count of students differs across forms, a weighted mean adjusts for the different counts across the forms. For instance, if twice as many students took one form as compared to another, this form would receive twice the weight in calculating the mean for the content area. Usually about the same numbers of students take each form, so in practice, the weighted mean is very similar to an unweighted mean.

The "Diff from 2013 State Mean Pct Correct" column displays performance relative to the 2013 state mean percent correct. Negative values indicate a score performance below the state mean percent correct, while positive values indicate performance above the state mean.

Figure 9-8 Sample Goal Summary Report for EOG Grade 8 Science



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2014-2015
Science Grade 8 Goal Summary Report
Regular test administration

SystemCode = SAMPLE SystemName = SAMPLE

	Scale Score Mean	Number of Valid Scores	Pct of Science Items per Form [1]	Weighted Mean Pct Correct	Diff from 2014 State Mean Pct Correct [2]
Science	249.1	1774	100.0		
State 2014 [3]	250.8	112560			
State 2013	250.2	109375			
Physical Science			26.7	51.7	-2.6
Matter: Properties and Change			16.7	53.3	-1.6
Energy: Conservation and Transfer			10.0	48.9	-4.5
Earth Science			25.0	55.4	-3.0
Earth Systems, Structures, and Processes			13.3	59.2	-2.6
Earth History			11.7	50.9	-3.6
Life Science			48.3	58.8	-4.0
Structures and Functions of Living Organisms			17.2	61.0	-3.9
Ecosystems			11.1	58.7	-5.5
Evolution and Genetics			13.3	59.3	-4.9
Molecular Biology			6.7	52.2	-0.1

[1] Domains may not sum to 100 due to rounding.

[2] The test forms used year to year may be different. Tests are equivalent at the total score level, not at the goal or objective level. Thus, forms from year to year may have more or less difficult items on a particular goal or objective.

[3] The goal summary report provides valid data about curriculum implementation when all multiple forms are administered within the same classroom/school/LEA, there are at least five students per form, and approximately equal numbers of students have taken each form. It is best to compare a group's weighted mean percent correct with the state weighted mean to determine how far above or below the state weighted mean the group has performed.

The Grade 8 Essential Standards for Science can be found at:
http://www.ncpublicschools.org/docs/acre/standards/new-standards/science/6-8.pdf

For example, student's average score for the content domain "Physical Science" is -2.6 score points lower than that in 2013. However, test forms used this year may be different from forms in 2013. Tests are equivalent at the total score level, not at the objective level. Thus, difficulty at domain or objective level may be different in this year's forms and those in 2013. *Interpretative Guides to the Winscan Score Reports for North Carolina Assessments* are available at http://www.ncpublicschools.org/accountability/testing/technicalnotes.

Chapter 10 Validity Evidences and Reports 2012–2015

This chapter presents summary validity evidence collected in support of the interpretation of EOG and EOC test scores. The first couple of sections in this chapter present validity evidence in support of the internal structure of these assessments. Evidence presented in these sections includes reliability, standard error estimates, and a classification consistency summary of reported achievement levels, and exploratory Principal Component analysis in support of the unidimensionality interpretation of EOG and EOC test scores. The final sections of the chapter documents validity-evidence based content summarized from the alignment study and evidence based on relation to other variables, and the last part presents a summary of procedures used to ensure EOG and EOC assessments are accessible and fair for all students.

10.1 Reliability Evidences of EOG and EOC Science

The internal consistency reliability estimate provides a sample base summary statistic that describes the proportion of the reported score that is true score variance. In order to justify valid use of test results in large-scale standardized assessments, evidence must be documented that shows test results are stable, consistent, and dependable across all subgroups of the intended population. A reliable test produces scores that are expected to be relatively stable if the test were to be administered repeatedly under similar conditions. Scores from a reliable test reflect examinees' expected ability in the construct being measured with very little error variance. Internal consistency reliability coefficients (in this case measured by Cronbach's alpha) range from zero to one, where a coefficient of one refers to perfectly reliable measures with no error. For high-stakes assessments, alpha estimates of 0.85 or higher are generally desirable. The Cronbach's alpha (Cronbach, 1951) can be calculated as:

$$\boldsymbol{\hat{\alpha} = \frac{\kappa}{\kappa-1}\left(1 - \frac{\Sigma\hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right)} \qquad \textit{(10-1)}$$

Where k is the number of items on the test form, $\hat{\sigma}_i^2$ is the variance of item i, and $\hat{\sigma}_X^2$ is the total test variance. It is worth noting here that reliability estimates, since they are sample based, are less informative in describing accuracy of individual students' scores.

Table 10.1 shows Cronbach alpha as a measure of reliability estimate for all EOG and EOC science forms by grade, form, major demographic variables, and special student group. Overall, across all forms, reliability estimates from the 2012–2013 population range from 0.90 to 0.92. Subgroup reliabilities for gender ranged from 0.89 to 0.93; for ethnicity they ranged from 0.86 to 0.91; and for the special student group they ranged from 0.79 to 0.92. These alphas suggest that the reliability of the NCSTP tests are reasonably high.

Table 10-1 Science EOG and Biology EOC Reliabilities by Subgroup

EOG/EOC and Form		Gender		Ethnicity[h]			Accommodation		All Students
		Female	Male	Black	Hispanic	White	SWD	ELL	
Science Grade 5	*A*	0.90	0.91	0.88	0.89	0.89	0.89	0.86	0.90
	B	0.89	0.90	0.88	0.88	0.88	0.89	0.86	0.90
	C	0.90	0.91	0.88	0.89	0.88	0.89	0.84	0.90
	M	0.90	0.91	0.88	0.89	0.89	0.88	0.84	0.91
	N	0.89	0.91	0.88	0.89	0.89	0.89	0.79	0.90
	O	0.89	0.90	0.87	0.89	0.89	0.89	0.86	0.90
Science Grade 8	*A*	0.91	0.93	0.89	0.91	0.91	0.91	0.85	0.91
	B	0.91	0.93	0.89	0.90	0.91	0.90	0.86	0.91
	C	0.91	0.92	0.89	0.89	0.91	0.90	0.85	0.91
	M	0.91	0.92	0.89	0.90	0.91	0.88	0.85	0.92
	N	0.90	0.92	0.88	0.90	0.90	0.88	0.85	0.92
	O	0.90	0.92	0.88	0.90	0.90	0.88	0.84	0.92
Biology	*A*	0.90	0.91	0.88	0.90	0.90	0.88	0.86	0.92
	B	0.90	0.91	0.87	0.89	0.89	0.87	0.81	0.92
	C	0.90	0.91	0.88	0.90	0.89	0.88	0.84	0.92
	M	0.91	0.91	0.88	0.90	0.90	0.87	0.84	0.92
	N	0.89	0.90	0.86	0.88	0.89	0.85	0.83	0.92
	O	0.90	0.91	0.88	0.90	0.89	0.88	0.86	0.92

[h] Reliabilities estimates are displayed only for major ethnic groups and special student groups investigated in DIF analysis with acceptable sample size.

10.2 Conditional Standard Error at Scale Score Cuts

The information provided by the standard error of measurement (SEM) for a given score is important because it assists in determining the accuracy of examinees' classifications. It allows a probabilistic statement to be made about an individual's test score. For example, if a score of 100 has an SEM of plus or minus two, then one can conclude that a student obtained a score of 100, which is accurate within plus or minus 2 points with 68% confidence. In other words, a 68% confidence interval for a score of 100 is 98–102. If that student were to be retested, his or her score would be expected to be in the range of 98–102 about 68% of the time.

The standard errors of measurement at the scale score cuts for achievement levels for the North Carolina EOG and EOC science assessments are provided in Table 10-2. For students with scores within 2 SD of the mean (95% of the students), standard errors are typically 2 to 3 points. For most of the EOG and EOC science scale scores, the standard error of measurement in the middle range of scores, particularly at the cut point between Level II and Level III, is generally around 3 points. Scores at the lower and higher ends of the scale (above the 97.5th percentile and below the 2.5th percentile) have standard errors of measurement of approximately 5 points. This is typical as the extreme scores are associated with fewer students, with less variability resulting in less measurement precision associated with those extreme scores.

The SEMs at Level 2 across forms and grades ranged from 3 to 4, and Level 3 and Level 4 ranged from 2 to 3. One useful application of the conditional SEMs is that it can be used to estimate a band of scores around any scale score or cut score where decision has to be precise. For example, the solid proficiency (Level 3) cut score for grade 5 science is 249. If a student obtained a scale score of 249, the SEM of 3 in Form A tells educators with 68% probability that the student's scale score could range from 246 to 252 (249±1*3), meaning that the student will likely not be Solid Proficient. Similarly, if an educator wants to be 95% confident of the decision, the scale score could range from 243 to 255 (249±2*3).

Table 10-2 Conditional Standard Errors at Achievement level Cuts and Hoss/Loss by Form and Grade Level

		LOSS		Level 2		Level 3		Level 4		Level 5		HOSS	
Grade	Form	LOSS	SE	Partial	SE	Sufficient	SE	Solid	SE	Superior	SE	HOSS	SE
5	A	217	5	242	3	249	3	252	3	263	3	278	5
	B	220	5	242	3	249	3	252	3	263	3	279	5
	C	217	5	242	3	249	3	252	3	263	4	277	5
	M	217	5	242	3	249	3	252	3	263	3	278	5
	N	220	5	242	3	249	3	252	3	263	3	279	5
	O	217	5	242	3	249	3	252	3	263	4	277	5
8	A	221	5	241	4	245	3	248	3	260	3	278	5
	B	221	5	241	4	245	3	248	3	260	3	278	5
	C	221	5	241	3	245	3	248	3	260	3	278	5
	M	221	5	241	3	245	3	248	3	260	3	278	5
	N	221	5	241	4	245	3	248	3	260	3	278	5
	O	221	5	241	4	245	3	248	3	260	3	279	5
Biology	A	218	5	243	3	250	3	252	2	261	3	275	5
	B	219	5	243	3	250	2	252	2	261	3	275	5
	C	217	5	243	3	250	2	252	2	261	3	274	5
	M	218	5	243	3	250	2	252	2	261	3	275	5
	N	219	5	243	3	250	2	252	2	261	3	275	5
	O	217	5	243	3	250	2	252	2	261	3	274	5

Note: LOSS = the lowest obtainable scale score; HOSS = the highest obtainable scale score; Partial=partial command; Sufficient=sufficient command; Solid=solid command; Superior=superior command

10.3 Evidence of Classification Consistency

The No Child Left Behind Act of 2001 and subsequent Race to the Top Initiative emphasized the measurement of adequate yearly progress (AYP) with respect to the percentage of students at or above performance standards set by states. With this emphasis on the achievement level classification, a psychometric interest could be how consistently and accurately assessment instruments can classify students into the achievement levels. The importance of classification consistency as a measure of the categorical decisions when the test is used repeatedly has been recognized in the Standard 2.16 of the *Standards* (AERA, APA, and NCME, 20144) which states that

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

The methodology used for estimating the reliability of achievement-level classification decisions, as described in Hanson and Brennan (1990) and Livingston and Lewis (1995), provides estimates of decision accuracy and classification consistency. The classification consistency refers to "the agreement between classifications based on two non-overlapping, equally difficult forms of the test," and decision accuracy refers to "the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known" (Livingston & Lewis, 1995, P. 178). That is, classification consistency refers to the agreement between two observed scores, while classification accuracy refers to the agreement between observed and true scores.

The analyses are implemented using the computer program BB-Class.[i] The program provides results for both the Hanson and Brennan (1990) and Livingston and Lewis (1995) procedures. Since the Hanson and Brennan (1990) procedures assume that a "test consists of *n* equally weighted, dichotomously-scored items," while the Livingston and Lewis (1995) procedures are intended to handle situations in which "(a) items are not equally weighted and/or (b) some or all of the items are polytomously scored" (Brennan, 2004, pp. 2–3), the analyses for EOG and EOC science followed the Hanson and Brennan (1990) or HB procedures.

Table 10-3 presents the decision accuracy and consistency indices for achievement levels at each grade. Overall, the values indicate good classification accuracy (ranging from 0.91 to 0.94) and consistency (from 0.87 to 0.92). For example, if Grade 5 Science students who were classified as Level 2 take a non-overlapping, equally difficult form a second time, 92% of them

[i] BB-Class is an ANSI C computer program that uses the beta-binomial model (and its extensions) for estimating classification consistency and accuracy. It can be downloaded from https://www.education.uiowa.edu/centers/casma/computer-programs#de748e48-f88c-6551-b2b8-ff00000648cd.

would still be classified in Level 2. Smaller standard error translates to a highly reliable measurement that will exhibit higher levels of classification consistency.

Table 10-3 Classification Accuracy and Consistency Results

	Level 2 Partial Command		Level 3 Sufficient Command		Level 4 Solid Command		Level 5 Superior Command	
Grade	Acc.	Con.	Acc.	Con.	Acc.	Con.	Acc.	Con.
5	0.94	0.92	0.91	0.88	0.91	0.87	0.92	0.90
8	0.94	0.91	0.92	0.89	0.91	0.88	0.93	0.91
Biology	0.93	0.90	0.91	0.88	0.91	0.88	0.94	0.89

Note: Acc. = Classification Accuracy; Con. = Classification Consistency

10.4 EOG and EOC Dimensionality Analysis

Evidence of overall dimensionality for EOG and EOC science assessments were explored using Principal Component Analysis (PCA). PCA is an exploratory technique that seeks to summarize observed variables using fewer linear dimensions referred to as components. The primary question in a PCA analysis is to determine the fewest number of reasonable dimensions or components that can explain most of the observed variance in the data. Two common criteria are used to decide the number of meaningful dimensions for a set of observed variables:

- Retain components whose eigenvalues are greater than the average of all the eigenvalues which is usually 1.
- Use scree a graph which is a plot of eigenvalues against and count the number of components above the natural linear break.

It is very common to rely on both criteria when evaluating the number of possible dimensions for a given variable.

To explore the dimensionality of NC EOG and EOC assessments, PCA were extracted from the tetrachoric correlation matrix for dichotomized response data to determine the number of meaningful components. Scree graphs from the PCA analysis by grade are shown in Figure 10-1 through Figure 10-3 for the first 16 components. The eigenvalue of the first component describes the amount of total variance accounted for by that component range from 15–20 and accounted for about 30% of total variance. The ratio of the first to second eigenvalue across grades ranged from approximately 6 to greater than 8 for some grades and forms. Based on the two evaluation criteria listed above, a strong case can be made for one dominant component to explain a significant amount of the total variance in the observed correlation matrices for EOG and EOC science forms. Further evaluation of the scree graph with the distinct break of the linear trend after the first dominant component presents enough exploratory evidence in support of the assumption of unidimensionality of EOG and EOC assessments.

Figure 10-1 Grade 5 Science Scree Plot of Operational Forms

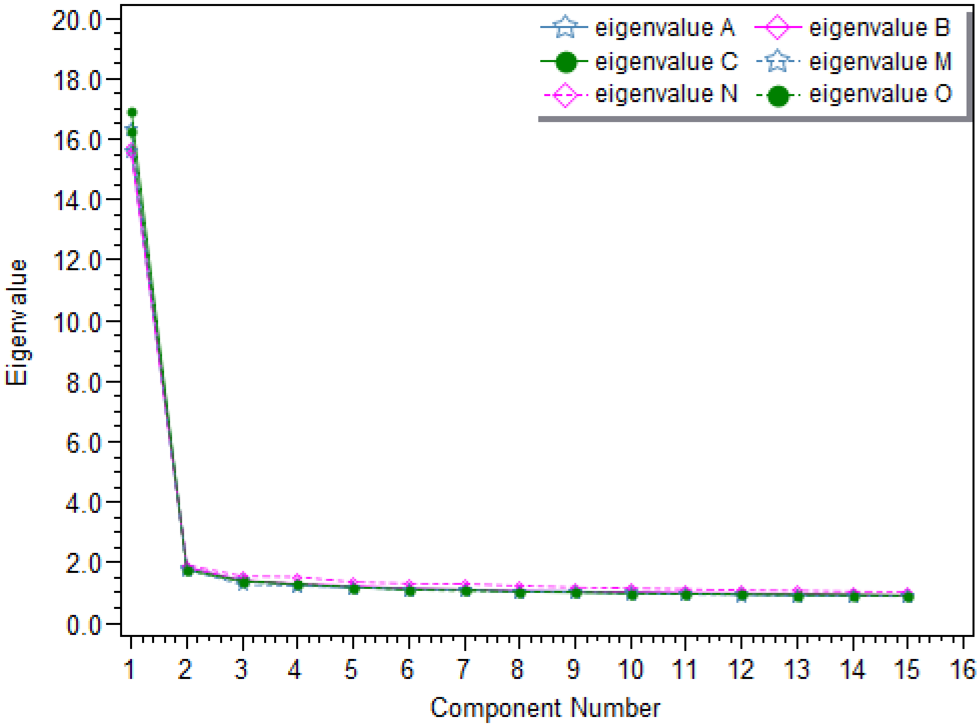


Figure 10-2 Grade 8 Science Scree Plot of Operational Forms

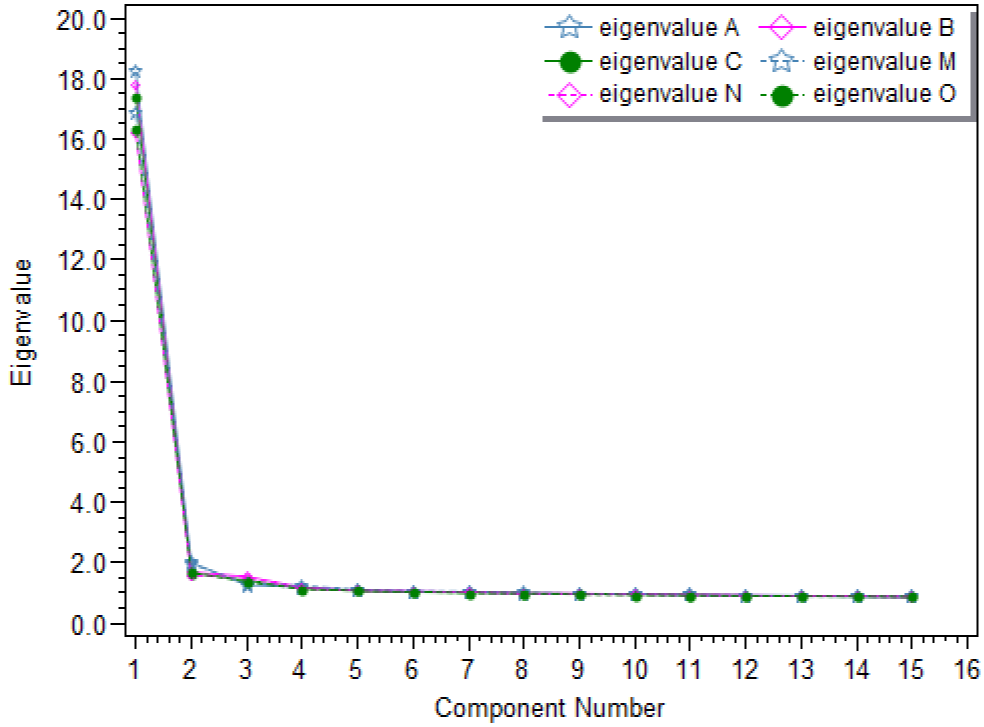
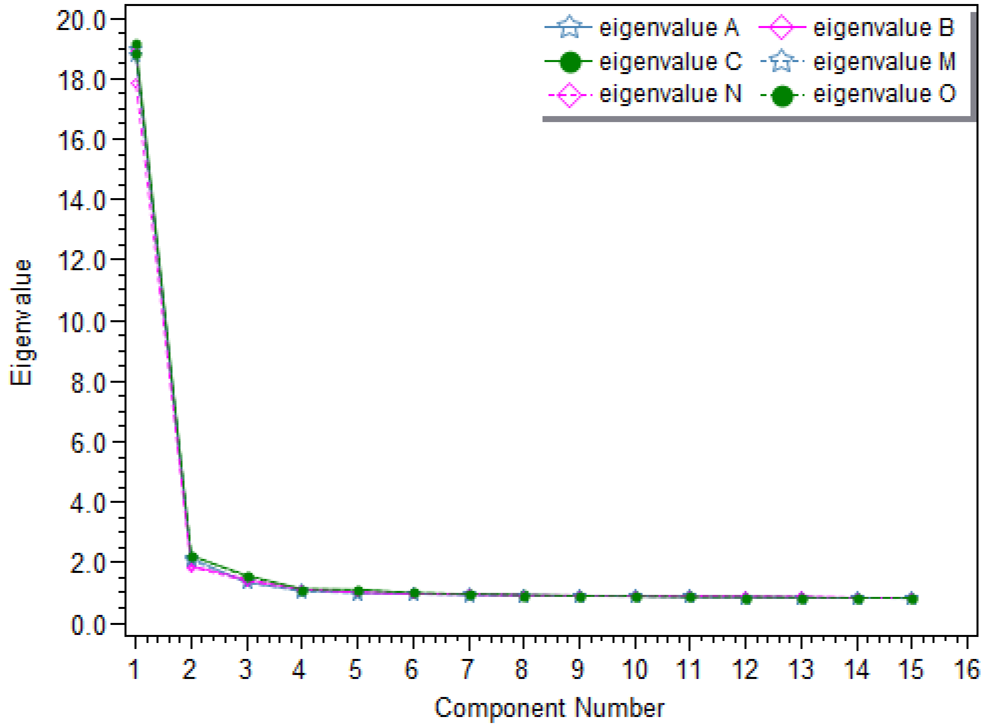


Figure 10-3 Biology Scree Plot of Operational Forms



10.5 Alignment Study

As a part of the larger effort to make a systemic examination of the state's standards-based reform efforts, the NCDPI commissioned the Wisconsin Center for Education Research (WCER) in September, 2014 to conduct an in-depth study of the alignment of the state's newly developed assessments for mathematics, reading, and science to new standards. The current report focuses explicitly on the relationship between new assessments and their respective alignment to content standards or curricular goals. Phase 2 of the study will examine the relationship between instructional practice and relevant content standards based upon a randomly selected representative sample of teachers in the state, while Phase 3 will examine the impact of students' opportunity to learn standards-based content on student achievement. The completed study will provide the state with a unique data set for modeling the performance of the standards-based system as depicted by the various data collection and analysis strategies employed for the study.

Specifically, the current report focuses on describing the alignment characteristics of the assessment program in North Carolina based upon analyses of 42 assessment forms covering state mathematics and reading assessments for grades 3, 4, 5, 6, 7, 8, and HS, as well as state science assessment forms, for grades 5, 8, and HS Biology. The complete report prepared by the WCER is available on the NCDPI website. An abbreviated version of the report with highlighted summaries for reading assessments is included in this section as a part of the validity evidence.

10.5.1 Rationale for Alignment Study

Standards-based educational reform has been the fundamental education model employed by states, and to a growing extent federal policy-makers, for twenty-plus years. Emerging out of the systemic research paradigm popular in the late eighties and early nineties, the standards-based model is essentially a systemic model influencing educational change. The standards-based system is based upon three fundamental propositions: 1) standards will serve as an explicit goal or target toward which curriculum planning, design and implementation will move; 2) accountability for students, teachers, and schools can be determined based upon student performance; and 3) standardized tests are aligned to the state content standards. Woven through these propositions is the notion of alignment, and the importance of it to the standards-based paradigm.

While examination of instructional alignment can help answer the first proposition, and alignment studies of assessments can help assure the third, neither of these can address whether the assumptions of the second proposition are justified. To do this, one must look at the role of both in explaining student achievement. Moreover, in order to address the overall effectiveness of the standards-based system as implemented in one or another location, one must be able to bring together compatible alignment indicators that span the domains of instruction, assessment, and student performance.

The Surveys of Enacted Curriculum (SEC) is unique among alignment methodologies in that it allows one to examine the interrelationships of instruction, assessments, and student performance using an objective, systematic, low-inference, and quantifiable approach to examining alignment issues. The SEC, though best known for its tools for describing instructional practice, provides a methodology and set of data collection and analysis procedures that permit examination of all three propositions in order to consider the relationships between each. This allows for a look at the standards-based system as a whole to determine how well the system is functioning.

This document reports on Phase I of a three phase study commissioned by the NCDPI to examine the effectiveness of the state's efforts to implement a newly structured standards-based system in the state. Phase I focuses on alignment of new assessments developed for EOG and EOC administered by the state from 2012–13. Phase II will focus on instructional alignment, and Phase III will examine student performance in light of students' opportunities to learn standards-based content given the assessments used to generate achievement results. Once all three phases have been completed, the state will have an in-depth look at the state's standards-based system, and a wealth of information in considering its continuing efforts to provide quality educational opportunities to the state's K–12 population.

10.5.2 What Is Alignment Analysis?

Alignment, in terms of characteristics of assessment and instruction, is inherently a question about relationships. How does A relate to B? However, that also means alignment is inherently an abstraction in the sense that it is not easily measurable. Moreover, as with most relationships, the answers aren't simply "yes" or "no," but rather a matter of degree. Relationships also tend to be multidimensional; they have more than a single aspect, dimension, or quality that is important to a full understand of the nature of the alignment relationship. All of these factors make alignment analysis a challenging activity.

Alignment measures in SEC are derived from content descriptions. That is alignment analyses report on the relationship between two multidimensional content descriptions. Each dimension of the two descriptions can then be compared, using procedures described below to derive a set of alignment-indicator measures that summarizes the quantitative relationship between any two content descriptions on any of the dimensions used for describing academic content. In addition to examination of each dimension independently, the method allows for examination of alignment characteristics at the intersection of all three dimensions employed, producing a summative "overall" alignment indicator that has demonstrated a predictive capacity in explaining the variation of students' opportunities to learn assessed content, otherwise referred to as predictive validity.

Content descriptions are described in more detail in Section III of the report. Note that two descriptions of academic content are collected in order to calculate and report alignment results: one a description of the content covered across a series of assessment forms for a particular grade level, and the other a description of the relevant academic content standards for the assessed grade and subject. These content descriptions are systematically compared to determine the alignment characteristics existing between the two descriptions, using a simple iterative algorithm that generates an alignment measure or index based on the relevant dimension(s) of the content being considered.

As mentioned, there are three dimensions to the content descriptions collected, and hence three dimensions upon which to look at the degree of alignment the analyses indicate. These indicator measures can be distilled further to a single overall alignment index (OAI) that summarizes the alignment characteristics of any two content descriptions at the intersection of the three dimensions of content embedded in the SEC approach. These dimensions and the yielded alignment indicators are described next.

10.5.3 The Dimensions of Alignment

SEC content descriptions are collected at the intersection of three dimensions: (1) topic coverage, (2) performance expectation, and (3) relative emphasis. These parallel the three alignment indices that measure the relationship between the two descriptions on one or another of these three dimensions: (1) balance of representation (BR), (2) Topical Coverage (TC), and (3) performance expectations (PE).

When considered in combination with one another, that is when all three dimensions are included in the alignment algorithm; a fourth, summary measure of 'overall alignment' can be calculated. The procedure for calculating alignment is discussed further in the report, as a discussion of what constitutes 'good' alignment using the SEC approach. In short, each alignment indicator is expressed on a scale with a range of 0 to 1.0—with 1.0 representing identical content descriptions (perfect alignment) and 0 indicating no content in common between the two descriptions, or perfect misalignment. A threshold measure is set at 0.5 for each of the four summary indicator measures. Above the 0.5 threshold, alignment is considered to be at an acceptable level; and below it is considered weak or questionable, indicating that a more detailed examination related to that indicator measure is warranted.

10.5.4 Content Analysis Workshop

Content descriptions used to generate visual displays like *Figure 10-1* through *Figure 10-3* in *Section 10.5.8 Alignment Results* were collected using a particular type of document analysis referred to as content analysis. All content analysis work was conducted using teams of content analysts (educators with K–12 content expertise) that received a half day of training at content analysis workshops where specific documents are then analyzed by content-analysis teams over a one or two day period.

North Carolina hosted a content-analysis workshop as part of the alignment study in January 2015 at the McKimmon Conference and Training Center in Raleigh, North Carolina. There 10 subject-based teams of content analysts were formed from more than 30 teachers and other content specialists and trained to conduct independent analyses of 51 assessment forms for mathematics, reading, and science for all assessed grades. Each team was led by a veteran analyst familiar with the process and able to facilitate the conversations among team members. The process involved both independent analysis and group discussion, though group consensus was not required.

The alignment analyses of any two content descriptions are based on detailed comparisons of the descriptive results collected during the content analysis process. While alignment results are based on a straightforward computational procedure and provide precise measures of the relationship between two descriptions, a simple visual comparison of two content maps is often sufficient to identify the key similarities and differences between any two descriptions. For example, a simple visual comparison of the two maps presented in **Error! Reference source not found.** suggests that, while distinctions can be identified, there is a generally similar structure to both that suggests reasonably good alignment of the two descriptions.

10.5.5 Balance of Representation

Among the three content dimensions on which alignment measures are based, two are directly measured and one is derived. That is two of the content dimensions are based upon observer/analyst reports of the occurrence of one or another content description. The derived measure concerns "how much" and is based on the number of reported occurrences for a specific description of content relative to the total number of reports making up the full content description. This yields a proportional measure, summing to 1.00. The SEC refers to this "how much" dimension as "balance of representation" (BR).

As a summary indicator, BR is calculated as the product of two values; the portion of the assessment that targets standards-based content, multiplied by the portion of standards-based content represented in the assessment. For example, if 90% of an assessment (i.e. 10% of the assessment covers content not explicitly referenced in the standards) covered 40% of the standards for a particular grade level (i.e., 60% of the content reflected in the standards was not reflected in the assessment), the BR measure would be 0.36. As with all the summary indicator measures, reported here, the "threshold" for an acceptable degree of alignment is 0.50 or higher. Our example would thus reflect a weak measure of alignment, given this threshold measure. The rationale for this 0.5 measure is discussed in Section II of the full report.

The influence of BR runs through all of the alignment indices, since the relative emphasis of content is the value used in making comparisons between content descriptions. In a very real sense, the dimensions of topic and performance expectation provide the structure for looking at alignment, while the balance of representation provides the values that get placed in that structure. This will become more apparent in the discussion on the calculation of alignment presented in Section II of the report.

For assessments, relative emphasis is expressed in terms of the proportion of score points attributed to one or another topic and/or performance expectation. The relative emphasis refers to the number of times a particular topic and/or performance expectation is noted across all the strands of a standard presented for a given grade and subject. Table 10-4 displays BR Index by grade levels for the NC EOG and EOC science assessments. The summary measures on BR for the assessed grades exceeded the 0.5 threshold. This one measure alone however provides insufficient information for making a judgment regarding alignment. It tells only part of the alignment story. Other indicators provide other perspectives for viewing alignment that help to fill out the full picture of the alignment relationship existing between assessments and standards.

Table 10-4 Balance of Representation Index by Grade

Grade	5	8	Biology
BR Index	0.77	0.54	0.83

10.5.6 Topic Coverage

The first dimension considered in most if not all alignment analyses, regardless of the methodology employed, concerns what Norman Webb (1997) calls categorical concurrence. For convenience, and to better fit the SEC terminology, this indicator is simply referred to as topic coverage (TC) and measures a seemingly simple question; does the topic or subtopic identified in one description match a topic or subtopic occurring in the other description?

Actually, there are a series of questions implied here, each relevant to a comparison of the topics covered in an assessment with those indicated in the relevant target standard:

1) Which topics in the assessment are also in the standards?
2) Which topics in the assessment are not in the standards?
3) Which topics in the standards are in the assessments?
4) Which topics in the standards are not in the assessment?

Each of these represents a distinctly different question that can be asked when comparing topic coverage. The algorithm used to calculate topical concurrence is sensitive to each of these questions, with the resulting index representing, in effect, a composite response to all four questions. Table 10-5 provides the summary alignment results for TC for each of the assessed grades in science. Once again the summary measures for this dimension also indicate above-threshold alignment results, suggesting that the science assessments are well aligned to the standards with respect to topic coverage.

Table 10-5 Topic Coverage Index by Grade

Grade	5	8	Biology
TC Index	0.73	0.63	0.70

10.5.7 Performance Expectations

The SEC taxonomies enable descriptions of academic content based on two dimensions ubiquitous to the field of learning: knowledge and skills. When referencing standards it is frequently summarized with the statement “what students should know and be able to do.” The “what students should know” part refers to topics, while the “be able to do” references expectations for student performance, or performance expectations for short. The SEC taxonomies enable the collection of content descriptions on both of these dimensions, and together form the alignment “target” for both assessments and curriculum.

Just as we can examine alignment with respect to topic coverage only, we can similarly examine the descriptions of performance expectations embedded in the content descriptions of assessments and standards. This alignment indicator is referred to as performance expectations (PE), and is based on the five categories of expectations for student performance employed by the SEC. While the labels vary slightly from subject to subject, the general pattern of expectations follows this general division:

1) Memorization/Recall,
2) Procedural Knowledge,
3) Conceptual Understanding,
4) Analysis, Conjecture and Proof, and
5) Synthesis, Integration and Novel Thinking.

Table 10-6 reports the performance expectation measure across assessed grade levels for science. It is expressed as an index with a range of 0 to 1, with 0.50 indicating acceptable alignment. As can be seen, all subjects/grades surpass this threshold.

Table 10-6 Performance Expectations Index by Grade

Grade	5	8	Biology
PE Index	0.75	0.74	0.72

10.5.8 Alignment Results

While the SEC approach to alignment allows reporting and consideration of the results along each of these three dimensions, the most powerful alignment measure results occur when all three dimensions are combined into an overall index measure that is sensitive to the dynamic interplay of all three dimensions by comparison of content descriptions at the intersection of all three dimensions. Overall alignment results are summarized in Table 10-7.

The resulting alignment index, just like the summary indices for each dimension reported separately, has a range of 0.00 to 1.00, with 0.50 or higher indicating adequate overall alignment. Despite the higher alignment index each for balance of representation, topic coverage, and performance expectation, the overall alignment indices for the grade levels are borderline above the acceptable range.

Table 10-7 Overall Alignment Index by Grade

Grade	5	8	Biology
OAI	0.55	0.56	0.52

Examples of content maps with content descriptions, relative emphasis, and performance expectations for the EOG and EOC assessments are shown in *Figure 10-4* through *Figure 10-6*. *Figure 10-4* indicates that the topics with the strongest emphasis in North Carolina’s grade 5 science standards (map to the right “Target Content Areas”) are energy, motion and forces, and meteorology, particularly at the performance level of communicate and memorize. A careful visual review of the two maps in *Figure 10-4* in terms of the three alignment dimensions indicates the following:

- Balance of Representation (BR): the two figures are shaped similarly, which indicates a good balance of representation for EOG grade 5 science assessments. This is also confirm by a BR index of 0.77 see Table 10-4**.**
- Topic Coverage (TC): topics with the strongest emphasis in both maps are energy, motion and forces, and meteorology, where the contour lines are closer together. This indicates the assessment blueprint is aligned to the content standards with respect to TC. The TC index for EOG grade 5 science is 0.73 above the threshold of 0.50, see Table 10-5.
- Performance Expectation (PE): PE focuses on what students should "be able to do" more generally summarized by DOK levels. From the grade 5 assessment map (left) the two strongest topics of emphasis are mostly assessed with communication and memorize and weak emphasis on analysis.

Note that the content description maps provided in the figures are displayed along three axes or dimensions; the Y-axis, represented by the list of science topics presented to the right of the image, the X-axis represented by the five categories of performance expectations running across the bottom of the image, and the Z-axis (displayed by contour lines and color bands), indicating the relative emphasis for each intersection of topic and performance expectation. These three dimensions form the foundational structure for describing and analyzing content using the SEC approach. Academic content is described in terms of the interaction of topic and performance expectations. By measuring each occurrence of some element of content (topic by performance expectation), a measure of the relative emphasis of each content topic, as it appears in the content description, can be obtained.

Figure 10-4 EOG Grade 5 Science Assessment and Standard Content Map

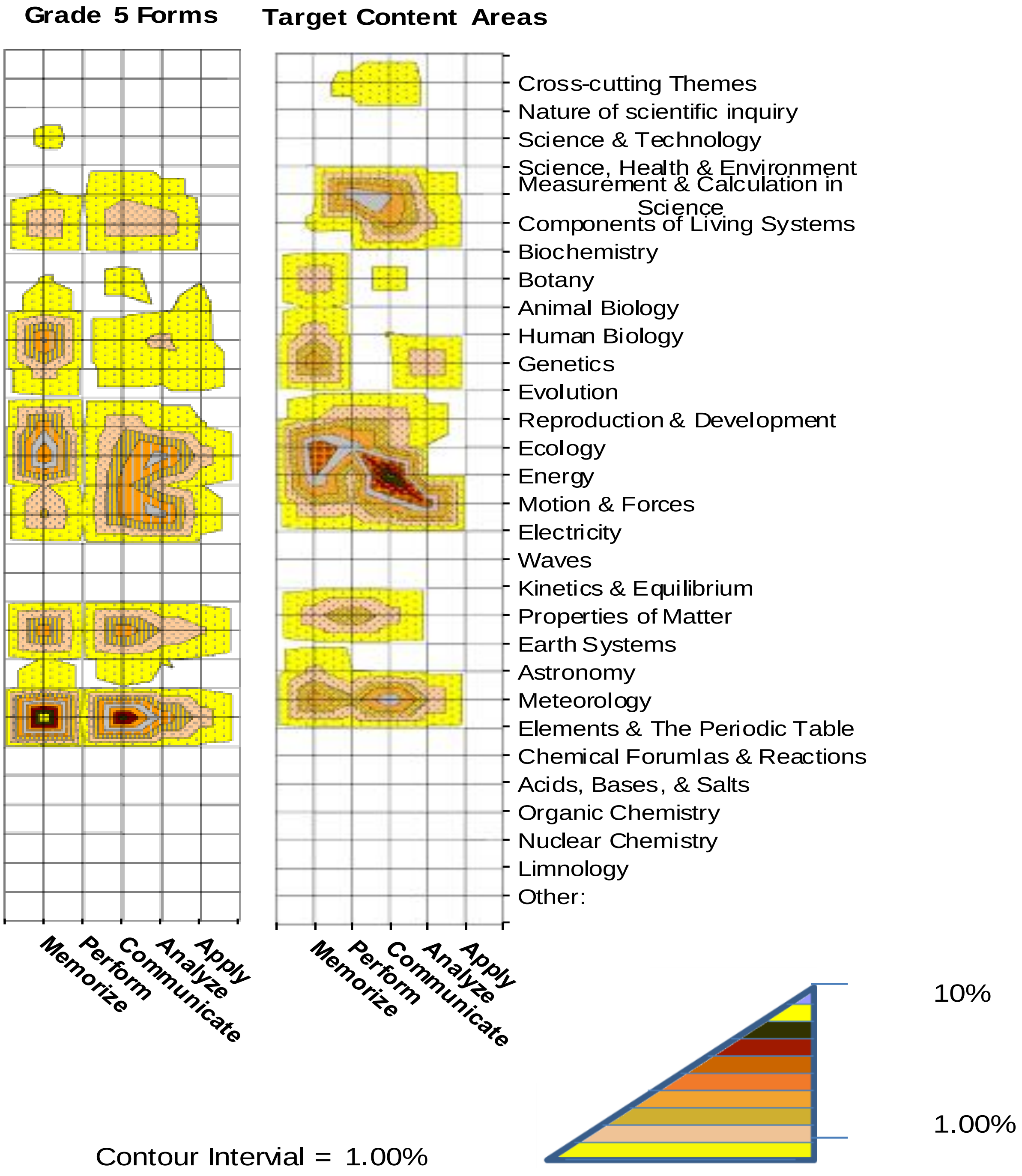


Figure 10-5 EOG Grade 8 Science Assessment and Standard Content Map

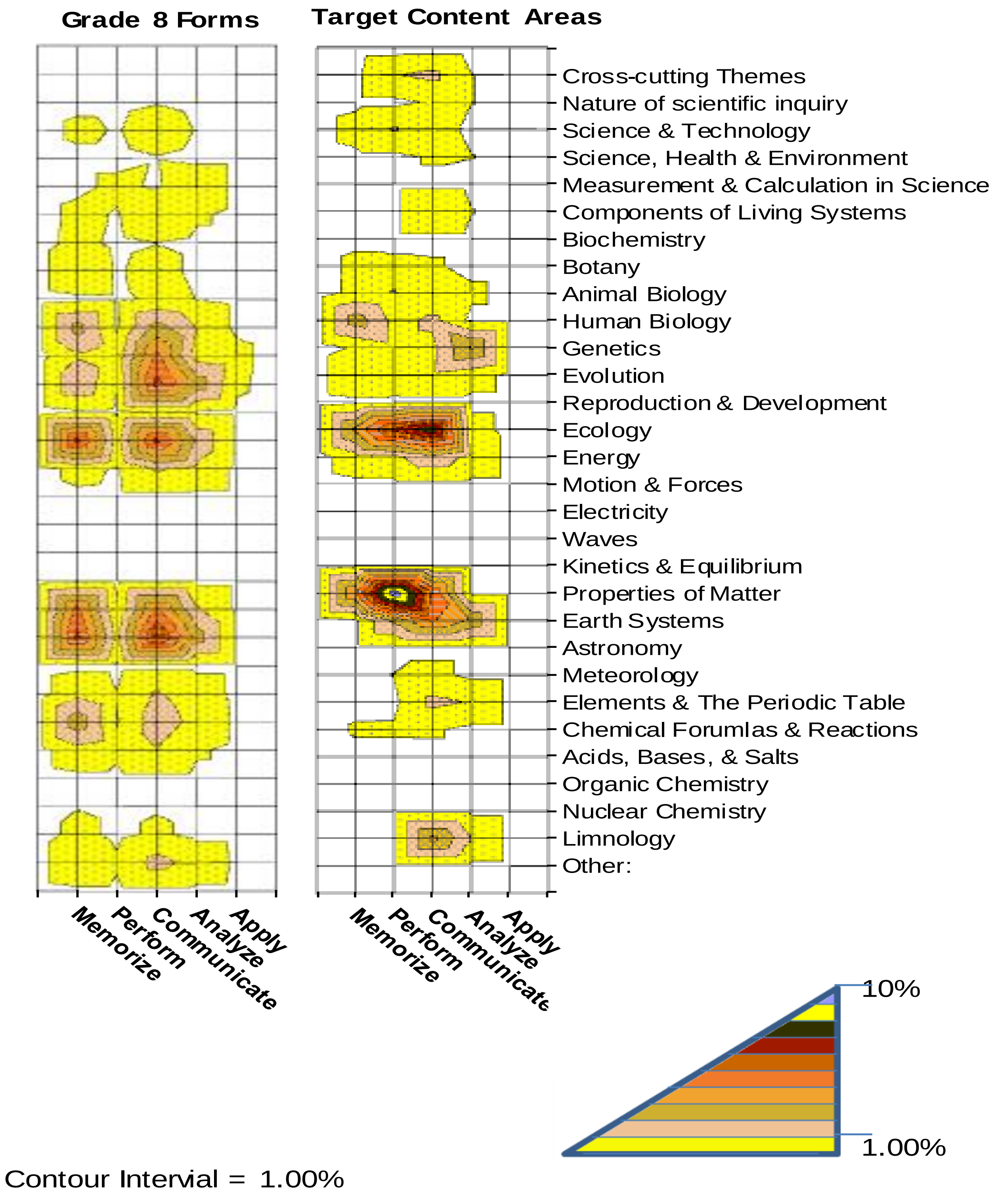
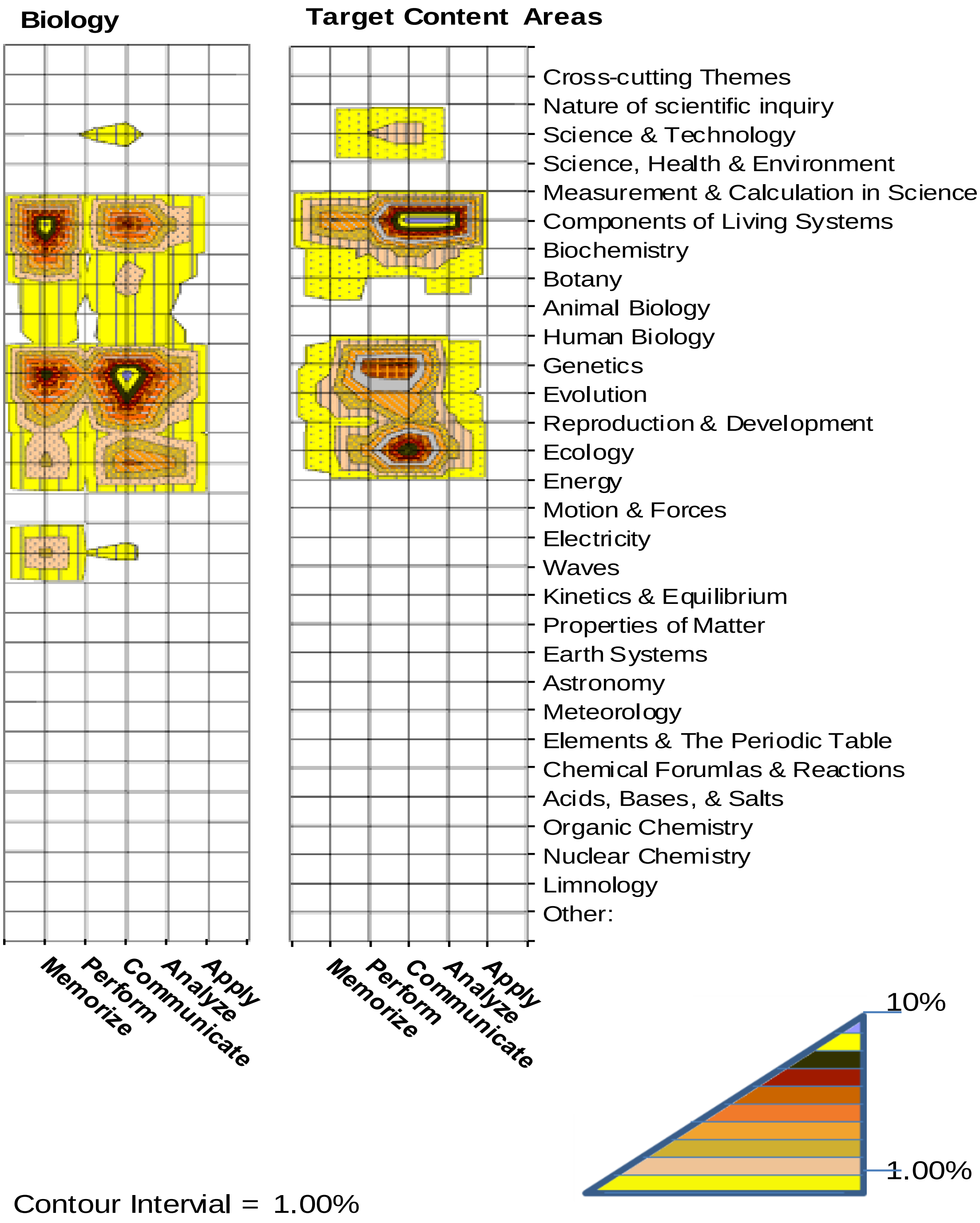


Figure 10-6 EOC Biology Assessment and Standard Content Map



10.5.9 Discussion of Alignment Study Findings

As indicated by the results from the Phase I alignment study presented in sections above, the science assessments used by the state across the grades covered by this study reveal acceptable levels of alignment. The results make clear that the design of the assessments attend to the content embedded in the standards, and the implementation of that design yielded assessment instruments with good alignment characteristics across the board as measured by the SEC methodology.

There are a number of mediating contextual issues that should be considered in making final determination of any alignment result. For example, the selection of an appropriate alignment target may justify a narrowing of the standards content considered for alignment purposes. Moreover, while the threshold measure provides a convenient benchmark against which to compare results, it is a measure selected by convention, and these measures are indicators of alignment that must be considered within the real-world contexts of assessment validity and economic feasibility.

Once student performance data has been collected (Phase III of the study), additional information will be available regarding the impact of the assessments' alignment characteristics on student performance, controlling for students' opportunity to learn standards-based (and/or) assessment-based content. Such analyses may provide additional data to assist state leaders in determining the adequacy of the state's assessment program.

The results reported here mark a good beginning for the larger study of which this alignment study represents only one part. With the collection of instructional practice data in Phase II, along with results of student performance on the assessment in Phase III, the analysis team will have the necessary data to better understand and describe the impact of instructional practice and assessment design on student achievement, thereby providing the means to determine the relative health of the state's assessment and instructional programs. Perhaps more importantly, the results from the full study will provide both teachers and educators with valuable information regarding the curriculum and assessment strategies employed in classrooms around the state and their impact on student learning.

10.5.10 Conclusion of the Phase I Alignment Study

This study collected and examined a comprehensive set of content descriptions covering the full span of the assessment instruments for science grades 5 and 8 EOG and Biology EOC. The resulting content descriptions provide a unique set of visual displays depicting assessed content and provide the Department a rich descriptive resource for reviewing and reflecting upon the assessment program being implemented throughout the state. Alignment analyses indicated that the science assessments administered by the state are, for the most part, reasonably aligned.

10.6 Evidence Regarding Relationship with External Variables

Analysis of the relationship of test scores to variables external to the test provides another important source of convergent and divergent validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs (AERA, APA, & NCME, 2014). Evidence regarding relationships with criteria (i.e., previously termed criterion-related validity) of a test indicates the effectiveness of a test in predicting an individual's behavior in a specific situation. The criterion for evaluating the performance of a test can be measured at the same time (concurrent validity) or at some later time (predictive validity).

For the North Carolina grades 5 and 8 Science EOG and Biology EOC tests, teachers' judgments of student achievement, expected grade, and assigned achievement levels all serve as sources of evidence of concurrent validity. The variables used in the analysis are as follows.

- **Teacher Judgment of Achievement Level**: For all students participating in the test, teachers were asked to evaluate their students' absolute abilities, external to the test, based on their knowledge of the students' achievements. The categories that teachers could use correspond to the achievement level descriptors.
- **Expected Grade**: Teachers were also asked to provide the letter grade that they anticipated each student would receive at the end of the grade or course.
- **Assigned Achievement Level**: the achievement level assigned to a student based on his or her test score, based on the cut scores previously described
- **Score**: the converted raw-score-to-scale-score value obtained by each examinee

The Pearson correlation coefficient for these variables ranged from 0.49 to 0.78 (see **Table 10-8**), a moderate correlation that sufficiently demonstrates that teachers can reasonably predict students' achievement level. Moreover, the correlation coefficients between the Science and Reading/math scale scores for a given grade were moderate ranging from 0.70 to 0.78, suggesting that those who do well in science also did reasonably well in ELA/Reading and math or vice versa. The lowest correlation (0.53, 0.49, and 0.57 respectively for grades 5, 8, and Biology) was observed for assigned achievement level and expected grade.

Table 10-8 Pearson Correlation Coefficient between Teacher expected Grade and Actual Grade EOG and EOC Science Tests

Grade	5	8	Biology
Teacher Judgment of Achievement Level by Assigned Achievement Level	0.60	0.59	0.60
Teacher Judgment of Achievement Level by Expected Grade	0.72	0.69	0.69
Teacher Judgment of Achievement Level by Scale Score	0.64	0.62	0.62
Assigned Achievement Level by Expected Grade	0.53	0.49	0.57
Expected Grade by Scale Score	0.56	0.52	0.59
Science/Biology by Math/Algebra I Scale Score	0.74	0.73	0.70
Science/Biology by ELA/English II Scale Score	0.78	0.75	0.72

Note: Correlation is significant at .01 level (2-tailed).

10.7 Fairness and Accessibility

10.7.1 Accessibility in Universal Design

To ensure fairness and accessibility for all eligible students for NC assessments, the principle of Universal Design was embedded throughout the development and design of EOG and EOC assessments. The EOG and EOC assessments measure what students know and are able to do as defined in the North Carolina State Content Standards. Assessments must ensure comprehensible access to the contents being measured to allow students to accurately demonstrate their standing in the contents assessed. In order to ensure items and assessments were developed with universal design principles, NCDPI organized a workshop “Plain English Strategies: Research, Theory, and Implications for Assessment development” in April 2011. Dr. Edynn Sato who was then Director of Research and English Learner Assessment at WestEd was invited to train NCDPI test development staff including curriculum staff as well as employees from NC-TOPS on universal design principles and writing in plain English language. The universal design principles were applied in every step of the test development, administration, and reporting.

Evidence of universal design principles applied in the development of EOG and EOC assessments, so that students could show what they know and are able to do so, has been documented throughout the item development and review, form review, and test administration sections in the report. More detailed descriptions regarding plain language in the item development process are presented in Section 3.3 of this report. Some of the Universal Design principles applied include:

- Precisely Defined Constructs
    - Direct match to objective being measured
- Accessible, Nonbiased Items[j]
    - Accommodations from the start (Braille, large print, oral presentation etc.)
    - Ensure that quality is retained in all items
- Simple, Clear Directions and Procedures
    - Presented in understandable language
    - Use simple, high frequency, and compound words
    - Use words that are directly related to content the student is expected to know
    - Omit words with double meanings or colloquialisms
    - Consistency in procedures and format in all content areas
- Maximum Legibility
    - Simple fonts
    - Use of white space
    - Headings and graphic arrangement
    - Direct attention to relative importance
    - Direct attention to the order in which content should be considered
- Maximum Readability: Plain Language
    - Increases validity to the measurement of the construct
    - Increases the accuracy of the inferences made from the resulting data
    - Active instead of passive voice
    - Short sentences

[j] See discussions on Bias review in chapter 4

    - Common, everyday words
    - Purposeful graphics to clarify what is being asked

- Accommodations
    - One item per page
    - Extended time for ELL Students
    - Test in separate room

10.7.2 Fairness in Access

As documented throughout chapter 3 and alignment evidence presented in Section 10.5 of this report, the NCDPI ensured that all assessment blueprints are aligned to agree upon content domains that are also aligned to the NC*ESS*. Assessments' content domain specifications and blueprint are published on the NCDPI public website with other relevant information regarding the development of EOG and EOC assessments. This ensures schools and students have exposure to content being targeted in the assessments and that the schools provide them with opportunity to learn.

Prior to the administration of the first operational form of EOG and EOC assessments, NCDPI also published released forms for every grade level which were constructed using the same blueprint as the operational forms. These release forms provided students, teachers, and parents with sample items and a general practice form similar to the operational assessment. These release forms also served as a resource to familiarize students with the various response format in the new assessments and the online platform. It is recommended that students be given the opportunity to view the large font and/or alternate background color versions of the online tutorial and released forms of the assessment (with the device to be used on test day) to determine which mode of administration is appropriate. A periodic table was provided to all students for their reference.

10.7.3 Fairness in Administration

Chapter 5 of this report documents the procedures put in place by the NCDPI to assure the administration of EOG and EOC assessments are standardized, fair and, secured for all students across the state. For each assessment NCDPI publishes an "Assessment Guide" which is the main training material for all test administrators across the state. These guides provide comprehensive details of key features about each assessment. Key information provided includes a general overview of each assessment which covers the purpose of the assessment, eligible students, testing window, and makeup testing options. Assessment guides also cover all preparations and steps that should be followed the day before testing, on test day, and after testing. Samples of answer sheets are also provided in the assessment guide. In addition to assessment guides used to train test administrators, NCDPI also publishes a "Proctor Guide" which is used by test coordinators to train proctors.

Computer-based assessments are available to all students in regular or large font and in alternate background colors (i.e., yellow, green, gray, black background with yellow text); however, the North Carolina Department of Public Instruction (NCDPI) recommends these options be considered only for students who routinely use similar tools (e.g., color acetate overlays, colored background paper, and large print text) in the classroom. It is recommended that students be given the opportunity to view the large font and/or alternate background color versions of the online tutorial and released forms of the assessment (with the device to be used on test day) to determine which mode of administration is appropriate.

Additionally, NCDPI recommends that the Online Assessment Tutorial should be used to determine students' appropriate font size (i.e., regular or large) and/or alternate background color for test day. These options must be entered in the student's interface questions (SIQ) before test day. The Online Assessment Tutorial can assist students, whose IEP or Section 504 Plan designates the Large Print accommodation, in determining if the large font will be sufficient on test day. If the size of the large font is not sufficient for a student because of his/her disability, this accommodation may be used in conjunction with the Magnification Devices accommodation, or a Large Print Edition of the paper and-pencil assessment may be ordered.

10.7.4 Fairness Across Forms and Modes

The *standards* (AERA, NCME & APA, 2014) states that "When multiple forms of a test are prepared, the same test specifications should govern all of the forms (*p82-83*)." It is imperative that when multiple forms are created from the same test blueprint, the resulting test scores from parallel forms should be comparable, and it should make no difference to students which form was administered. For EOG and EOC assessments, parallel forms were created based on the same content and statistical specifications. As shown in Chapter 7 all parallel forms were constructed and matched to have the same CTT and IRT properties of average p-value, reliability, and closely aligned TCCs, as well as CSEM. Meeting these criteria ensured that the test forms are essentially parallel. Moreover, these forms were spiraled within the class to obtain equivalent samples for calibration and scaling. This ensured that each form was administered to a random equivalent sample of student across the state. Any difference in form difficulty was accounted for during separate group calibration as the random group data design ensured all parameters were put onto the same IRT scale and separate raw-to-scale tables were created to adjust for any form differences.

In order to ensure that scores from forms administered across mode (paper and computer) were comparable, DIF sweep procedure was implemented during item analysis. The DIF sweep procedure flags items that show a significant differential item parameter between computer and paper modes. These items, even though identical, are treated as unique items during joint calibration of computer and paper forms. The process involved two steps; in step 1, items were calibrated in each mode separately, and their estimated item parameters were evaluated. If the estimated parameters showed no evidence of mode effect, then the two sets of responses were concurrently calibrated to estimate the final item parameters. If the estimated parameters showed a sign of mode effect, then in step 2, those items that exhibited no DIF were considered anchor, and a separate set of item parameters were estimated for each item by mode that exhibited DIF. This process ensured that the item parameters and test scores are in a common IRT scale and mode effects are accounted for. Finally, the resulting item parameters were used to create a separate raw-to-scale score table for each form by modes.

As a part of the continuous validity framework adopted, the NCDPI has plans to conduct a comprehensive comparability study of mode effects. The methodology will be based on selecting random matched samples using the propensity score matching procedure using relevant matching variables. The results from the two equivalent samples will be evaluated in terms of item parameter estimates and their impact on raw-to-scale score conversion as well as proficiency classifications.

Furthermore, to ensure equitable access for students taking computer-based forms across devices and comparability of scores, the NCDPI has set minimum device requirements that will guarantee all items and forms will exhibit acceptable functionality as intended. These requirements were based on review of industry standards and usability studies and research findings conducted with other national testing programs. The NCDPI device requirements for EOG and EOC computer-based assessments includes:

- Minimum screen size of 9.5 inches
- Minimum screen resolution of 1024 x 768
- iPads must use Guided Access or a Mobile Device management system to restrict the iPad to only run the NCTest iPad App.
- Screen capture capabilities must be disabled.
- Chrome App on desktops and laptops requires the Chrome Browser version 43 or higher.
- Windows machines must have a minimum of 512 MB of RAM.
- Pentium 4 or newer processor for Windows machines and Intel for MacBooks

In addition to the technical specification of devices, the NCDPI also conducted a review of each sample item across devices i.e., laptops, iPad, and desktops, to make sure items are rendered as intended. Reviews also checked to make functionalities of the test platform, such as audio files, large font, and high contrast versions. The technical specifications may be reviewed at https://center.ncsu.edu/nc/course/view.php?id=361.

Glossary of Key Terms

The terms below are defined by their application in this document and their common uses in the North Carolina Testing Program. Some of the terms refer to complex statistical procedures used in the process of test development. In an effort to avoid the use of excessive technical jargon, definitions have been simplified; however, they should not be considered exhaustive.

Accommodations		Changes made in the format or administration of the test to provide options to test takers who are unable to take the original test under standard test conditions.
Achievement levels		Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum classified by broad ranges of performance.
Asymptote		An item statistic that describes the proportion of examinees that endorsed a question correctly but did poorly on the overall test. Asymptote for a theoretical four-choice item is 0.25 but can vary somewhat by test.
Biserial correlation		The relationship between an item score (right or wrong) and a total test score.
Cut scores		A specific point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.
Dimensionality		The extent to which a test item measures more than one ability.

Embedded test model		Using an operational test to field test new items or sections. The new items or sections are “embedded” into the new test and appear to examinees as being indistinguishable from the operational test.
Equivalent forms		Statistically insignificant differences between forms (i.e., the red form is not harder).
Field test		A collection of items to approximate how a test form will work. Statistics produced will be used in interpreting item behavior/performance and allow for the calibration of item parameters used in equating tests.
Foil counts		Number of examinees that endorse each foil (e.g. number who answer “A,” number who answer “B,” etc.).
Item response theory		A method of test item analysis that takes into account the ability of the examinee and determines characteristics of the item relative to other items in the test. The NCDPI uses the 3-parameter model, which provides slope, threshold, and asymptote.
Item tryout		A collection of a limited number of items of a new type, a new format, or a new curriculum. Only a few forms are assembled to determine the performance of new items and not all objectives are tested.

Mantel-Haenszel		A statistical procedure that examines the differential item functioning (DIF) or the relationship between a score on an item and the different groups answering the item (e.g. gender, race). This procedure is used to identify individual items for further bias review.
Operational test		Test is administered statewide with uniform procedures, full reporting of scores, and stakes for examinees and schools.
p-value		Difficulty of an item defined by using the proportion of examinees who answered an item correctly.
Parallel form		Covers the same curricular material as other forms.
Percentile		The score on a test below which a given percentage of scores fall.
Pilot test		Test is administered as if it were "the real thing" but has limited associated reporting or stakes for examinees or schools.
Raw score		The unadjusted score on a test determined by counting the number of correct answers.
Scale score		A score to which raw scores are converted by numerical transformation. Scale scores allow for comparison of different forms of the test using the same scale.

Slope		The ability of a test item to distinguish between examinees of high and low ability.
Standard error of measurement		The standard deviation of an individual's observed scores, usually estimated from group data.
Test blueprint		The testing plan, which includes the numbers of items from each objective that are to appear on a test and the arrangement of objectives.
Threshold		The point on the ability scale where the probability of a correct response is fifty percent. Threshold for an item of average difficulty is 0.00.

References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing.* Washington, D.C.: Author.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the Beta-Binomial model for classification consistency and accuracy*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment (CASMA).

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling* [Computer software]. Lincolnwood, IL: Scientific Software International.

Cai, L. (2012). *flexMIRT^{TM} version 1.88: A numerical engine for multilevel item factor analysis and test scoring.* [Computer software]. Seattle, WA: Vector Psychometric Group.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications, Inc.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed.). Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 22*(3), 297-334.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Kluwer-Nijhoff Publishing.

Hanson, B.A. & Brennan, R.L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, 27(4), 345-359.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Lewis, D. M., Green, D. R., Mitzel, H.C., Baum, K. & Patz, R.J. (1998). *The Bookmark standard setting procedure: Methodology and recent implementations*. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Linn, R. L. (2002). The measurement of student achievement in international studies. In A. C. Porter & A. Gamoran (Eds). *Methodological Advances in Large-Scale Cross-National Education Surveys* (pp. 25-57). Washington, DC: Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education, National Academy Press.

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

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

SAS Institute, Inc. (1985). *SAS User's Guide: Statistics, Version 5 Edition.* Cary, NC: Author.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, NJ: Erlbaum.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy.* (Synthesis Report 41). Minneapolis, MN: National Center on Educational Outcomes. Retrieved [January 25, 2016] from http://www.cehd.umn.edu/nceo/onlinepubs/Synthesis41.html

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2005). *Web Alignment Tool*. Wisconsin Center of Educational Research. University of Wisconsin-Madison. Retrieved [January, 2016] from http://wat.wceruw.org/index.aspx

Testing Code of Ethics

Introduction

In North Carolina, standardized testing is an integral part of the educational experience of all students. When properly administered and interpreted, test results provide an independent, uniform source of reliable and valid information, which enables:

- *students* to know the extent to which they have mastered expected knowledge and skills and how they compare to others;
- *parents* to know if their children are acquiring the knowledge and skills needed to succeed in a highly competitive job market;
- *teachers* to know if their students have mastered grade-level knowledge and skills in the curriculum and, if not, what weaknesses need to be addressed;
- *community leaders and lawmakers* to know if students in North Carolina schools are improving their performance over time and how the students compare with students from other states or the nation; and
- *citizens* to assess the performance of the public schools.

Testing should be conducted in a fair and ethical manner, which includes:

Security

- assuring adequate security of the testing materials before, during, and after testing and during scoring
- assuring student confidentiality

Preparation

- teaching the tested curriculum and test-preparation skills
- training staff in appropriate testing practices and procedures
- providing an appropriate atmosphere

Administration

- developing a local policy for the implementation of fair and ethical testing practices and for resolving questions concerning those practices
- assuring that all students who should be tested are tested
- utilizing tests which are developmentally appropriate
- utilizing tests only for the purposes for which they were designed

Scoring, Analysis and Reporting

- interpreting test results to the appropriate audience
- providing adequate data analyses to guide curriculum implementation and improvement

Because standardized tests provide only one valuable piece of information, such information should be used in conjunction with all other available information known about a student to assist in improving student learning. The administration of tests required by applicable statutes and the use of student data for personnel/program decisions shall comply with the *Testing Code of Ethics* (16 NCAC 6D .0306), which is printed on the next three pages.

.0306 TESTING CODE OF ETHICS

(a) This Rule shall apply to all public school employees who are involved in the state testing program.

(b) The superintendent or superintendent's designee shall develop local policies and procedures to ensure maximum test security in coordination with the policies and procedures developed by the test publisher. The principal shall ensure test security within the school building.

(1) The principal shall store test materials in a secure, locked area. The principal shall allow test materials to be distributed immediately prior to the test administration. Before each test administration, the building level test coordinator shall accurately count and distribute test materials. Immediately after each test administration, the building level test coordinator shall collect, count, and return all test materials to the secure, locked storage area.

(2) "Access" to test materials by school personnel means handling the materials but does not include reviewing tests or analyzing test items. The superintendent or superintendent's designee shall designate the personnel who are authorized to have access to test materials.

(3) Persons who have access to secure test materials shall not use those materials for personal gain.

(4) No person may copy, reproduce, or paraphrase in any manner or for any reason the test materials without the express written consent of the test publisher.

(5) The superintendent or superintendent's designee shall instruct personnel who are responsible for the testing program in testing administration procedures. This instruction shall include test administrations that require procedural modifications and shall emphasize the need to follow the directions outlined by the test publisher.

(6) Any person who learns of any breach of security, loss of materials, failure to account for materials, or any other deviation from required security procedures shall immediately report that information to the principal, building level test coordinator, school system test coordinator, and state level test coordinator.

(c) Preparation for testing.

(1) The superintendent shall ensure that school system test coordinators:

(A) secure necessary materials;

(B) plan and implement training for building level test coordinators, test administrators, and proctors;

(C) ensure that each building level test coordinator and test administrator is trained in the implementation of procedural modifications used during test administrations; and

(D) in conjunction with program administrators, ensure that the need for test modifications is documented and that modifications are limited to the specific need.

(2) The principal shall ensure that the building level test coordinators:

(A) maintain test security and accountability of test materials;

(B) identify and train personnel, proctors, and backup personnel for test administrations; and

(C) encourage a positive atmosphere for testing.

(3) Test administrators shall be school personnel who have professional training in education and the state testing program.

(4) Teachers shall provide instruction that meets or exceeds the standard course of study to meet the needs of the specific students in the class. Teachers may help students improve test-taking skills by:

(A) helping students become familiar with test formats using curricular content;

(B) teaching students test-taking strategies and providing practice sessions;

(C) helping students learn ways of preparing to take tests; and

(D) using resource materials such as test questions from test item banks, testlets and linking documents in instruction and test preparation.

(d) Test administration.

(1) The superintendent or superintendent's designee shall:

(A) assure that each school establishes procedures to ensure that all test administrators comply with test publisher guidelines;

(B) inform the local board of education of any breach of this code of ethics; and

(C) inform building level administrators of their responsibilities.

(2) The principal shall:

(A) assure that school personnel know the content of state and local testing policies;

(B) implement the school system's testing policies and procedures and establish any needed school policies and procedures to assure that all eligible students are tested fairly;

(C) assign trained proctors to test administrations; and

(D) report all testing irregularities to the school system test coordinator.

(3) Test administrators shall:

(A) administer tests according to the directions in the administration manual and any subsequent updates developed by the test publisher;

(B) administer tests to all eligible students;

(C) report all testing irregularities to the school system test coordinator; and

(D) provide a positive test-taking climate.

(4) Proctors shall serve as additional monitors to help the test administrator assure that testing occurs fairly.

(e) Scoring. The school system test coordinator shall:

(1) ensure that each test is scored according to the procedures and guidelines defined for the test by the test publisher;

(2) maintain quality control during the entire scoring process, which consists of handling and editing documents, scanning answer documents, and producing electronic files and reports. Quality control shall address at a minimum accuracy and scoring consistency.

(3) maintain security of tests and data files at all times, including:

(A) protecting the confidentiality of students at all times when publicizing test results; and

(B) maintaining test security of answer keys and item-specific scoring rubrics.

(f) Analysis and reporting. Educators shall use test scores appropriately. This means that the educator recognizes that a test score is only one piece of information and must be interpreted together with other scores and indicators. Test data help educators understand educational patterns and practices. The superintendent shall ensure that school personnel analyze and report test data ethically and within the limitations described in this paragraph.

(1) Educators shall release test scores to students, parents, legal guardians, teachers, and the media with interpretive materials as needed.

(2) Staff development relating to testing must enable personnel to respond knowledgeably to questions related to testing, including the tests, scores, scoring procedures, and other interpretive materials.

(3) Items and associated materials on a secure test shall not be in the public domain. Only items that are within the public domain may be used for item analysis.

(4) Educators shall maintain the confidentiality of individual students. Publicizing test scores that contain the names of individual students is unethical.

(5) Data analysis of test scores for decision-making purposes shall be based upon:

(A) dissagregation of data based upon student demographics and other collected variables;

(B) examination of grading practices in relation to test scores; and

(C) examination of growth trends and goal summary reports for state-mandated tests.

(g) Unethical testing practices include, but are not limited to, the following practices:

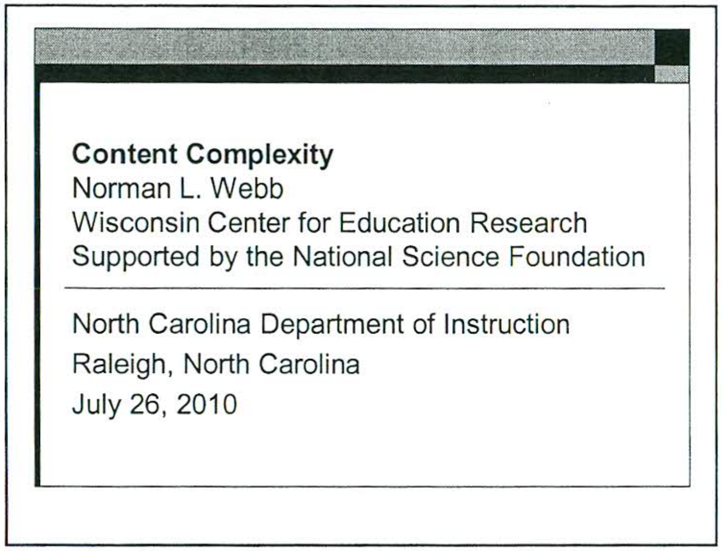
(1) encouraging students to be absent the day of testing;
(2) encouraging students not to do their best because of the purposes of the test;
(3) using secure test items or modified secure test items for instruction;
(4) changing student responses at any time;
(5) interpreting, explaining, or paraphrasing the test directions or the test items;
(6) reclassifying students solely for the purpose of avoiding state testing;
(7) not testing all eligible students;
(8) failing to provide needed modifications during testing, if available;
(9) modifying scoring programs including answer keys, equating files, and lookup tables;
(10) modifying student records solely for the purpose of raising test scores;
(11) using a single test score to make individual decisions; and
(12) misleading the public concerning the results and interpretations of test data.

(h) In the event of a violation of this Rule, the SBE may, in accordance with the contested case provisions of Chapter 150B of the General Statutes, impose any one or more of the following sanctions:

(1) withhold ABCs incentive awards from individuals or from all eligible staff in a school;
(2) file a civil action against the person or persons responsible for the violation for copyright infringement or for any other available cause of action;
(3) seek criminal prosecution of the person or persons responsible for the violation; and
(4) in accordance with the provisions of 16 NCAC 6C .0312, suspend or revoke the professional license of the person or persons responsible for the violation.

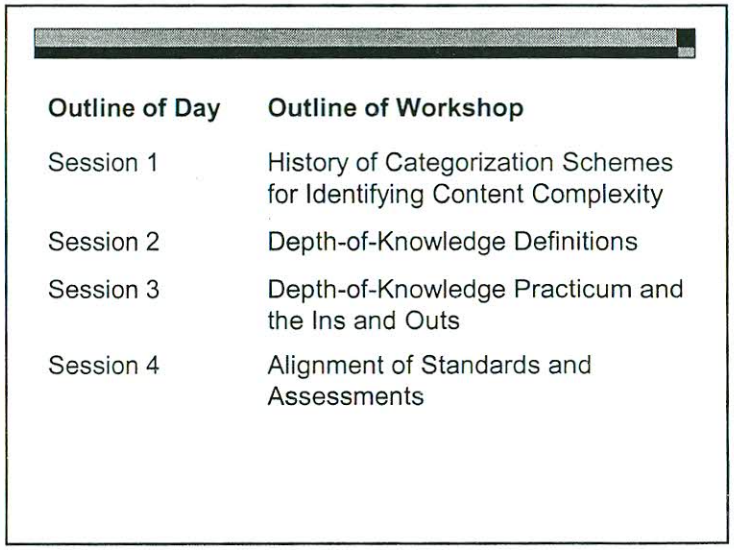
History Note: Authority G.S. 115C-12(9)c.; 115C-81(b)(4);
Eff. November 1, 1997;
Amended Eff. August 1, 2000.

APPENDIX 3-A NORM WEBB TRAINING-CONTENT COMPLEXITY

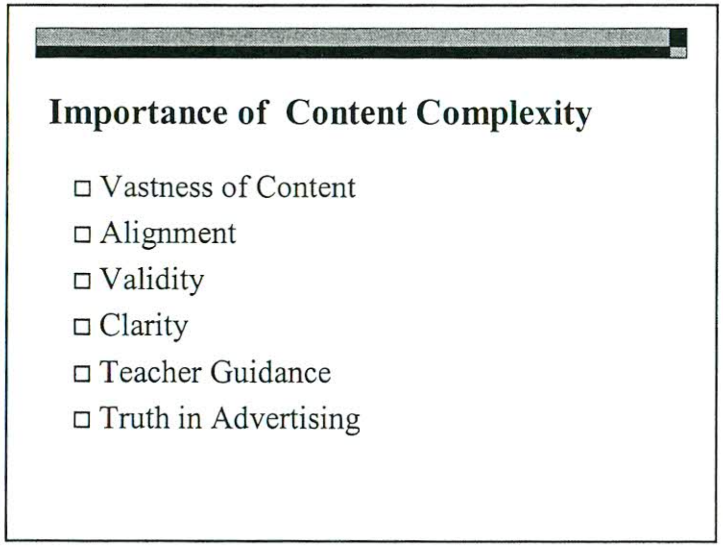


Content Complexity
Norman L. Webb
Wisconsin Center for Education Research
Supported by the National Science Foundation

North Carolina Department of Instruction
Raleigh, North Carolina
July 26, 2010

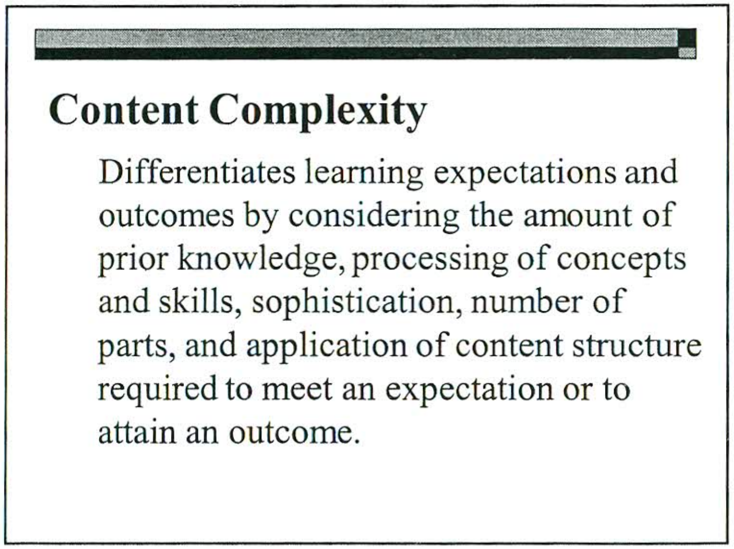


Outline of Day	Outline of Workshop
Session 1	History of Categorization Schemes for Identifying Content Complexity
Session 2	Depth-of-Knowledge Definitions
Session 3	Depth-of-Knowledge Practicum and the Ins and Outs
Session 4	Alignment of Standards and Assessments



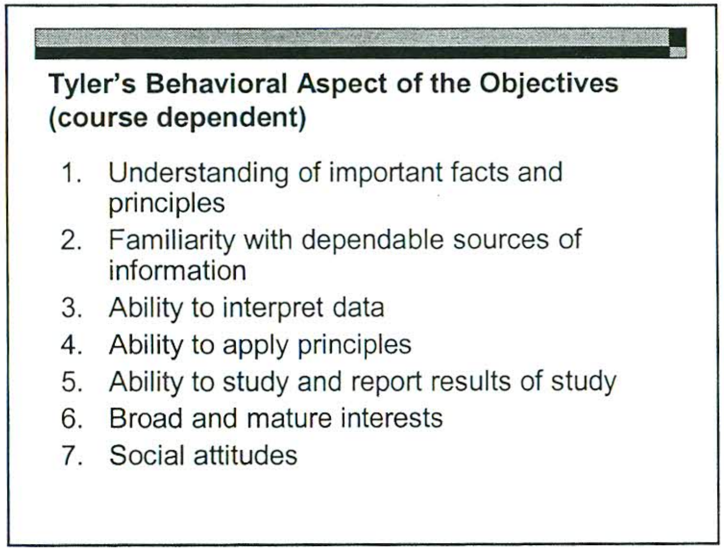
Importance of Content Complexity

- □ Vastness of Content
- □ Alignment
- □ Validity
- □ Clarity
- □ Teacher Guidance
- □ Truth in Advertising



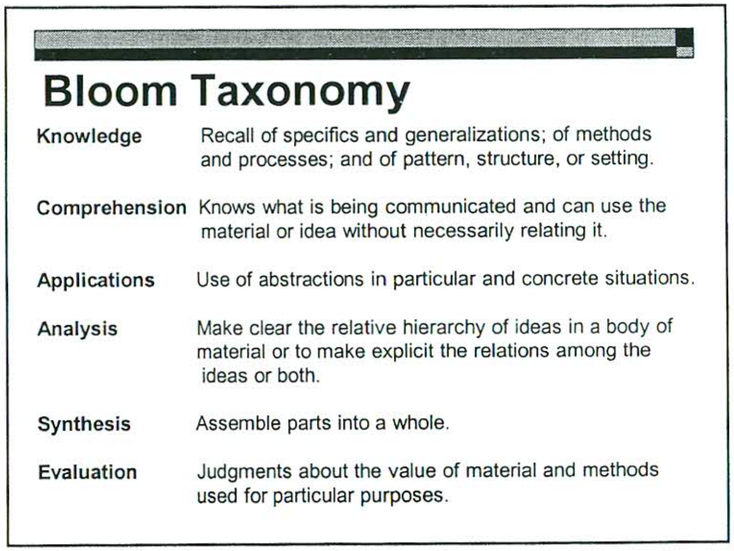
Content Complexity

Differentiates learning expectations and outcomes by considering the amount of prior knowledge, processing of concepts and skills, sophistication, number of parts, and application of content structure required to meet an expectation or to attain an outcome.



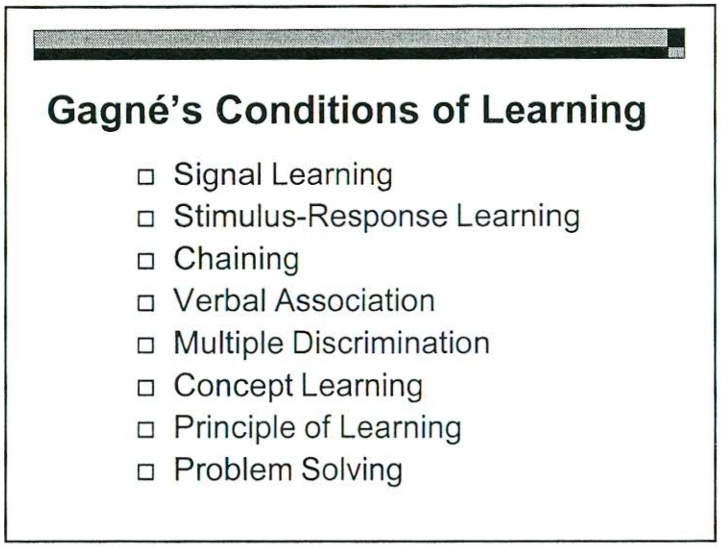
Tyler's Behavioral Aspect of the Objectives (course dependent)

1. Understanding of important facts and principles
2. Familiarity with dependable sources of information
3. Ability to interpret data
4. Ability to apply principles
5. Ability to study and report results of study
6. Broad and mature interests
7. Social attitudes



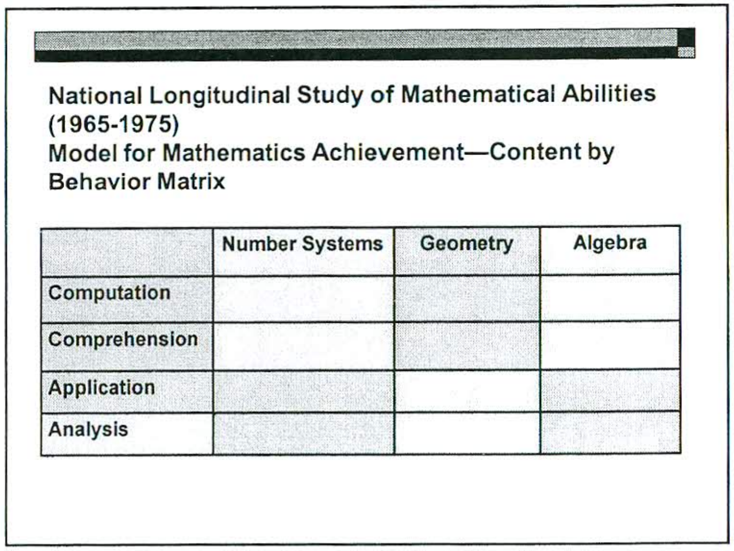
Bloom Taxonomy

Knowledge	Recall of specifics and generalizations; of methods and processes; and of pattern, structure, or setting.
Comprehension	Knows what is being communicated and can use the material or idea without necessarily relating it.
Applications	Use of abstractions in particular and concrete situations.
Analysis	Make clear the relative hierarchy of ideas in a body of material or to make explicit the relations among the ideas or both.
Synthesis	Assemble parts into a whole.
Evaluation	Judgments about the value of material and methods used for particular purposes.



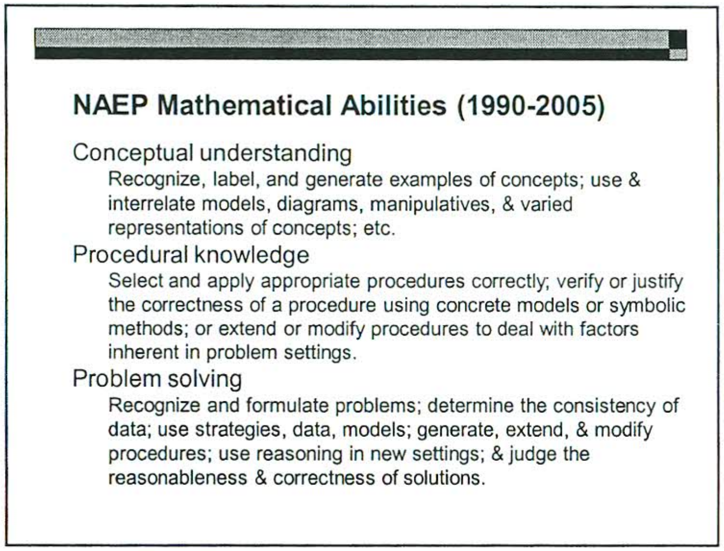
Gagné's Conditions of Learning

- □ Signal Learning
- □ Stimulus-Response Learning
- □ Chaining
- □ Verbal Association
- □ Multiple Discrimination
- □ Concept Learning
- □ Principle of Learning
- □ Problem Solving



National Longitudinal Study of Mathematical Abilities (1965-1975)
Model for Mathematics Achievement—Content by Behavior Matrix

	Number Systems	Geometry	Algebra
Computation		▨	
Comprehension		▨	
Application	▨		▨
Analysis	▨		▨



NAEP Mathematical Abilities (1990-2005)

Conceptual understanding

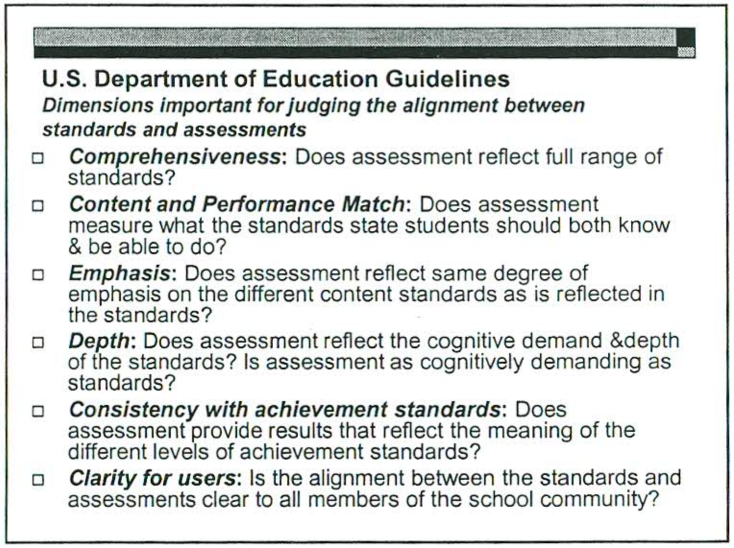
Recognize, label, and generate examples of concepts; use & interrelate models, diagrams, manipulatives, & varied representations of concepts; etc.

Procedural knowledge

Select and apply appropriate procedures correctly; verify or justify the correctness of a procedure using concrete models or symbolic methods; or extend or modify procedures to deal with factors inherent in problem settings.

Problem solving

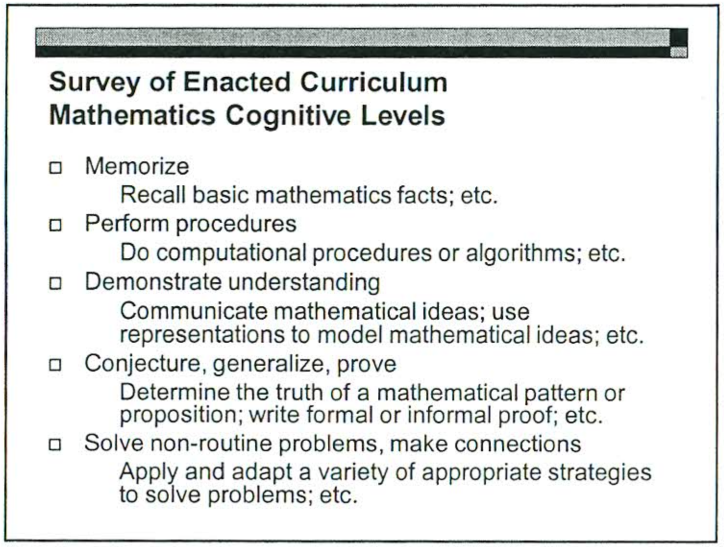
Recognize and formulate problems; determine the consistency of data; use strategies, data, models; generate, extend, & modify procedures; use reasoning in new settings; & judge the reasonableness & correctness of solutions.



U.S. Department of Education Guidelines

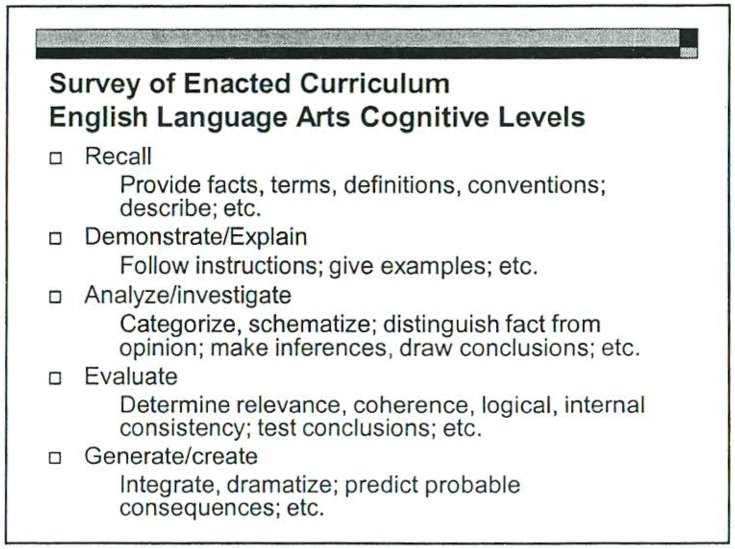
Dimensions important for judging the alignment between standards and assessments

- □ ***Comprehensiveness*:** Does assessment reflect full range of standards?
- □ ***Content and Performance Match*:** Does assessment measure what the standards state students should both know & be able to do?
- □ ***Emphasis*:** Does assessment reflect same degree of emphasis on the different content standards as is reflected in the standards?
- □ ***Depth*:** Does assessment reflect the cognitive demand &depth of the standards? Is assessment as cognitively demanding as standards?
- □ ***Consistency with achievement standards*:** Does assessment provide results that reflect the meaning of the different levels of achievement standards?
- □ ***Clarity for users*:** Is the alignment between the standards and assessments clear to all members of the school community?



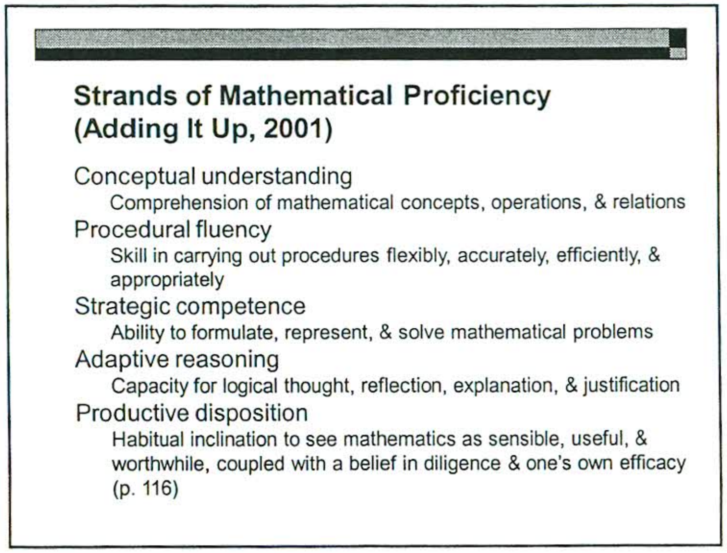
Survey of Enacted Curriculum Mathematics Cognitive Levels

- □ Memorize
  - Recall basic mathematics facts; etc.
- □ Perform procedures
  - Do computational procedures or algorithms; etc.
- □ Demonstrate understanding
  - Communicate mathematical ideas; use representations to model mathematical ideas; etc.
- □ Conjecture, generalize, prove
  - Determine the truth of a mathematical pattern or proposition; write formal or informal proof; etc.
- □ Solve non-routine problems, make connections
  - Apply and adapt a variety of appropriate strategies to solve problems; etc.



Survey of Enacted Curriculum English Language Arts Cognitive Levels

- □ Recall
  - Provide facts, terms, definitions, conventions; describe; etc.
- □ Demonstrate/Explain
  - Follow instructions; give examples; etc.
- □ Analyze/investigate
  - Categorize, schematize; distinguish fact from opinion; make inferences, draw conclusions; etc.
- □ Evaluate
  - Determine relevance, coherence, logical, internal consistency; test conclusions; etc.
- □ Generate/create
  - Integrate, dramatize; predict probable consequences; etc.



Strands of Mathematical Proficiency (Adding It Up, 2001)

Conceptual understanding

Comprehension of mathematical concepts, operations, & relations

Procedural fluency

Skill in carrying out procedures flexibly, accurately, efficiently, & appropriately

Strategic competence

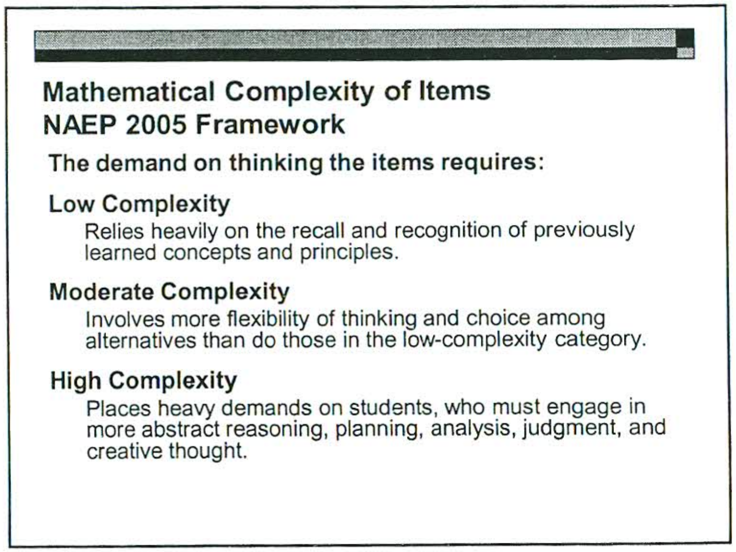
Ability to formulate, represent, & solve mathematical problems

Adaptive reasoning

Capacity for logical thought, reflection, explanation, & justification

Productive disposition

Habitual inclination to see mathematics as sensible, useful, & worthwhile, coupled with a belief in diligence & one's own efficacy (p. 116)



Mathematical Complexity of Items NAEP 2005 Framework

The demand on thinking the items requires:

Low Complexity

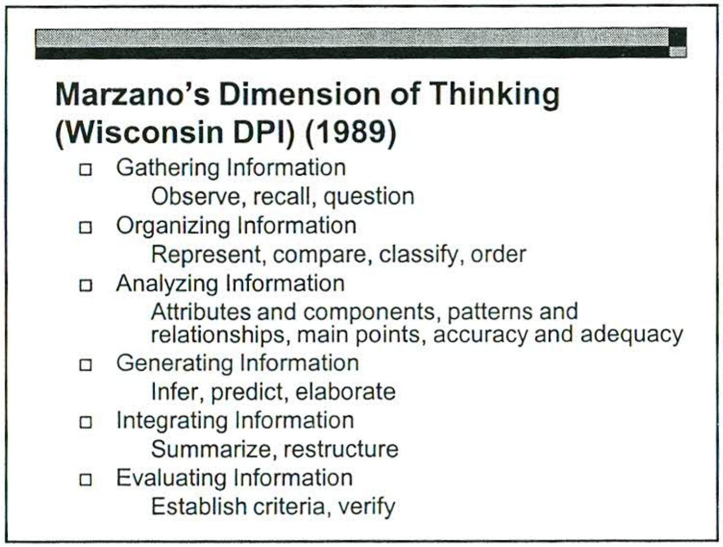
Relies heavily on the recall and recognition of previously learned concepts and principles.

Moderate Complexity

Involves more flexibility of thinking and choice among alternatives than do those in the low-complexity category.

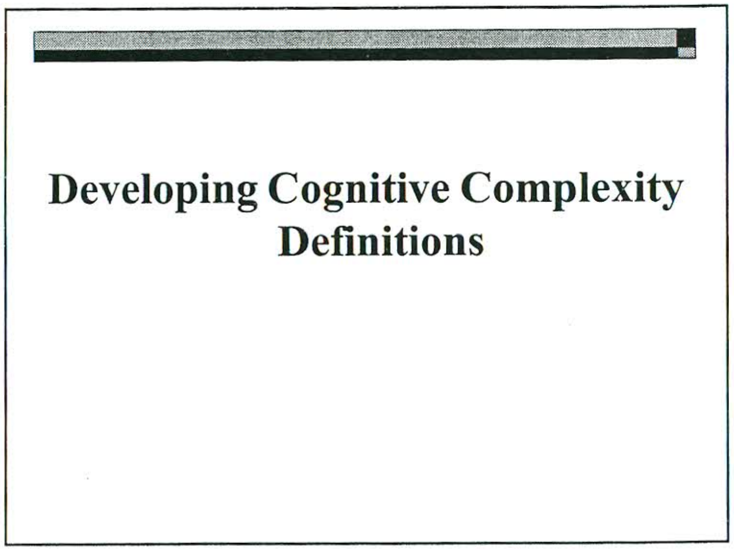
High Complexity

Places heavy demands on students, who must engage in more abstract reasoning, planning, analysis, judgment, and creative thought.

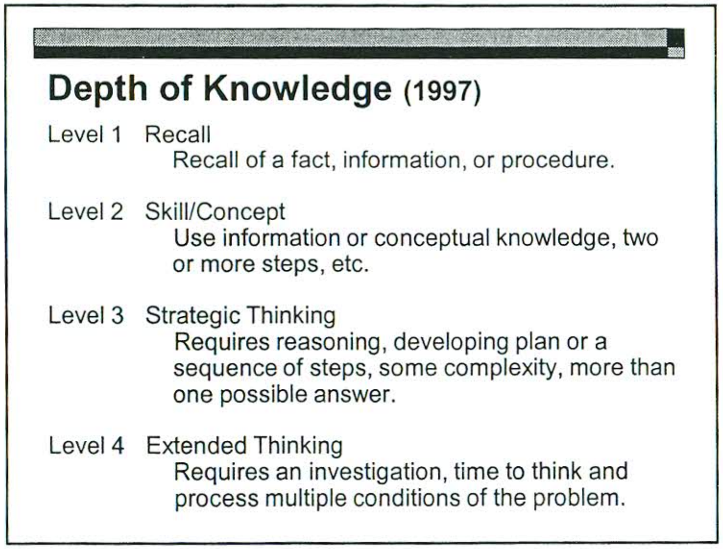


Marzano's Dimension of Thinking (Wisconsin DPI) (1989)

- □ Gathering Information
  Observe, recall, question
- □ Organizing Information
  Represent, compare, classify, order
- □ Analyzing Information
  Attributes and components, patterns and relationships, main points, accuracy and adequacy
- □ Generating Information
  Infer, predict, elaborate
- □ Integrating Information
  Summarize, restructure
- □ Evaluating Information
  Establish criteria, verify



Developing Cognitive Complexity Definitions



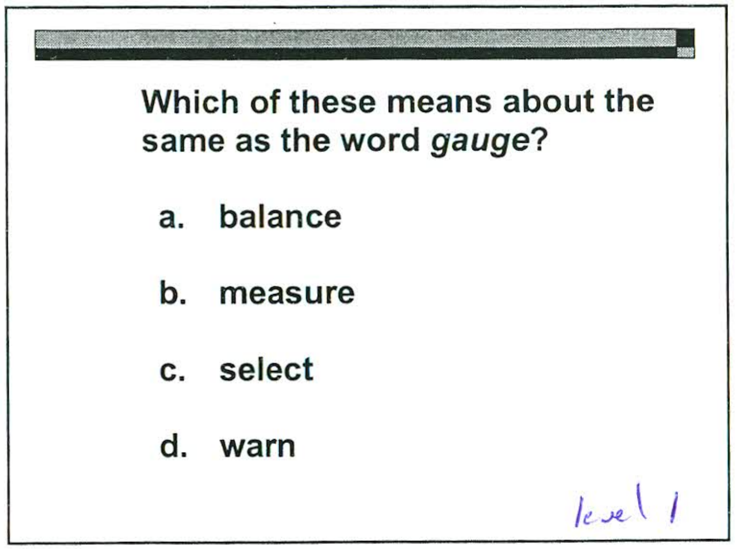
Depth of Knowledge (1997)

Level 1 Recall
Recall of a fact, information, or procedure.

Level 2 Skill/Concept
Use information or conceptual knowledge, two or more steps, etc.

Level 3 Strategic Thinking
Requires reasoning, developing plan or a sequence of steps, some complexity, more than one possible answer.

Level 4 Extended Thinking
Requires an investigation, time to think and process multiple conditions of the problem.



Which of these means about the same as the word *gauge*?

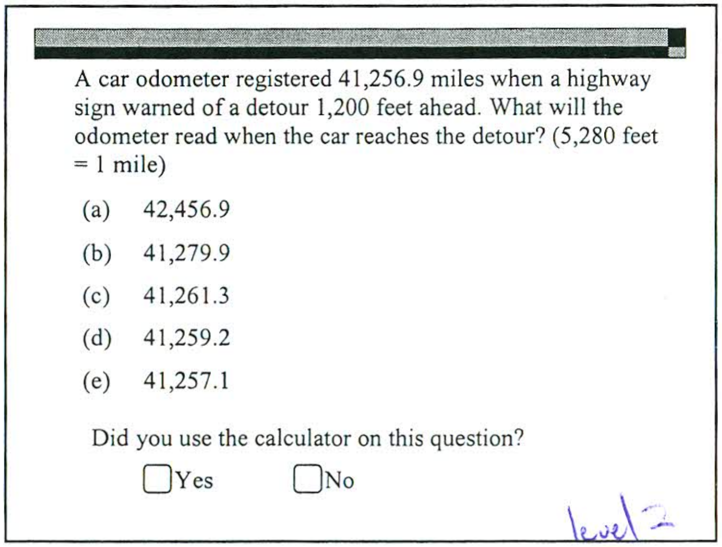
a. balance

b. measure

c. select

d. warn

level 1



A car odometer registered 41,256.9 miles when a highway sign warned of a detour 1,200 feet ahead. What will the odometer read when the car reaches the detour? (5,280 feet = 1 mile)

(a) 42,456.9

(b) 41,279.9

(c) 41,261.3

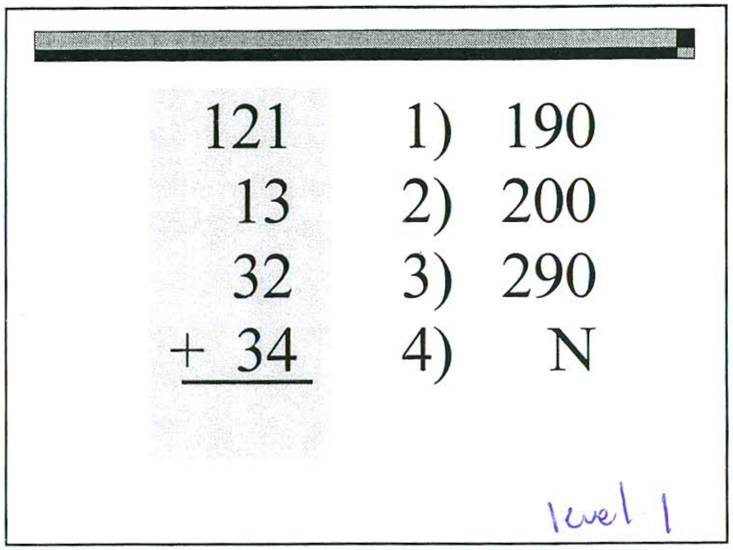
(d) 41,259.2

(e) 41,257.1

Did you use the calculator on this question?

□Yes □No

level 2



$$\begin{array}{r} 121 \\ 13 \\ 32 \\ +\ 34 \\ \hline \end{array}$$

1) 190
2) 200
3) 290
4) N

level 1

Which of these conclusions is best supported by information from the passage?

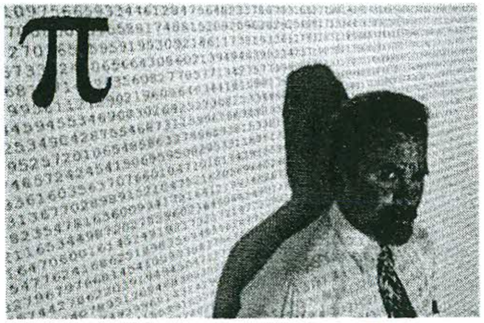
a. If a candidate meets the personal and educational qualifications and is in fair physical shape, his or her chances of becoming an agent are very good.

b. Compared with other law enforcement agencies in the country, the F.B.I. has a low success rate for tracking down and apprehending suspected offenders.

c. The job of an agent is not for everyone; it takes someone with special training who is not afraid of danger and doesn't mind being socially isolated at times.

d. The life of a federal investigator is not as interesting as most people think; agents spend most of their time working at desks.

It is Still A Level 1



Marc Umile poses for a picture in front of a projection of the string of numbers knows as pi in Philadelphia, Friday, March, 2, 2006. Umile is among a group of people fascinated with pi, a number that has been computed to more than a trillion decimal places. He has recited pi to 12,887 digits, perhaps the U.S. record. (AP Photo/Matt Rourke)

Depth of Knowledge Framework for the Wisconsin Knowledge and Concepts Examinations Re-alignment Study

TerraNova Thinking Skill	Descriptor	Depth of Knowledge Levels			
		1—Recall of Information	2—Basic Reasoning	3—Complex Reasoning	4—Extended Reasoning
Gathering Information	Observe	✓			
	Recall	✓			
	Question	✓	✓		
Organizing Information	Represent	✓	✓		
	Compare		✓		
	Classify		✓		
	Order		✓		
Analyzing Information	Attributes & Components	✓	✓		
	Patterns & Relationships		✓		
	Main Points		✓		
	Accuracy & Adequacy		✓		
Generating Information	Infer		✓	✓	
	Predict		✓	✓	
	Elaborate		✓	✓	
Integrating Information	Summarize				
	Restructure		✓	✓	
Evaluating Information	Establish Criteria		✓	✓	
	Verify		✓	✓	

Hess's Bloom's & DOK Levels

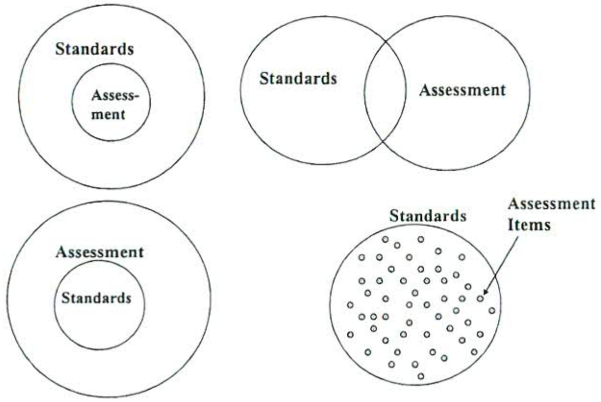
Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 Recall & Reproduction	Level 2 Skills & Concepts	Level 3 Strategic Thinking/ Reasoning	Level 4 Extended Thinking
Remember	▨			
Understand	▨	▨	▨	▨
Apply	▨	▨	▨	▨
Analyze	▨	▨	▨	▨
Evaluate			▨	▨
Create	▨	▨	▨	▨

Review DOK Definitions and Sample Objectives and Items

Alignment Process

- □ Identify Standards and Assessments
- □ Select 6-8 Reviewers (Content Experts)
- □ Train Reviewers on DOK Levels
- □ Part I: Code DOK Levels of the Standards/Objectives
- □ Part II: Code DOK Levels and Corresponding Objectives of Assessment Items

Degree of Alignment



Specific Criteria

Content Focus

A. Categorical Concurrence

B. Depth-of-Knowledge Consistency

C. Range-of-Knowledge Correspondence

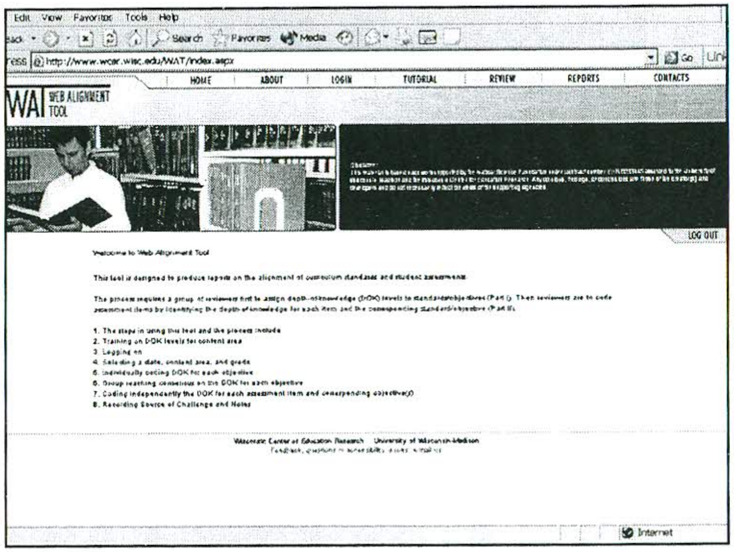
D. Balance of Representation

Alignment Levels Using the Four Criteria

Alignment Level	Categorical Concurrence	Depth of Knowledge	Range of Knowledge	Balance of Representation
Acceptable	6 item per standard	50%	50%	0.70
Weak		40% - 49%	40% - 49%	.60 - .69
Unacceptable	Less than 6 items per standard	Less than 40%	Less than 40%	Less than .60

Coding Process Tips

- □ One Primary Objective and up to Two Secondary Objectives (if necessary)
- □ Source of Challenge (a correct/incorrect response for the wrong reason)
- □ Notes (any insights to share)
- □ Consider Full Range of Standards
- □ Use generic objectives sparingly



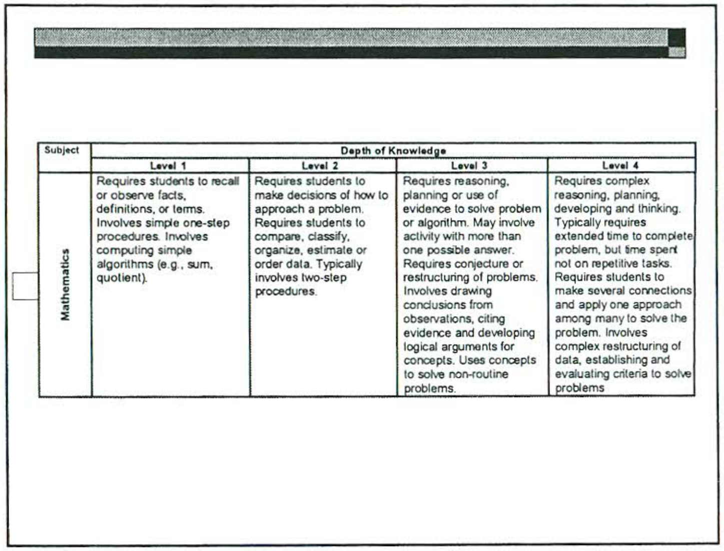
Welcome to Web Alignment Tool

This tool is designed to produce reports on the alignment of curriculum standards and student assessments.

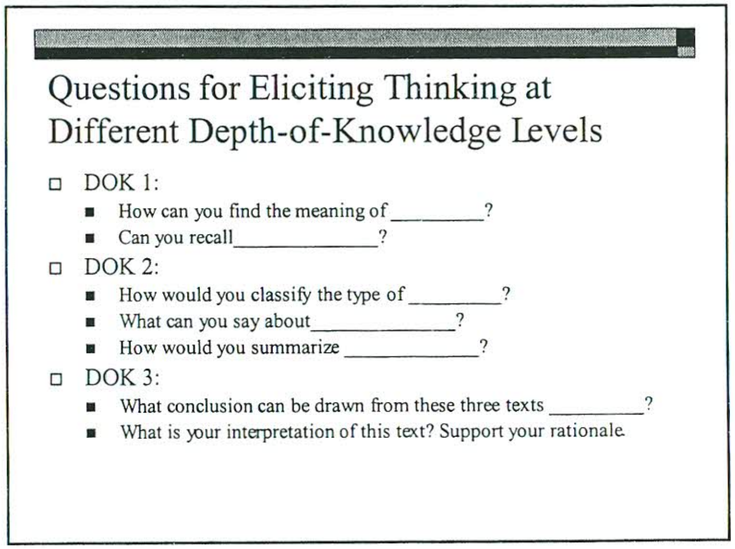
The process requires a group of reviewers first to assign depth-of-knowledge (DOK) levels to standards/objectives (Part I). Then reviewers are to code assessment items by identifying the depth-of-knowledge for each item and the corresponding standard/objective (Part II).

1. The steps in using this tool and the process include
2. Training on DOK levels for content area
3. Logging on
4. Selecting a date, content area, and grade
5. Individually coding DOK for each objective
6. Group reaching consensus on the DOK for each objective
7. Coding independently the DOK for each assessment item and corresponding objective(s)
8. Recording Source of Challenge and Notes

Wisconsin Center of Education Research University of Wisconsin-Madison
Feedback, questions or accessibility issues: e-mail us



Subject	Depth of Knowledge			
	Level 1	Level 2	Level 3	Level 4
Mathematics	Requires students to recall or observe facts, definitions, or terms. Involves simple one-step procedures. Involves computing simple algorithms (e.g., sum, quotient).	Requires students to make decisions of how to approach a problem. Requires students to compare, classify, organize, estimate or order data. Typically involves two-step procedures.	Requires reasoning, planning or use of evidence to solve problem or algorithm. May involve activity with more than one possible answer. Requires conjecture or restructuring of problems. Involves drawing conclusions from observations, citing evidence and developing logical arguments for concepts. Uses concepts to solve non-routine problems.	Requires complex reasoning, planning, developing and thinking. Typically requires extended time to complete problem, but time spent not on repetitive tasks. Requires students to make several connections and apply one approach among many to solve the problem. Involves complex restructuring of data, establishing and evaluating criteria to solve problems

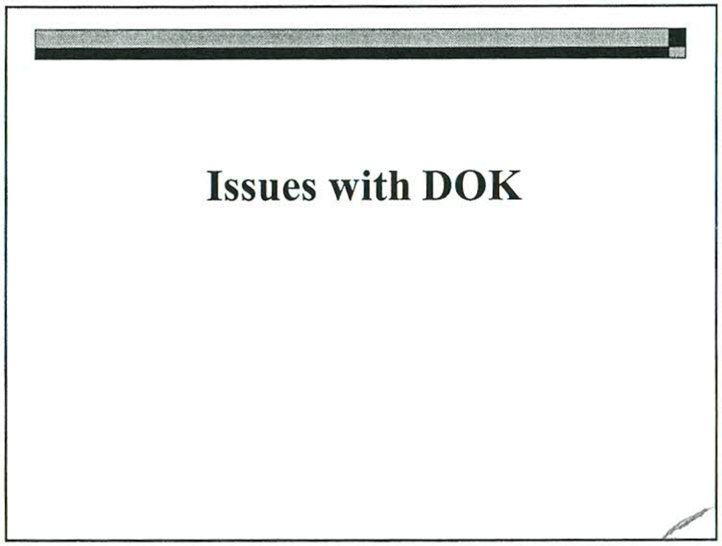


Questions for Eliciting Thinking at Different Depth-of-Knowledge Levels

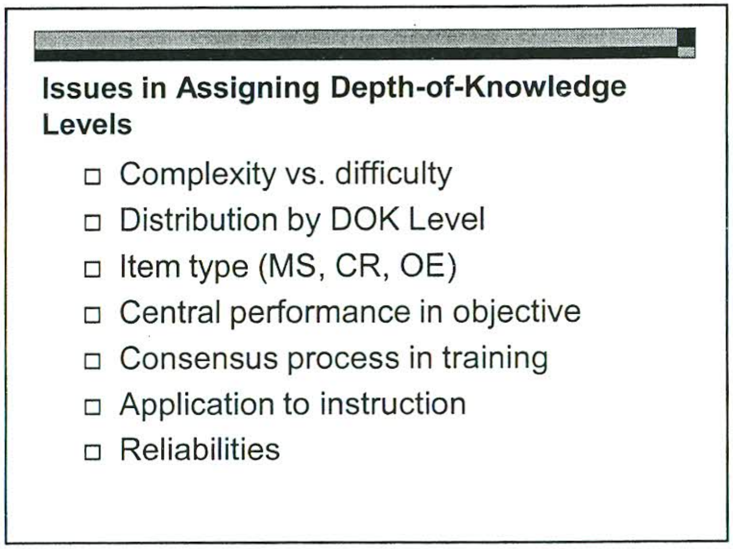
□ DOK 1:
- How can you find the meaning of_________?
- Can you recall_______________?

□ DOK 2:
- How would you classify the type of _________?
- What can you say about_______________?
- How would you summarize _____________?

□ DOK 3:
- What conclusion can be drawn from these three texts _________?
- What is your interpretation of this text? Support your rationale.



Issues with DOK



Issues in Assigning Depth-of-Knowledge Levels

□ Complexity vs. difficulty

□ Distribution by DOK Level

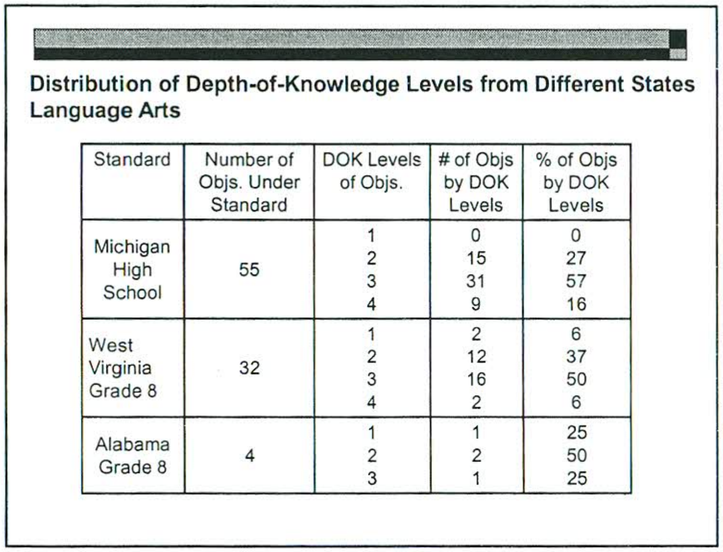
□ Item type (MS, CR, OE)

□ Central performance in objective

□ Consensus process in training

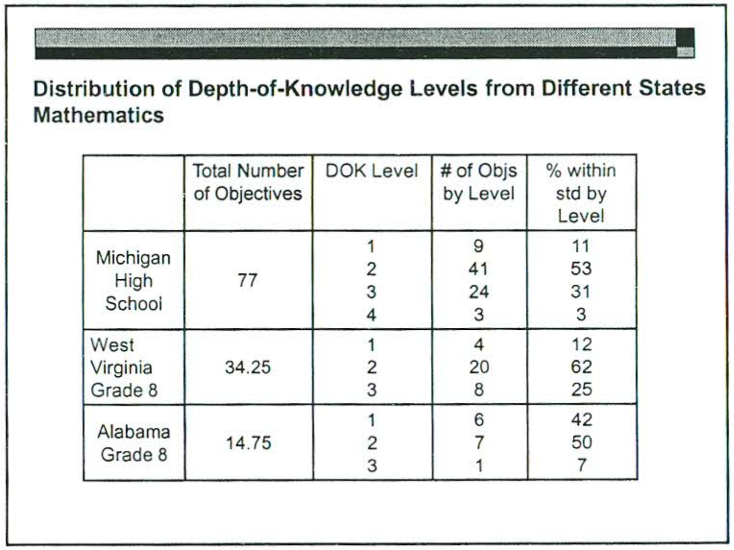
□ Application to instruction

□ Reliabilities



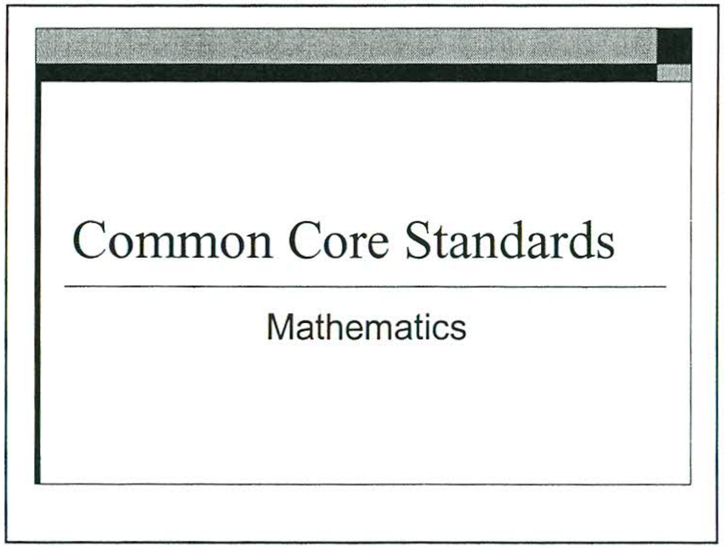
Distribution of Depth-of-Knowledge Levels from Different States Language Arts

Standard	Number of Objs. Under Standard	DOK Levels of Objs.	# of Objs by DOK Levels	% of Objs by DOK Levels
Michigan High School	55	1 2 3 4	0 15 31 9	0 27 57 16
West Virginia Grade 8	32	1 2 3 4	2 12 16 2	6 37 50 6
Alabama Grade 8	4	1 2 3	1 2 1	25 50 25



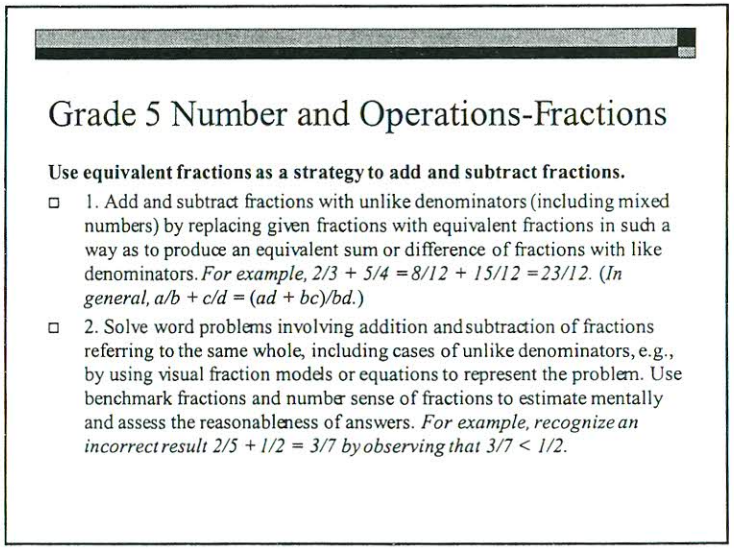
Distribution of Depth-of-Knowledge Levels from Different States Mathematics

	Total Number of Objectives	DOK Level	# of Objs by Level	% within std by Level
Michigan High School	77	1 2 3 4	9 41 24 3	11 53 31 3
West Virginia Grade 8	34.25	1 2 3	4 20 8	12 62 25
Alabama Grade 8	14.75	1 2 3	6 7 1	42 50 7



Common Core Standards

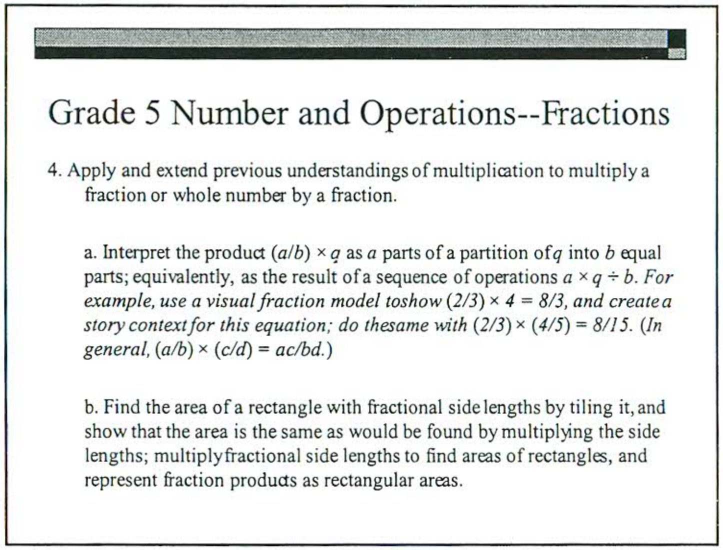
Mathematics



Grade 5 Number and Operations-Fractions

Use equivalent fractions as a strategy to add and subtract fractions.

- □ 1. Add and subtract fractions with unlike denominators (including mixed numbers) by replacing given fractions with equivalent fractions in such a way as to produce an equivalent sum or difference of fractions with like denominators. *For example, 2/3 + 5/4 = 8/12 + 15/12 = 23/12. (In general, a/b + c/d = (ad + bc)/bd.)*
- □ 2. Solve word problems involving addition and subtraction of fractions referring to the same whole, including cases of unlike denominators, e.g., by using visual fraction models or equations to represent the problem. Use benchmark fractions and number sense of fractions to estimate mentally and assess the reasonableness of answers. *For example, recognize an incorrect result 2/5 + 1/2 = 3/7 by observing that 3/7 < 1/2.*

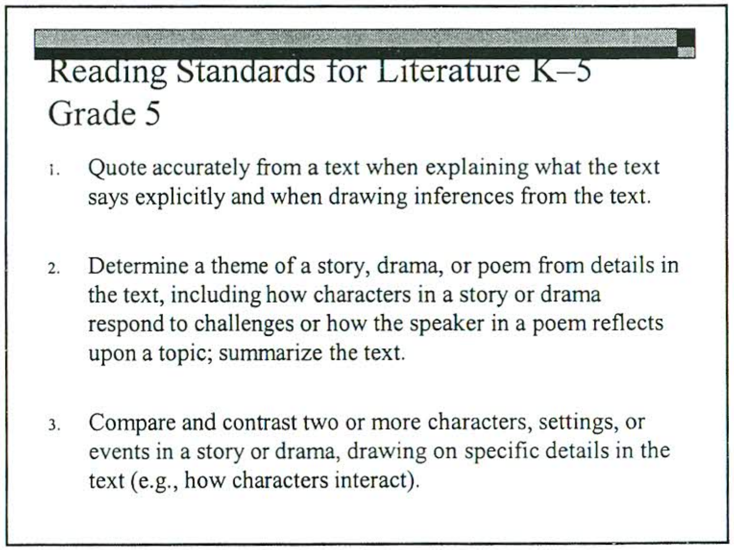


Grade 5 Number and Operations--Fractions

4. Apply and extend previous understandings of multiplication to multiply a fraction or whole number by a fraction.

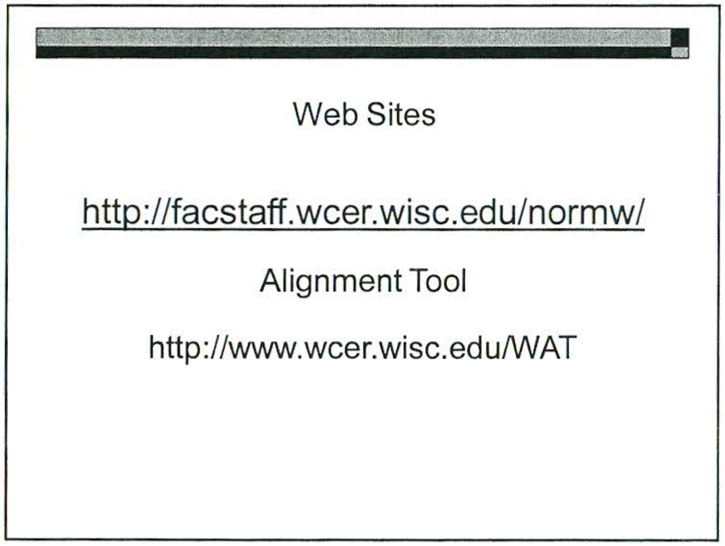
a. Interpret the product $(a/b) \times q$ as a parts of a partition of q into b equal parts; equivalently, as the result of a sequence of operations $a \times q \div b$. *For example, use a visual fraction model toshow* $(2/3) \times 4 = 8/3$*, and createa story contextfor this equation; do thesame with* $(2/3) \times (4/5) = 8/15$*. (In general,* $(a/b) \times (c/d) = ac/bd$*.)*

b. Find the area of a rectangle with fractional side lengths by tiling it, and show that the area is the same as would be found by multiplying the side lengths; multiplyfractional side lengths to find areas of rectangles, and represent fraction products as rectangular areas.



Reading Standards for Literature K–5 Grade 5

1. Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.
2. Determine a theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarize the text.
3. Compare and contrast two or more characters, settings, or events in a story or drama, drawing on specific details in the text (e.g., how characters interact).



Web Sites

http://facstaff.wcer.wisc.edu/normw/

Alignment Tool

http://www.wcer.wisc.edu/WAT

North Carolina Essential Standards for Science

End-of-Grade Grades 5 and 8 Science Assessments End-of-Course Biology Assessment

North Carolina Assessment Specifications

Purpose of the Assessments

- Edition 4 Grades 5 and 8 science assessments and the High School Biology assessments will measure students' proficiency on the *Essential Standards* for Science, adopted by the North Carolina State Board of Education in February 2010.
- NC State Board of Education policy GCS-C-003 (http://sbepolicy.dpi.state.nc.us/) directs schools to use the results from all operational EOC assessments as at least twenty percent (20%) of the student's final course grade.
- Assessment results will be used for school and district accountability under the READY Accountability Model and for federal reporting purposes.

Curriculum Cycle

- February 2010: North Carolina State Board of Education adoption of the Essential Standards for Science
- 2010-2011: Item development for the Next Generation of Assessments, Edition 4
- 2011-2012: Administration of stand-alone field tests of Edition 4 assessments
- 2012-2013: Operational administration of Edition 4 aligned to the Essential Standards for Science

Standards

- The North Carolina *Essential Standards* for Science are posted at: http://www.ncpublicschools.org/acre/standards/new-standards/ .
- Grade 5, grade 8 and High School Biology have a set of content standards.
- Each essential standard has associated clarifying objectives.
- The *Essential Standards* and its clarifying objectives were written using the framework *A Taxonomy for Learning, Teaching, and Assessing—A Revision of Bloom's Taxonomy of Educational Objectives (RBT).*
- The unifying concepts within each set of essential standards provide a context for teaching both science content and scientific-process skill goals.

- The *Essential Standards* for Science for Grades 5 and 8 were written to include content from each of the three branches of science: Life Science (L), Earth Science (E), and Physical Science (P). The unifying concepts for Grades 5 and 8 include:
    - Forces and Motion;
    - Matter: Properties and Change;
    - Energy: Conservation and Transfer;
    - Earth Systems, Structures and Processes;
    - Earth History;
    - Structures and Functions of Living Organisms;
    - Ecosystems;
    - Evolution and Genetics; and
    - Molecular Biology.
- The *Essential Standards* for Biology were written to provide a deeper understanding of the life science content learned throughout Grades K–8. The unifying concepts for Biology include:
    - Structure and Function of Living Organisms,
    - Ecosystems,
    - Evolution and Genetics, and
    - Molecular Biology.

Prioritization of Standards

The North Carolina Department of Public Instruction invited teachers to collaborate and develop recommendations for a prioritization of the standards indicating the relative importance of each standard, the anticipated instructional time, and the appropriateness of the standard for a multiple-choice item format. Subsequently, curriculum and test development staff from the North Carolina Department of Public Instruction met to review the results from the teacher panels and to develop weight distributions across the domains for each grade level. See Tables 1–3.

Table 1: Weight Distributions for Grade 5 Science

Unifying Concept	**Grade 5 Science**
Forces and Motion	13–15%
Matter: Properties and Change	12–14%
Energy: Conservation and Transfer	11–13%
Earth Systems, Structures and Processes	15–17%
Structures and Functions of Living Organisms	14–16%
Ecosystems	14–16%
Evolution and Genetics	13–15%
Total	**100%**

Table 2: Weight Distributions for Grade 8 Science

Unifying Concept	**Grade 8 Science**
Matter: Properties and Change	14–16%
Energy Conservation and Transfer	10–12%
Earth Systems, Structures and Processes	13–15%
Earth History	11–13%
Structure and Function of Living Organisms	19–23%
Ecosystems	9–11%
Evolution and Genetics	11–13%
Molecular Biology	8–10%
Total	**100%**

Table 3: Weight Distributions for Biology

Unifying Concept	**Biology**
Structure and Function of Living Organisms	18–22%
Ecosystems	18–22%
Evolution and Genetics	43–53%
Molecular Biology	15–19%
Total	**100%**

Cognitive Rigor and Item Complexity

Assessment items will be designed, developed, and classified to ensure that the cognitive rigor of the operational test forms align to the cognitive complexity and demands of the North Carolina *Essential Standards* for Science. These items will require students to not only recall information, but also apply concepts and skills and make decisions.

Types of Items

- The Grades 5 and 8 science and High School Biology assessments will consist of four-response-option multiple-choice items and technology-enhanced items (online administration only). All items will be worth one point each.
- The Grade 8 End-of-Grade (EOG) Science assessment requires access to a periodic table of the elements. It can be downloaded at http://www.ncpublicschools.org/accountability/testing/releasedforms.
- The ***NCEXTEND1*** alternate assessments for science will consist of fifteen performance-based, multiple choice items. All items will be worth one point each.
- Appendices A-C show the number of operational items for each clarifying objective administered on assessments. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*.

Delivery Mode

- Grades 5 and 8 science assessments will be designed for an online administration but will also be available in a paper-and-pencil format.

- The High School Biology assessment will be designed for an online administration but will also be available in a paper-and-pencil format.

- ***NCEXTEND1*** is an alternate assessment designed for students with significant cognitive disabilities whose IEP specifies an assessment aligned to the Extended Content Standards and based on alternate academic achievement standards. The ***NCEXTEND1*** assessments will be designed for paper/pencil administrations with online data entry by the assessor. The Extended Content Standards may be reviewed at http://www.ncpublicschools.org/acre/standards/extended/.

- End-of-grade and end-of-course assessments are only provided in English. Native language translation versions are not available.

Appendix A
Grade 5 Science
Number of Operational Items by Clarifying Objective

The following table shows the number of operational items for each clarifying objective. Note that future coverage of objectives could vary within the constraints of the content category weights in *Tables 1-3*. Some objectives not designated with tested items (i.e., "–") may be a prerequisite standard, may be tested within the context of another standard or may be included as an embedded field test item. The objectives for may be reviewed at http://www.ncpublicschools.org/acre/standards/new-standards/.

Grade 5 Science	Number of Operational Items Per Objective*
Forces and Motion 5.P.1.1	3-4
5.P.1.2	3-4
5.P.1.3	0-1
5.P.1.4	0-1
Matter: Properties and Change 5.P.2.1	4-6
5.P.2.2	1
5.P.2.3	2-3
Energy: Conservation and Transfer 5.P.3.1	2-4
5.P.3.2	2-3
Earth Systems, Structures and Processes 5.E.1.1	2
5.E.1.2	3
5.E.1.3	5
Structures and Functions of Living Organisms 5.L.1.1	4-5
5.L.1.2	5-6
Ecosystems 5.L.2.1	1-2
5.L.2.2	3-4
5.L.2.3	5
Evolution and Genetics 5.L.3.1	2-4
5.L.3.2	4-5

* Some objectives not designated with tested items (i.e., "–") may be a prerequisite standard, may be tested within the context of another standard or may be included as an embedded field test item.

Appendix B
Grade 8 Science
Number of Operational Items by Clarifying Objective

The following table shows the number of operational items for each clarifying objective. Note that future coverage of objectives could vary within the constraints of the content category weights in *Tables 1-3*. Some objectives not designated with tested items (i.e., "–") may be a prerequisite objective, may be tested within the context of another objective or may be included as an embedded field test item. The objectives for may be reviewed at http://www.ncpublicschools.org/acre/standards/new-standards/.

Grade 8 Science	Number of Operational Items Per Objective*
Matter: Properties and Change 8.P.1.1	2
8.P.1.2	3
8.P.1.3	2
8.P.1.4	3
Energy: Conservation and Transfer 8.P.2.1	3
8.P.2.2	3
Earth Systems, Structures and Processes 8.E.1.1	3
8.E.1.2	2
8.E.1.3	2-3
8.E.1.4	0-1
Earth History 8.E.2.1	3
8.E.2.2	4
Structures and Functions of Living Organisms 8.L.1.1	3-4
8.L.1.2	1-2
8.L.2.1	6
Ecosystems 8.L.3.1	1
8.L.3.2	2-3
8.L.3.3	2-3
Evolution and Genetics 8.L.4.1	4
8.L.4.2	4
Molecular Biology 8.L.5.1	2
8.L.5.2	2

* Some objectives not designated with tested items (i.e., "–") may be a prerequisite standard, may be tested within the context of another standard or may be included as an embedded field test item.

Appendix C
Biology
Number of Operational Items by Clarifying Objective

The following table shows the number of operational items for each clarifying objective. Note that future coverage of objectives could vary within the constraints of the content category weights in *Tables 1-3*. Some objectives not designated with tested items (i.e., "–") may be a prerequisite standard, may be tested within the context of another standard or may be included as an embedded field test item. The objectives for may be reviewed at http://www.ncpublicschools.org/acre/standards/new-standards/.

Biology	Number of Operational Items Per Objective*
Structure and Functions of Living Organisms 1.1.1	2
1.1.2	1-2
1.1.3	3
1.2.1	2
1.2.2	3
1.2.3	1
Ecosystems 2.1.1	1
2.1.2	1-2
2.1.3	2-3
2.1.4	1
2.2.1	2-3
2.2.2	3-4
Evolution and Genetics 3.1.1	3
3.1.2	1-2
3.1.3	1
3.2.1	1-2
3.2.2	1
3.2.3	3-4
3.3.1	3
3.3.2	2
3.3.3	–
3.4.1	2
3.4.2	3
3.4.3	1
3.5.1	2
3.5.2	2

Biology	Number of Operational Items Per Objective*
Molecular Biology 4.1.1	2
4.1.2	2
4.1.3	1
4.2.1	2
4.2.2	2

* Some objectives not designated with tested items (i.e., “–”) may be a prerequisite standard, may be tested within the context of another standard or may be included as an embedded field test item.

APPENDIX 3-C PLAIN ENGLISH TRAINING

Hope Lung

Subject:	Plain English Strategies Workshop
Location:	Room 150
Start:	Thu 4/28/2011 8:30 AM
End:	Thu 4/28/2011 4:00 PM
Recurrence:	(none)
Meeting Status:	Meeting organizer
Organizer:	Audrey Martin-McCoy

As previously announced, the plain English strategies workshop will be held on April 28. Attached you will find a draft agenda for the day.

The workshop will be held in room 150 of the Education Building, 8:30 am - 4:00 pm.

Audrey

Audrey Martin-McCoy, Ph.D.
Education Testing/Accountability Consultant
Testing Policy and Operations Section/Division of Accountability Services
North Carolina Department of Public Instruction
6314 Mail Service Center
Raleigh, NC 27699-6314

All e-mail correspondence to and from this address is subject to the North Carolina Public Records Law, which may result in monitoring and disclosure to third parties, including law enforcement.

>>> Audrey Martin-McCoy 03/16/11 11:22 AM >>>
A workshop will be offered in an attempt to extend and refine our knowledge and use of plain English language practices in test construction. The workshop will be facilitated by Dr. Edynn Sato. Edynn is Director of Research and English Learner Assessment with the Assessment and Standard Development Services Program at West Ed. She is also the Director of Special Populations at the Assessment and Accountability Comprehensive Center at West Ed.

The training workshop will focus on the latest research in the area of plain English practices and examine its use in our current training used for our item writers/editors and in released state test forms. In sum, this is an opportunity to build and/or re-evaluate how we go about developing plain English test items. Follow up conference calls will be scheduled after the workshop to foster continued understanding of concepts discussed.

The workshop will be held on April 28, 2011, from 8:30 am to 4:00 pm in room 150 at the Education Building. Lunch is on your own from 11:30 am to 12:30 pm. A draft agenda will be sent within the next two weeks. Personnel from DPI ESL, Accountability, and NCSU - TOPS will be invited to attend.

Please save this date and time. Let me know if you have questions.

Audrey

Assessment and Accountability
Comprehensive Center
AACC [illegible]

ARCC
Appalachia Regional
Comprehensive Center | at EDVANTIA

WORKSHOP

Plain English Strategies:
Research, Theory, and Implications for Assessment Development

Agenda

April 28, 2011

Workshop Objective: To provide participants with information about plain English strategies that will inform and support the effective application of these practices in the state's test item development process.

8:30 — 8:45 am	Welcome and Introductions *Shirley Carraway, ARCC- NC Liaison* *Audrey Martin-McCoy, NCDPI*
8:45 — 10:00 am	Introduction to Plain English: Research, Theory, and the Accessibility Context *Edynn Sato, AACC- WestEd* Director *Rachel Lagunoff, AACC – WestEd*
10:00 — 10:15 am	Break
10:15 — 11:30 am	Introduction to Plain English: Research, Theory, and the Accessibility Context (Continued) *Edynn Sato and Rachel Lagunoff*
11:30 am — 12:30 pm	Lunch
12:30 pm — 3:30 pm	Application of Plain English Strategies: Implications for Item Development and Related Training *Edynn Sato and Rachel Lagunoff*
3:30 pm — 4:00 pm	Discussion of Possible Next Steps *NCDPI Staff*

Plain English Strategies
Application of Plain English Strategies: Implications for Item Development

WORKSHOP

Examples of applying research-based Plain English strategies to test items

Research Findings	Practical Recommendations	Examples
Words that are short (simple morphologically) tend to be more familiar and, therefore, easier.	Use simple words; use high-frequency words; only use compound words and words with prefixes or suffixes that are likely to be familiar. Exception: words that are directly related to content the student is expected to know	Change *utilize* to *use* Even though *chair* is EDL 2 and *man* is EDL 1, *chairman* is EDL 7, so may not be familiar; both *base* and *baseball* are EDL 3, so likely to be equally familiar. *Proper* is EDL 5, but *improper* is EDL 8, so *im-* is likely to be an unfamiliar prefix; *happy* is EDL 1, and *unhappy* is EDL 2, so *un-* is likely to be a familiar prefix.
Passages with words that are familiar (simple semantically) are easier to understand.	Use familiar words. Omit or define words with double meanings or colloquialisms.	Change *go off* to *leave*, *explode*, or *start to ring* Even seemingly simple words can have multiple meanings, e.g., *fine* (feeling, weather, hair or line, penalty, etc.). Even seemingly simple words can have colloquial or idiomatic uses, e.g., *hop in, blow up, get it.*

Research Findings	Practical Recommendations	Examples
Longer sentences tend to be more complex syntactically and, therefore, more difficult to comprehend.	Retain Subject-Verb-Object structure for statements. Begin questions with question words. Avoid clauses and phrases.	Change *At which of the following times* to *When* Change *A report that contains 64 papers* to *He needs 64 sheets of paper for each report*
Long items tend to pose greater difficulty.	Remove unnecessary expository material.	Change *The weights of four different bookbags are recorded in the chart above. According to the chart, which bookbag is the heaviest?* to *Look at the chart below. Which bookbag weighs the MOST?*
Complex sentences tend to be more difficult than simple or compound sentences.	Keep to the present tense, use active voice, avoid the conditional mode, and avoid starting with sentence clauses.	Change *The weights of 3 objects were compared* to *Sandra compared the weights of 3 objects* Change *If Lee delivers x newspapers* to *Lee delivers x newspapers*

Suggested Strategies for Ensuring Maximum Test Item Readability and Comprehensibility

Strategy	Example
Avoid irregularly spelled words	Words such as *trough* or *feign* may be difficult to read
Use generic terms and familiar proper names with simple spelling	Use *tree* instead of *pine* or *oak*; use *Jeff* instead of *Geoffrey* and *Ellen* instead of *Eleanor*
Avoid multiple terms for the same concept	Do not use both *children* and *kids* in an item or a set of items; in items based on a reading passage, use the same term as in the passage
Make sure all noun-pronoun relationships are clear	In the stem *Scientists think bears are most dangerous when they are*, replace *they* with *the bears*
Put important context first	When time and setting are important to the sentence, place them at the beginning of the sentence; put the location of information in a passage at the beginning of the stem (e.g., *In the 1800s*; *In the second paragraph*)
When possible, write closed stems that end with a question mark	If the answer choices are complete sentences, a closed stem is usually possible; if words are repeated at the beginning of answer choices, an open stem may be preferable

References

Abedi, J. et al. (2005). *Language Accommodations for English Learners in Large-Scale Assessments: Bilingual Dictionaries and Linguistic Modification.* (CSE Report 666). Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Brown, P.J. (1999). *Findings of the 1999 plain language field test.* University of Delaware, Newark, DE: Delaware Education Research and Development Center.

Gaster, L., & Clark, C. (1995). *A guide to providing alternate formats.* West Columbia, SC: Center for Rehabilitation Technology Services. (ERIC Document No. ED 405689)

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved April 25, 2011, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Evaluating Items for Plain English: Sample Items

SAMPLE A
Reading Comprehension
Grade 3
Selection: *Hamish McBean and His Sheep*

2. Which words from the selection ***best*** help the reader picture the setting?

SAMPLE B
Reading Comprehension
Grade 3
Selection: *Lots of Kids Live Here*

9. Which completes the chart?

kids	young goats
does	female goats
bucks	?

A old goats

B male goats

C mother goats

D newborn goats

SAMPLE C
Reading Comprehension
Grade 5
Selection: *Seneca Oil and Early America*

18. According to the selection, what was one effect of the Senecas' mixing petroleum with paint, particularly during a time of war?

SAMPLE D
Reading Comprehension
Grade 8
Selection: *Here's to Ears*

15. Why is impaired hearing called "auditory isolation"?

A It has a single cause.

B It does not involve other body systems.

C It cuts people off from their environment.

D It keeps sound waves from reaching the auditory nerve.

SAMPLE E
Mathematics—Calculator Inactive
Grade 3

2. There are 20 seeds in a package. If 5 seeds are put in each flower pot, how many flower pots are needed to plant all of the seeds?

SAMPLE F
Mathematics—Calculator Active
Grade 4

17. The bread truck makes deliveries to a store 3 days each week. Each delivery has 45 loaves of bread. Which expression could be used to determine the number of loaves of bread delivered in 5 weeks?

SAMPLE G
Mathematics—Calculator Active
Grade 6

29. Marsha wants to find out how other students at her school get to school each day. Which of the following groups, if surveyed, would give her the ***most accurate*** sample of the student body?

SAMPLE H
Algebra I

44. A computer is purchased for $1,200 and depreciates at $140 per year. Which linear equation represents the value, V, of the computer at the end of t years?

Language for Achievement—A Framework for Academic English Language

Handout description:

The *Language for Achievement Framework* (page 2) is theory and research based, and aspects of the framework have been used in the evaluation and development of English language proficiency (ELP) standards and assessments in a number of states, as well as in examinations of linkage or correspondence between state ELP and academic content standards (i.e., to identify aspects of English language needed to facilitate student access to and meaningful engagement with academic content).

This handout also includes a *taxonomy* (page 3) that focuses on academic language functions (as opposed to, for example, social language and linguistic skills) that is intended to serve for the language domain the role that Bloom's taxonomy, for example, serves for the cognitive domain—Bloom's taxonomy serves as a classification system for thinking behaviors that are important to the learning process (Forehand, 2005; Hancock, 1994; Kreitzer & Madaus, 1994; Seddon, 1978). The taxonomy provides a structure for arranging content learning objectives according to the academic language necessary for students to meet a content objective, or set of related objectives. The taxonomy can inform the development of *language progressions* which place the academic language skills and knowledge of the taxonomy on a developmental continuum, reflecting a progression from the most basic and foundational English language skills and knowledge to the most advanced and developed language skills and knowledge relevant to accessing and achieving rigorous academic content. Therefore, the taxonomy has important implications for instructional practices that can support the language related to academic achievement not only of EL students but of *all* students working to meet more rigorous and higher academic expectations.

Also associated with the framework are rubrics related to language complexity (pages 4-6). The language demands represented in the framework (i.e., academic vocabulary and grammar, functions, spoken and written text, classroom discourse) interact with language complexity.

Information presented in this handout is intended for the following purposes:

- to help analyze the content and language in standards, assessment tasks, and instructional materials;
- to help make explicit the expectations (cognitive, language) of students;
- to help inform instructional planning and practice so that they are intentional and appropriate in supporting students' progress (cognitive, linguistic) toward proficiency and achievement; and
- to serve as a tool for cross-disciplinary discussions related to appropriately addressing the content and language needs of English learner students and facilitating their achievement in school.

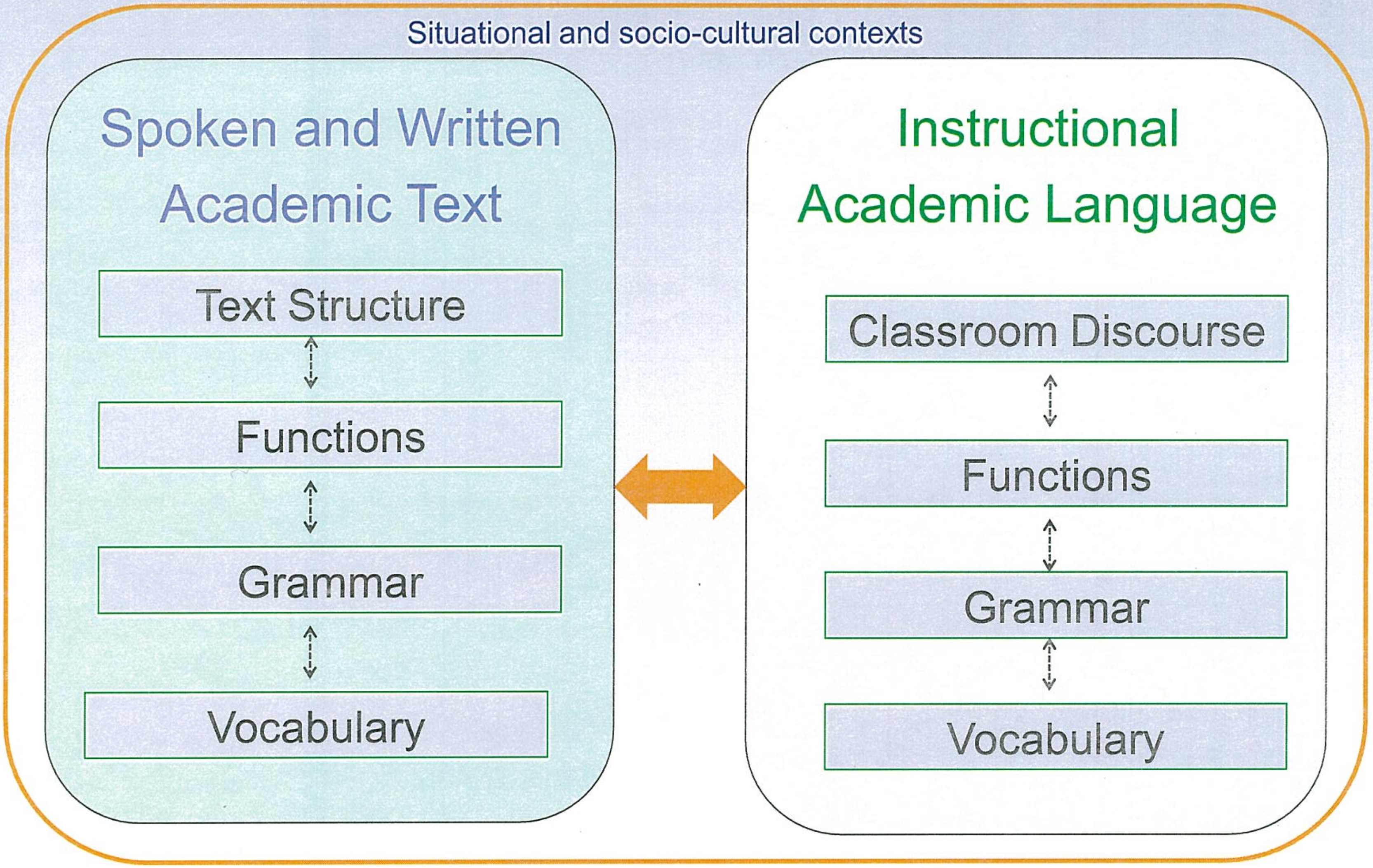
For more information, please contact Dr. Edynn Sato at WestEd (esato@wested.org; 415-615-3226).

Notes:

- For use and distribution of information contained in this packet, please contact Dr. Edynn Sato (contact information listed above).
- The information in this handout was originally developed for research purposes. The information is not necessarily comprehensive (e.g., list of functions).

Draft: October 2010

Language for Achievement: Overview



Situational and socio-cultural contexts

Spoken and Written Academic Text

Text Structure

Functions

Grammar

Vocabulary

Instructional Academic Language

Classroom Discourse

Functions

Grammar

Vocabulary

Additional considerations include: receptive (listening, reading) and productive (speaking, writing) language; language complexity

Sato & Lagunoff (2010)

Language for Achievement—Taxonomy: Academic English Language Functions

Academic English Language Function		*Operational Definition—The language needed to **engage with and achieve** in the content (standard or item) consists of the use of:*
A	Identification	a word or phrase to name an object, action, event, idea, fact, problem, need, or process.
	Labeling	a word or phrase to name an object, action, event, or idea.
	Enumeration	words or phrases to name distinct objects, actions, events, or ideas in a series, set, or in steps.
B	Classification	words, phrases, or sentences to assign/associate an object, action, event, or idea to the category or type to which it belongs.
	Sequencing	words, phrases, or sentences to express the order of information (e.g., a series of objects, actions, events, ideas). Discourse markers include adverbials such as *first*, *next*, *then*, *finally*.
	Organization	words, phrases, or sentences to express relationships between/among objects, actions, events, or ideas, or the structure or arrangement of information. Discourse markers include coordinating conjunctions such as *and*, *but*, *yet*, *or*, and adverbials such as *first*, *next*, *then*, *finally*.
C	Comparison/ Contrast	words, phrases, or sentences to express similarities and/or differences, or to distinguish between two or more objects, actions, events, or ideas. Discourse markers include coordinating conjunctions *and*, *but*, *yet*, *or*, and adverbials such as *similarly*, *likewise*, *in contrast*, *instead*, *despite this*.
D	Inquiring	words, phrases, or sentences to solicit information (e.g., *yes-no* questions, *wh*-questions, statements used as questions).
E	Description	word, phrase, or sentence to express or observe the attributes or properties of an object, action, event, idea, or solution.
F	Definition	word, phrase, or sentence to express the meaning of a given word, phrase, or expression.
G	Explanation	phrases or sentences to express the rationale, reasons, causes, or relationships related to one or more actions, events, ideas, or processes. Discourse markers include coordinating conjunctions *so*, *for*, and adverbials such as *therefore*, *as a result*, *for that reason*.
H	Retelling	phrases or sentences to relate or repeat information. Discourse markers include coordinating conjunctions such as *and*, *but*, and adverbials such as *first*, *next*, *then*, *finally*.
	Summarization	phrases or sentences to express important facts or ideas and relevant details about one or more objects, actions, events, ideas, or processes. Discourse structures include: beginning with an introductory sentence that specifies purpose or topic.
I	Interpretation	phrases, sentences, or symbols to express understanding of the intended or alternate meaning of information.
J	Analyzing	phrases or sentences to indicate parts of a whole and/or the relationship between/among parts of an action, event, idea, or process. Relationship verbs such as *contain*, *entail*, *consist of*, partitives such as *a part of*, *a segment of*, and quantifiers such as *some*, *a good number of*, *almost all*, *a few*, *hardly any* often are used.
K	Generalization	phrases or sentences to express an opinion, principle, trend, or conclusion that is based on facts, statistics, or other information, and/or to extend that opinion/principle/etc. to other relevant situations/contexts/etc.
	Inferring	words, phrases, or sentences to express understanding of implied/implicit based on available information. Discourse markers include inferential logical connectors such as *although*, *while*, *thus*, *therefore*.
	Prediction	words, phrases, or sentences to express an idea or notion about a future action or event based on available information. Discourse markers include adverbials such as *maybe*, *perhaps*, *obviously*, *evidently*.
	Hypothesizing	phrases or sentences to express an idea/expectation or possible outcome based on available information. Discourse markers include adverbials such as *generally*, *typically*, *obviously*, *evidently*.
L	Argumentation	phrases or sentences to present a point of view with the intent of communicating or supporting a particular position or conviction. Discourse structures include expressions such as *in my opinion*, *it seems to me*, and adverbials such as *since*, *because*, *although*, *however*.
	Persuasion	phrases or sentences to present ideas, opinions, and/or principles with the intent of creating agreement around or convincing others of a position or conviction. Discourse markers include expressions such as *in my opinion*, *it seems to me*, and adverbials such as *since*, *because*, *although*, *however*.
	Negotiation	phrases or sentences to engage in a discussion with the purpose of creating mutual agreement from two or more different points of view.
M	Synthesizing	phrases or sentences to express, describe, or explain relationships among two or more ideas. Relationship verbs such as *contain*, *entail*, *consist of*, partitives such as *a part of*, *a segment of*, and quantifiers such as *some*, *a good number of*, *almost all*, *a few*, *hardly any* often are used.
N	Critiquing	phrases or sentences to express a focused review or analysis of an object, action, event, idea, or text.
O	Evaluation	phrases or sentences to express a judgment about the meaning, importance, or significance of an action, event, idea, or text.
P	Symbolization & Representation	symbols, numerals, and letters, to represent meaning within a conventional context (e.g., ÷, -, CO_2, >, Δ, π, cos, $y=3x+4$, $c^2=a^2+b^2$, $h/2(b_1+b_2)$, *cat* vs. cat).
Z	No Academic Language Function	Item or standard does not contain *any* academic language functions; may contain linguistic skills (e.g., phonemic awareness, syllabication).

Note: This taxonomy focuses on academic language functions and does not address the identification or definition of linguistic skills (e.g., phonology, morphology).

Language for Achievement—Language Complexity

The *Language for Achievement* language demands (i.e., academic vocabulary and grammar, functions, spoken and written text, classroom discourse) interact with language complexity. Language complexity, as used in this framework, is defined below.

Vocabulary and Grammar

Lower Complexity	Higher Complexity
• Semantically simple words and phrases • Common, high-frequency words and phrases • Simple, high-frequency morphological structures (e.g., common affixes, common compound words)	• Semantically complex words and phrases (e.g., multiple-meaning words, idioms, figurative language) • Specialized or technical words and phrases • Complex, higher level morphological structures (e.g., higher level affixes and compound words)
• Short, simple sentences with limited modifying words or phrases • SVO sentence structure; simple verb and noun phrase constructions • Simple, familiar modals (e.g., *can*) • Simple *wh-* and *yes/no* questions • Direct (quoted) speech • Verbs in present tense, simple past tense, and future with *going to* and *will* • Simple, high-frequency noun, adjective, and adverb constructions	• Compound and complex sentences; longer sentences with modifying words, phrases, and clauses • High level phrase and clause constructions (e.g., passive constructions, gerunds and infinitives as subjects and objects, conditional constructions) • Multiple-meaning modals, past forms of modals • Complex *wh-* and *yes/no* question constructions, tag questions • Indirect (reported) speech • Present, past, and future progressive and perfect verb structures • Complex, higher level noun, adjective, and adverb constructions

Functions

Lower Complexity	Higher Complexity
• Length ranges from a word to paragraphs • No/little variation in words and/or phrases in sentences/paragraphs; consistent use of language • Repetition of key words/phrases/sentences *reinforces* information • Language is used to present critical/central details • No/little abstraction; language reflects more literal/concrete information; illustrative language is used; language is used to define/explain abstract information • Graphics and/or relevant text features reinforce critical information/details • Mostly common/familiar words/phrases; no/few uncommon words/phrases, compound words, gerunds, figurative language, and/or idioms • Language is organized/structured • Mostly simple sentence construction • No/little passive voice • Little variation in tense • Mostly one idea/detail per sentence • Mostly familiar construction (e.g., 's for possessive; s and es for plural) • Mostly familiar text features (e.g., bulleted lists, bold face)	• Length ranges from a word to paragraphs • Some variation in words and/or phrases in sentences/paragraphs • Repetition of key words/phrases/sentences *introduces new* or *extends* information • Language is used to present critical/central details, but non-essential detail also is presented • Some abstraction; language *may or may not* be used to define/explain abstract information; illustrative language *may or may not* be used; technical words/phrases are used • Graphics and/or relevant text features *may or may not* reinforce critical information/details • Some common/familiar words/phrases; some uncommon words/phrases, compound words, gerunds, figurative language, and/or idioms • Language *may or may not* be organized/structured • Varied sentence construction, including complex sentence construction • Some passive voice • Variation in tense • Multiple ideas/details per sentence • Some less familiar/irregular construction • Some less familiar text features (e.g., pronunciation keys, text boxes)

Spoken and Written Texts

Lower Complexity	Higher Complexity
• Short texts, or longer texts chunked into short sections (words, phrases, single sentences, short paragraphs) • No or little variation of words/phrases in sentences/paragraphs • Repetition of key words/phrases reinforces information • One idea/detail per sentence; only critical/central ideas included • No or little abstraction; mostly literal/concrete information; abstract information is defined or explained • Visual aids, graphics, and/or text features reinforce critical information/details • Common text features (e.g. bulleted lists, boldface font)	• Long texts (long lists of words/phrases, a series of sentences, long paragraphs, multiple-paragraph texts) • Variation of words/phrases in sentences/paragraphs • Repetition of key words/phrases introduces new information or extends information • Multiple ideas/details per sentence; non-essential ideas included • Some or much abstraction that is not explicitly defined or explained • Visual aids, graphics, and/or text features may not reinforce critical information/details • Higher level text features (e.g., pronunciation keys, text boxes)

Classroom Discourse

Lower Complexity	Higher Complexity
• Semantically simple words and phrases • Common, high-frequency words and phrases • Simple, high-frequency morphological structures (e.g., common affixes, common compound words)	• Semantically complex words and phrases (e.g., multiple-meaning words, idioms, figurative language) • Specialized or technical words and phrases • Complex, higher level morphological structures (e.g., higher level affixes and compound words)
• Short, simple sentences with limited modifying words or phrases • SVO sentence structure; simple verb and noun phrase constructions • Simple, familiar modals (e.g., can) • Simple wh- and yes/no questions • Direct (quoted) speech • Verbs in present tense, simple past tense, and future with going to and will • Simple, high-frequency noun, adjective, and adverb constructions Note: To the extent that spoken "texts" (planned, connected utterances) are used in classroom discourse, elements of lower complexity spoken text, as defined previously, apply here.	• Compound and complex sentences; longer sentences with modifying words, phrases, and clauses • High level phrase and clause constructions (e.g., passive constructions, gerunds and infinitives as subjects and objects, conditional constructions) • Multiple-meaning modals, past forms of modals • Complex wh- and yes/no question constructions, tag questions • Indirect (reported) speech • Present, past, and future progressive and perfect verb structures • Complex, higher level noun, adjective, and adverb constructions Note: To the extent that spoken "texts" (planned, connected utterances) are used in classroom discourse, elements of higher complexity spoken text, as defined previously, apply here.

Definition from the *Framework for High-Quality ELP Standards and Assessments* (AACC, 2009):
Academic language, broadly defined, includes the language students need to meaningfully engage with academic content within the academic context. This should *not* be interpreted to suggest that separate word lists and/or definitions of content-related language should be developed for each academic subject. Rather, academic language includes the words, grammatical structures, and discourse markers needed in, for example, describing, sequencing, summarizing, and evaluating — these are language demands (skills, knowledge) that facilitate student access to and engagement with grade-level academic content. These academic language demands are different from cognitive demands (e.g., per Bloom's taxonomy). Although there may not be just one accepted definition of academic language, there are a good number of resources available that address the issue of academic language and may be considered in the development of state ELP standards and assessments. For example: Aguirre-Munoz, Parks, Benner, Amabisca, & Boscardin, 2006; Bailey, 2007; Bailey, Butler, & Sato, 2007; Butler, Bailey, Stevens, Huang, & Lord, 2004; Chamot & O'Malley, 1994; Cummins, 1980; Cummins, 2005; Halliday, 1994; Sato, 2007; Scarcella & Zimmerman, 1998; Schleppegrell, 2001.

For a free download of the *Framework for High-Quality ELP Standards and Assessments*, go to http://www.aacompcenter.org/cs/aacc/print/htdocs/aacc/resources_sp.htm.

From: http://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=92

Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets

Edynn Sato, Stanley Rabinowitz, Carole Gallagher, and Chun-Wei Huang

REL West's study on middle school math assessment accommodations found that simplifying the language—or linguistic modification—on standardized math test items made it easier for English Language learners to focus on and grasp math concepts, and thus was a more accurate assessment of their math skills.

The results contribute to the body of knowledge informing assessment practices and accommodations appropriate for English language learner students.

The study examined students' performance on two sets of math items—both the originally worded items and those that had been modified. Researchers analyzed results from three subgroups of students—English learners (EL), non-English language arts proficient (NEP), and English language arts proficient (EP) students.

Key results include:

- Linguistically modifying the language of mathematics test items did not change the math knowledge being assessed.
- The effect of linguistic modification on students' math performance varied between the three student subgroups. The results also varied depending on how scores were calculated for each student.
- For each of the four scoring approaches analyzed, the effect of linguistic modification was greatest for EL students, followed by NEP and EP students.

Note: The following pages are excerpted from the full report which is available at: http://ies.ed.gov/ncee/edlabs/projects/project.asp?ProjectID=92

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

REL WEST
Regional Educational Laboratory
At WestEd

NCEE 2009–4079

U.S. DEPARTMENT OF EDUCATION

Accommodations for English Language Learner Students: the Effect of Linguistic Modification of Math Test Item Sets

Final Report

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

REL WEST
Regional Educational Laboratory
At WestEd

REL WEST
Regional Educational Laboratory
At WestEd

Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets

June 2010

Authors:

Edynn Sato, Principal Investigator
WestEd

Stanley Rabinowitz, Principal Investigator
WestEd

Carole Gallagher, Senior Research Associate
WestEd

Chun-Wei Huang, Senior Research Analyst
WestEd

Project Officer:

Ok-Choon Park
Institute of Education Sciences

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

NCEE 2009-4079
U.S. Department of Education

Contents

Tables

Figures

Appendix D. Guide for developing a linguistically modified assessment

> [This guide was followed to linguistically modify the items used in this study. Experts in mathematics, linguistics, measurement, curriculum and instruction, and the English language learner student population were convened to discuss linguistic modification strategies and their application. These experts possessed advanced degrees (such as an M.A. or Ph.D.), had classroom teaching experience, and assessment development experience. The selection of items, the linguistic modification of items, and the creation of the item sets used in this study occurred over the equivalent of a period of approximately three weeks and followed generally accepted item development procedures including verification of content alignment, appropriateness for the student population, and freedom from bias and sensitivity issues.]

For all students, access to test content is necessary to ensure the validity of assessment results.[35] Valid assessments are especially critical if results are used to inform classroom instruction or for accountability purposes. When access is constrained in some way (for example, linguistically or cognitively), students may be prevented from fully demonstrating what they know and can do, and the test score may underestimate or misrepresent students' achievement. To assess English language learner students' knowledge of academic content, it is critical to determine whether their academic performance reflects their understanding of the targeted content or their lack of English language proficiency. There is an interaction between how assessed content is presented in test items and what English language learner students need in order to access that content. This interaction affects the validity of the assessment results and the interpretation of those results.

Linguistic modification of test items is an approach for addressing the particular access needs of English language learner students so that test performance is attributable less to English language proficiency and more to knowledge and skills related to the tested content. The approach outlined below is intended to help researchers in this study consider key characteristics of the content and the student population as they develop linguistically modified test items. The three steps in this process are:

- Define the domain and constructs of tested content.
- Define the English language learner population that will be tested.
- Apply and evaluate linguistic modification strategies to test items.

[35] Information in this appendix is drawn from Sato (2008).

Step 1: define the domain and constructs

Articulate the purpose of the assessment. Consider the range of ways the assessment results will be used and the intended outcomes of testing.

Recommended specialists for this step

Given the purpose of the assessment and the population assessed, this step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge about the English language learner student population).

Purpose

The assessment results will be used for the following purpose(s):

____

Assessed academic content domain

The assessment will measure students' knowledge of:

____

____

Considerations
Is this test appropriate for the target content domain? To what degree do content domain characteristics align with the intended purpose of this assessment?

Assessed constructs—content and skills

More specifically, the assessment will measure the following constructs (content and skills) related to the domain:

____

____

Considerations
Do the content and skills assessed in the set of linguistically modified test items reflect the intended breadth, depth, and range of complexity of the assessed domain? Are the verbs used in the state standards statements specific enough to guide assessment development (for example, "identify," "describe," "compare" vs. the more vague "know," "understand")? If the latter, how are students expected to demonstrate their knowledge and skills?

Content-related language—language demands

The following language demands are associated with the content and skills that will be assessed (see tables E1 and E2 in appendix E for a list of language demands—linguistic skills and academic language functions):

Considerations
Have students' linguistic skills and academic language functions both been considered?
Is the range of language demands in the linguistically modified items consistent with the breadth, depth, and range of complexity of the assessed content domain?

Content-related language—specific vocabulary and terminology

The following vocabulary and terminology are specific to the grade-level content assessed; therefore, they should not be linguistically modified:

Considerations
Is the vocabulary and terminology identified consistent with the intent of the grade-level content standards?

Step 2: define the population and student subgroups

Articulate the key characteristics and access needs of the English language learner student population. Since this group of students is especially diverse and heterogeneous, it may be necessary to identify key subgroups of students within the state.

Recommended specialists for this step

Given the purpose of the assessment and the population assessed, this step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge about English language learner students).

Student population

The target English language learner population can be characterized as follows
(see appendix E for a description of English language learner students):

____

____

Student access needs

Document the access needs of the target English language learner student population, taking into account characteristics such as:

Context
What topics, themes, locations, situations, illustrations, and such are familiar to these students?

____

Words, phrases, sentences
What written vocabulary is familiar to these students? What phrases are familiar to these students? What sentence structures are familiar to these students? What tenses (for example, present, past) and constructions (for example, plural _s, possessive _'s) are familiar to these students? What proper nouns are familiar to students as a result of their classroom reading?

____

____

Format/Style
With what formats/styles are these students familiar (for example, bulleted lists, text boxes, underlining for emphasis)? How is information typically presented to these students during instruction?

____

____

Step 3: apply and evaluate linguistic modification strategies

Determine which content and item types lend themselves to linguistic modification. Then develop and evaluate each test item according to the following dimensions: context, graphics, vocabulary/wording, sentence structure, and format/style (see table D1 for linguistic modification guidelines and strategies for each dimension).

Recommended specialists for this step

This step is best conducted by a team that includes content specialists, assessment specialists, curriculum and instruction specialists, English language development specialists, and population specialists (that is, individuals with specialized knowledge of the English language learner population).

Categorize target content and item types

Sort content/test items into one of the following three categories of eligibility for linguistic modification. Within each eligibility category, group content standards and test items by content strand (for example, measurement or algebra for mathematics).

- Definitely eligible.
- Definitely not eligible.
- Possibly eligible.

Considerations
A test item's appropriateness for linguistic modification is associated with the quantity of construct-irrelevant language in that test item; the greater the quantity of construct-irrelevant language, the greater the likelihood that the item can be linguistically modified effectively for English language learner students. There also is a greater likelihood that construct-irrelevant language can be linguistically modified without significantly changing the assessed construct (for example, mathematics achievement).

Apply linguistic modification guidelines and strategies

For content/items that are eligible and possibly eligible for linguistic modification, systematically apply the relevant guidelines and strategies presented in table D1 (that is, context, graphics, vocabulary/wording, sentence structure, format/style).

Considerations
The team of specialists who are linguistically modifying items need specialized training to ensure that they are appropriately applying linguistic modification guidelines. It is important to ensure the guidelines are accurately and consistently applied during item development and that the intended construct, cognitive complexity, and language demands specified in the grade-level standards have not been significantly altered.

Follow checklist for evaluating the linguistically modified items

For each item, verify that:

- The construct being tested has not changed.
- The cognitive complexity of the item is appropriate.
- The following elements in the linguistically modified item maximize English language learner students' linguistic access:
  - Context.

  - Graphics.
  - Vocabulary/wording.
  - Sentence structure.
  - Format/style.

Methods used to verify that the test item has been appropriately linguistically modified include:

- Expert verification (for example, by a technical advisory committee, content and bias review committee, or independent external reviewer) that the construct has not changed and that the cognitive complexity of the item is appropriate.

- Statistical analyses (for example, analysis of variance, differential item functioning analysis, or factor analysis).

- Cognitive interviews.

Table D1. Linguistic modification guidelines and strategies

Desirable characteristics	Notes on approaches and criteria
Item context	
• Familiar to students. • No cultural or linguistic bias. • Minimal construct (no irrelevant words or phrases).	• The context situates the problem (and may include description of relationship or interaction between location and time). • In the body of the report, context is often described in relation to its complexity and as part of biased or construct-irrelevant information that should be pruned out. Recommendations: o Remove passive voice construction in original item. o Remove past tense and conditional in original item. o Break stem into shorter, less complex sentences (sometimes a series of shorter sentences can create a story line or present a more familiar context/situation to students). • Context can provide description that helps make abstract or highly generalized situations more concrete and relevant. Simply stated, it helps to ground the content being tested. Context that facilitates access for English language learner students is expressed in concrete language, illustrative language, and illustrations/graphics.

Desirable characteristics	**Notes on approaches and criteria**
Item graphics	
• Familiar to students. • No cultural or linguistic bias. • Symbols, legends, and key vocabulary relevant to the construct and familiar to English language learner students. • Consistent graphic and labeling/naming conventions • Supportive of English language learner student understanding of assessed content.	• Graphics include diagrams, tables, charts, drawings, graphs, pictures, and maps. • Student knowledge about certain graphics is required and assessed in mathematics. • Graphics allow for reduced amount or complexity of language in a test item. Use of graphics in test items should serve a clear purpose. Otherwise they may be misleading or distracting. For example, graphics may be used to: o Clarify key aspects of the content/construct assessed. o Clarify construct-relevant context. o Clarify a mathematical operation. o Indicate what the student is expected to do. o Help students shift from one context to another within an assessment (for example, from one type of test item to another). o Allow students to reinforce or verify understanding of key information in test item. o Simplify the structure of a test item that requires a number of operations or steps (for example, through bulleted lists or a diagram of the complete problem that accurately reflects the problem in its totality). • Some criteria that can be used to evaluate the need for a graphic include: o Does the graphic clarify construct-irrelevant information? If so, it may not be necessary. It might be better to revise or delete the construct-irrelevant information. o Does the graphic support the test item context without requiring additional written text? o Does the graphic accurately represent the full complexity of the problem? If not, it may be misleading. o Is the graphic consistent with the key content/construct of the item?

Desirable characteristics	**Notes on approaches and criteria**
Item vocabulary/wording	
• High-frequency words. • Common and familiar words. • Relevant technical terms that reflect language of the content standards and academic English language. • Technical terms defined, as appropriate. • Naming conventions consistent with graphics/stimuli. • Construct-irrelevant vocabulary/phrases at or below grade level.	• Careful selection of vocabulary and phrases can simplify sentence structure. The amount and complexity of language should be balanced with the amount of information necessary for student to understand/access the item. The goal is to make the language as clear and straightforward as possible, while still providing the amount and complexity of information necessary to communicate the targeted content of the test item. • Some general guidelines: o Use precise language. Appropriate language modification does not simply mean using common or familiar vocabulary. o Consider language used in the content standards and academic English language . o Repeat key words/phrases in the test item that students need to understand the item and respond to it. o Do not automatically provide synonyms for a key word. This may not be helpful, especially if a test item is already long or complex. Although providing synonyms may be helpful during instruction, it may not be useful in assessment items. o Use words/phrases consistently within the context of the item and consider consistency of terms within a strand—for example, reading or measurement). Support this use with context-familiar content-based abbreviations and make explicit connections between terms/abbreviations. • If possible, avoid using: o Ambiguous words or unnecessary words with multiple meanings. o Irregularly spelled words. o Proper nouns that are irrelevant or not meaningful to the population. o Words that are both nouns and verbs (for example, carpet, value, cost); however, if a choice needs to be made, use the word only as a noun. o Hyphenated and compound words o Gerunds. o Relative pronouns (for example, which, who, that) without a clear antecedent.

Desirable characteristics	**Notes on approaches and criteria**
Item sentence structure	
• Familiar, common sentence structure. • Complexity of sentence structure at or below grade level. • Key information presented first or early in the test item. • One sentence per idea for complex test items.	• To reduce the complexity of a sentence in a test item: o Identify the agent (that is, the person or object carrying out the action) to construct sentences that use active voice (and avoid passive voice). o Make sure that the verb in a sentence follows the subject as closely as possible. o Remove introductory phrases that are irrelevant to the construct being tested. o Use conventional constructions (for example, apostrophes for possessives and “s” or “es” for plurals. o Use proper nouns that students are familiar and are grade-level appropriate. o Use clear grammatical structures. • To reduce language load: o Change past or future tense verb forms to present tense. o Change passive verb forms to active verb forms. o Change complex sentence structure to subject-verb-object structure. o Shorten any long nominals/names/phrases (for example, “last year's class vice-president” to “a student leader”). o Replace compound sentences with two separate sentences, especially when making comparisons. o Shorten or delete long prepositional phrases. o Replace conditional clauses with separate sentences. o Change the order of a clause within a sentence. o Remove or rephrase relative clauses. o Rephrase questions framed in negative terms. • Make sure the following are clear. o Noun-pronoun relationships. o Antecedent references.

Desirable characteristics	**Notes on approaches and criteria**
Item format/style	
• Clear parts of the item/question. • Explicit order of operations. • Relevant and appropriate distinctions. • Segmented or shortened long problem statements.	• Place test item elements in the following order: (1) text that introduces the graphic; (2) graphic; and (3) the test item stem. • Format for emphasis of key words/terms (highly construct-relevant), using bold, ALL CAPS, and underline to call English language learner students' attention to them. • Consider whether blocks of text (that is, a paragraph) may be necessary and appropriate for presenting a test item. This depends on the construct assessed, the complexity of the information needed by the student to respond to the item, and the centrality of the context to the construct. Suggested strategies to help English language learner students process such text include: o Bulleted lists. o Indenting key information. o Emphasizing key words/terms. o Using graphics.

Source: Sato 2008.

Key terms

This section described key terms used in the discussion of linguistically modified assessments for training item developers.

Access

To maximize student access to the content being assessed on an achievement test (for example, mathematics), text in the item that is not directly related to the targeted construct (that is, construct-irrelevant text) is minimized or removed. Doing so facilitates students' ability to demonstrate their construct-relevant knowledge and skills and reduces or eliminates sources of construct-irrelevant variance (construct irrelevance) in test results among students. In other words, when access is constrained, it can result in the measurement of sources of variance that are not related to the intended test content. If student access to tested content is restricted, students cannot fully demonstrate what they know and can do; subsequently, test results underestimate their level of content achievement (underrepresentation).

In this study the construct-irrelevant factors that constrain access to tested content for English language learner students are examined to support development of mathematics test items that maximize students' ability to show what they know and can do in mathematics.

Accommodation vs. modification

An accommodation is a change in testing conditions that is implemented to increase accessibility of test content to a specific student population. Such changes are deemed fair and reasonable when standardized administration conditions do not provide an equal opportunity for all students to demonstrate what they know and can do (Abedi & Lord 2001; Butler & Stevens 2001; Holmes & Duron 2000; National Research Council 2002, 2004). It is assumed that the same construct is being assessed with and without the accommodation. An accommodation is intended to minimize or remove the effects on test performance of construct-irrelevant factors that may contribute to, for example, the underrepresentation of student achievement in the content area.

A modification is an adjustment to the test itself, the administration conditions, or the content standards for assessment. While modification may improve access to the test content for a specific student population in a fair and reasonable manner, it significantly alters the construct being assessed. Examples of test modifications include allowing students with specific disabilities to use calculators on mathematics computation items (when general education students cannot) or allowing the reading comprehension portions of a test to be read aloud to English language learner students.

In traditional psychometric practice, accommodations may affect the performance of its intended referent group only, while remaining construct-neutral to nonaccommodated students—that is,

characteristics. However, evaluation can be done only at the discourse level. A critical reading and assignment of meaning requires minimum language beyond the word or sentence level.

the accommodation should benefit the student needing the accommodation but should have no effect on those not needing the accommodation.

However, research-based test design practices (for example, universal design, simplified language in items and associated text) suggest that all student groups may benefit from item development strategies designed to minimize construct-irrelevant variance. So, for this study an accommodation may be considered valid, even if all groups benefit from its use, if evidence collected suggests that:

- The construct/content assessed was not significantly altered.
- The performance of the group targeted for accommodation (that is, English language learner students) improves at a greater rate than that of their English-proficient counterparts.

English language learner students

English language learner students are "national-origin-minority students[39] who cannot speak, read, write, or comprehend English well enough to participate meaningfully in and benefit from the schools' regular education program" (U.S. Department of Education, Office of Elementary and Secondary Education 1999, p. 60). No Child Left Behind legislation (including Title III) refers to this population as "limited English proficient" (U.S. Department of Education, Office of Elementary and Secondary Education 2000).

This study's analyses included only students in grades 7 and 8 who identified themselves as "Hispanic" or who identified Spanish as their first language or the language spoken in their home. Recruitment efforts targeted Spanish-speaking English language learner students who scored at the mid- to high range of English language proficiency to ensure that their command of the English language was at a level sufficient to benefit from the linguistic modification.

Linguistic modification

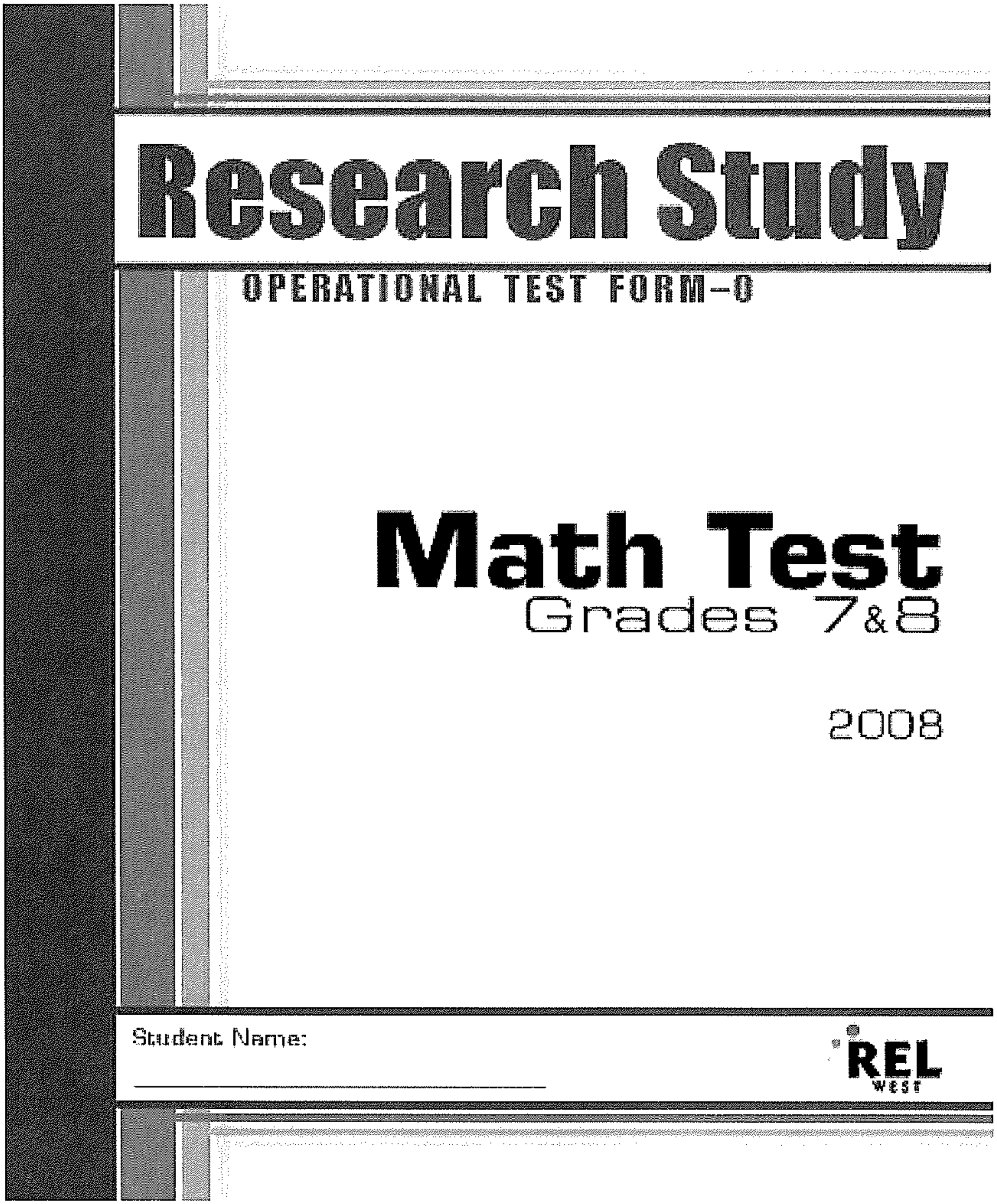
Linguistic modification is a theory- and research-based process in which the language in test items, directions, and response options is modified in ways that clarify and simplify the text without simplifying or significantly altering the construct assessed. To facilitate comprehension, linguistic modification reduces construct-irrelevant language demands (for example, semantic and syntactic complexity) of text through strategies such as reduced sentence length and complexity, use of common or familiar words, and use of concrete language (Abedi et al. 2005; Abedi, Lord, & Plummer 1997; Sireci, Li, & Scarpati 2002).

Linguistic modification is not simply good editing practice and does not result in simpler items. Rather, it is a linguistically based, systematic means for targeting, reducing, and removing the irrelevant variance in test performance that is attributable to individual differences in English proficiency so that English language learner students can fully demonstrate what they know and

[39] "National origin minority" can include students born in the United States.

can do in that content area. By minimizing the language load, a source of construct-irrelevant variance, English language learner students' access to construct-relevant content is enhanced.

Appendix N. Operational item set—original



Research Study

OPERATIONAL TEST FORM–O

Math Test

Grades 7&8

2008

Student Name:

____

REL
WEST

3. Fifteen boxes each containing 8 radios can be repacked in 10 larger boxes each containing how many radios?

A. 8

B. 12

C. 80

D. 120

7. What is 4 hundredths written in decimal notation?

A. 0.004

B. 0.04

C. 0.400

D. 4.00

10. If Jill is driving at 65 miles per hour, what is her approximate speed in kilometers per hour? (1 mile ≈ 1.6 kilometers)

A. 16

B. 41

C. 104

D. 173

11. A certain reference file contains approximately one billion facts About how many millions is that?

A. 1,000,000

B. 100,000

C. 10,000

D. 1,000

12. A car odometer registered 41,256.9 miles when a highway sign warned of a detour 1,200 feet ahead. What will the odometer read when the car reaches the detour? (5,280 feet = 1 mile)

A. 42,456.9

B. 41,261.3

C. 41,259.2

D. 41,257.1

14. The mean distance from Venus to the Sun is 1.08×10^8 kilometers. Which of the following quantities is equal to this distance?

A. 10,800,000 kilometers

B. 108,000,000 kilometers

C. 1,080,000,000 kilometers

D. 10,800,000,000 kilometers

15. If the values of the expressions below are plotted on a number line, which expression would be closest to five?

A. $|-4|$

B. $|-18|$

C. $|7|$

D. $|16|$

17. A sweater originally cost $37.50. Last week, Moesha bought it at 20% off.

How much was deducted from the original price?

A. $7.50

B. $17.50

C. $20.00

D. $30.00

20. A landscaper estimates that landscaping a new park will take 1 person 48 hours. If 4 people work on the job and they each work 6-hour days, how many days are needed to complete the job?

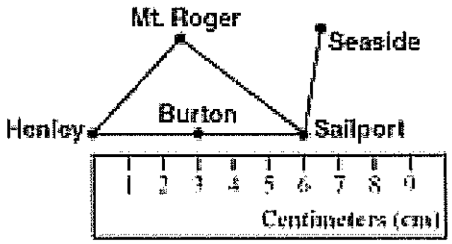
A. 2 days

B. 4 days

C. 6 days

D. 8 days

24. Javier is using a ruler and a map to measure the distance from Henley to Sailport.



The actual distance from Henley to Sailport is 120 kilometers (km). What scale was used to create the map?

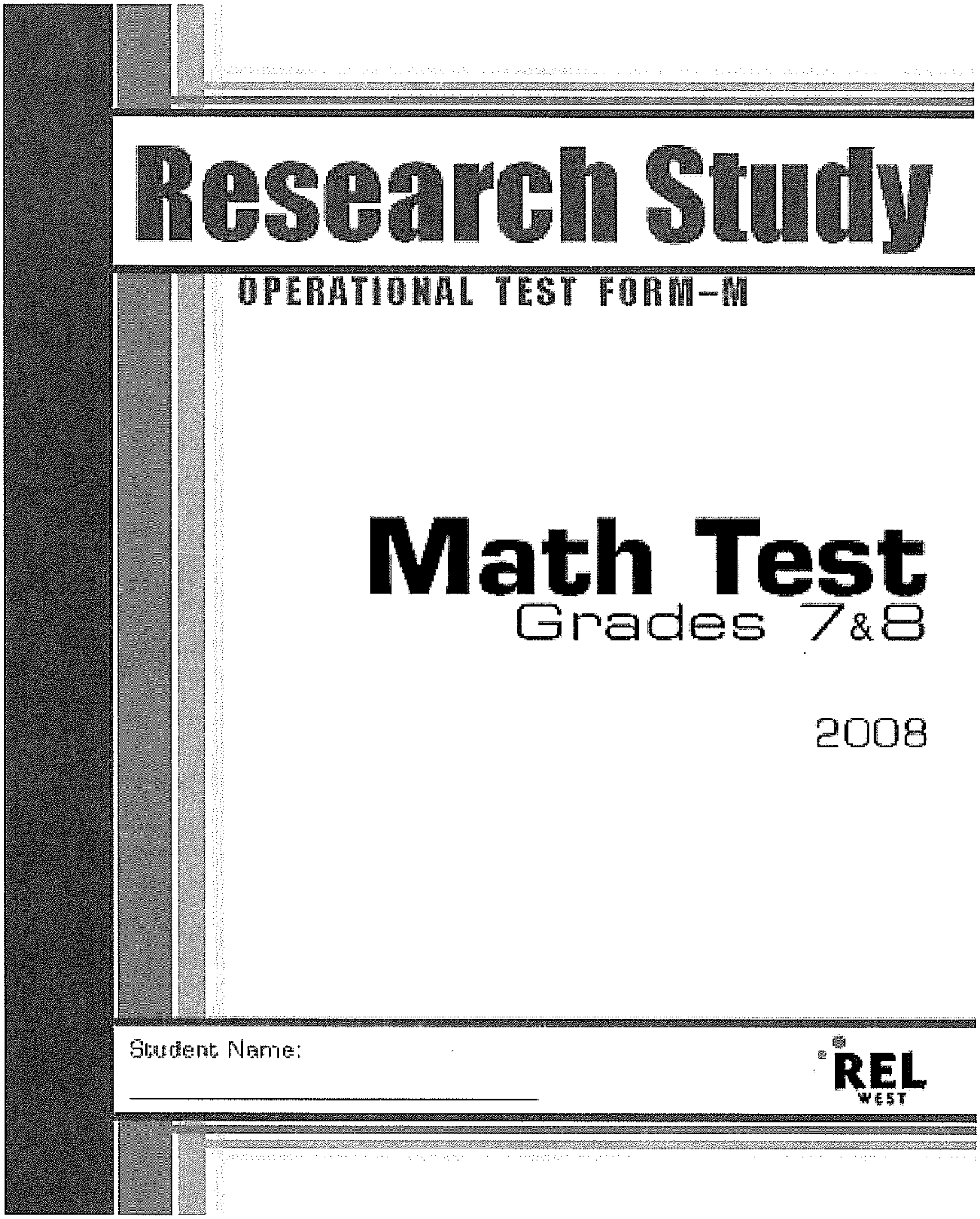
A. 1 cm = 6 km

B. 1 cm = 12 km

C. 1 cm = 15 km

D. 1 cm = 20 km

Appendix O. Operational item set—linguistically modified



Research Study

OPERATIONAL TEST FORM-M

Math Test

Grades 7&8

2008

Student Name: ______________________

REL WEST

3. A student works in a store.

- She unpacks 15 boxes.
- Each box contains 8 radios.
- She repacks the radios in 10 larger boxes.
- Each box contains the same number of radios.

How many radios are in each larger box?

A. 8

B. 12

C. 80

D. 120

7. 4 hundredths = ________

A. 0.004

B. 0.04

C. 0.400

D. 4.00

10. 65 miles per hour is about _____ kilometers per hour
(1 mile ≃ 1.6 kilometers)

A. 16

B. 41

C. 104

D. 173

11. How many millions is 1 billion?

A. 1,000,000

B. 100,000

C. 10,000

D. 1,000

12. A car's mileage is 41,256.9 miles.
The car travels 1,200 feet to an exit.
What is the car's mileage at the exit?
(5,280 feet = 1 mile)

A. 42,456.9

B. 41,261.3

C. 41,259.2

D. 41,257.1

14. Which distance equals
1.08×10^8 kilometers?

A. 10,800,000 kilometers

B. 108,000,000 kilometers

C. 1,080,000,000 kilometers

D. 10,800,000,000 kilometers

15. Which value is closest to five on a number line?

A. $|-4|$

B. $|-18|$

C. $|7|$

D. $|16|$

17. A girl wants to buy a sweater on sale.

- The regular price is $37.50.
- The discount is 20% of the regular price.

What is the amount of the discount?

A. $7.50

B. $17.50

C. $20.00

D. $30.00

20. A manager hires students to do a job.

- She estimates that 1 student needs 48 hours to do the job.
- She hires 4 students to do the job together.
- Each student works 6 hours per day.

What is the total number of days the 4 students need to do the job?

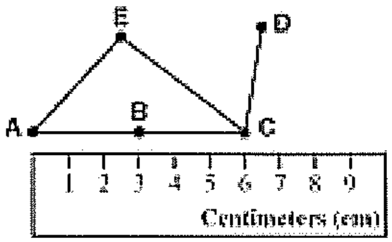
A. 2 days

B. 4 days

C. 6 days

D. 8 days

24. Look at the map and ruler below. The diagram below shows the distance from Point A to Point C on a map.



The actual distance from Point A to Point C is 120 kilometers (km). What is the scale of the map?

A. 1 cm = 6 km

B. 1 cm = 12 km

C. 1 cm = 15 km

D. 1 cm = 20 km

WestEd

Plain English Workshop
Template for Plain English Item Evaluation

Item Number: ____________________________

Level of Cognitive Complexity	Language that <u>should not</u> be simplified or changed	Language that can/should be simplified or changed

Evaluation of Item Elements for Plain English: Accessibility of Content		
Item Context	Item Graphics	Item Vocabulary/ Wording

Evaluation of Item Elements for Plain English: Accessibility of Content		
Item Sentence Structure	Item Format/ Style	Other/Comments

Revised Item:

has

19. ~~When he left the pizza restaurant,~~ Joseph had 25 pizzas to deliver. At his first stop, he delivered five pizzas ~~to a party~~. At his second stop, he delivered half of the remaining pizzas ~~to a school~~. At each remaining stop, he delivered one pizza. How many stops did Joseph make to deliver the 25 pizzas?

s s s

A 3

B 10

C 12

D 25

present tense too much info.

20. Morgan's family made a large pizza for lunch ~~on Saturday~~. Morgan ate $\frac{3}{12}$ of the pizza. Megan ate $\frac{1}{6}$ of the pizza, and Emma ate $\frac{1}{12}$ of the pizza. Their parents ate $\frac{1}{3}$ of the pizza. How much pizza was left?

A $\frac{1}{12}$

B $\frac{1}{6}$

C $\frac{6}{12}$

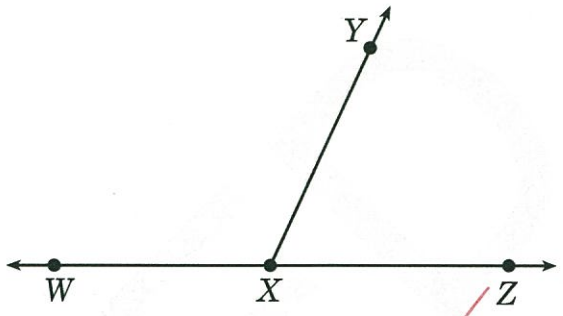
D $\frac{5}{6}$

~~was~~ is not eaten

Morgan v/s Megan ⇓ many

Tense?

21. ***About*** how many degrees is the measure of $\angle WXY$?



A 20°

B 60°

C 120°

D 160°

✓

22. Joey was looking at a square, a rectangle, and a right triangle. What is the total number of angles for all of the polygons, and how many are right angles?

asking two quests

A 11 angles, 8 right angles

B 11 angles, 9 right angles

C 12 angles, 8 right angles

D 12 angles, 9 right angles

present tense

What is the quotient?

19. Cara used this multiplication table to help her find the quotient for 112 ÷ 14.

Is the table necessary?

Multiplication Table

×	10	11	12	13	14	15	16
6	60	66	72	78	84	90	96
7	70	77	84	91	98	105	112
8	80	88	96	104	112	120	128
9	90	99	108	117	126	135	144
10	100	110	120	130	140	150	160
11	110	121	132	143	154	165	176

What answer should Cara get?

A 16

B 11

C 8

D 7

large

20. Mrs. Jones has some baskets of strawberries to sell. She has 52 baskets each containing 3 pounds of strawberries and 48 smaller baskets each containing 2 pounds of strawberries. ***About*** how much will her strawberries weigh in all?

A 250 pounds

B 200 pounds

C 150 pounds

D 100 pounds

Bullets?

present

cuts

21. Sallie ba~~ked~~ 4 apple pies and ~~cut~~ each of them into sixths. ~~If~~ S she ~~served~~ she serves $3\frac{1}{2}$ pies. H how many slices of pie did Sallie serve?

A 24

B 21

C 18

D 9

22. ~~Clint's teacher asked him to~~ write two fractions that are equivalent to $\frac{2}{5}$. ~~If Clint did this problem correctly, which answer did Clint write?~~

Context makes it harder

A $\frac{2}{10}$ and $\frac{4}{10}$

B $\frac{4}{10}$ and $\frac{6}{10}$

C $\frac{2}{10}$ and $\frac{20}{100}$

D $\frac{4}{10}$ and $\frac{40}{100}$

16. Which chart shows below the rule that the output value is two less than the input value?

A

Input	Output
5	7
8	10
11	13
12	14

B

Input	Output
5	3
8	4
11	9
12	10

C

Input	Output
5	10
8	16
11	22
12	24

D

Input	Output
5	3
8	6
11	9
12	10

17. The bread truck makes deliveries to a store 3 days each week. Each delivery has 45 loaves of bread. Which expression could be used to determine the number of loaves of bread delivered in 5 weeks?

A 3×5

B $45 \div (3 \times 5)$

C 45×3

D $45 \times 3 \times 5$

yard

18. Michael cuts grass for $15.00 per lawn. He cuts 2 lawns each day for 6 days a week. How much will Michael earn in 2 weeks?

A $390

B $360

C $180

D $90

has

1. The library ~~has~~ 7,126 books. The library will purchase exactly one hundred more books. How many books will ~~the library~~ have after the books are purchased?

Buys

present tense

A 7,136

B 7,137

C 7,226

D 8,126

2. There are 20 seeds in a package. If 5 seeds are put in each flower pot, how many flower pots are needed to plant all of the seeds?

A 4

B 5

C 15

D 25

chairs

3. A box of candy has 12 rows. There are 6 pieces of candy in each row. How many pieces of candy are in the box?

room

A 6

B 18

C 62

D 72

Needs a model or works better as a model

one day

OR

more?

4. On Saturday, 2,759 people went to the afternoon concert and 6,387 people went to the night concert. ***About*** how many people went to the concert on Saturday?

A 4,000

B 6,000

C 8,000

D 9,000

has

pennies

5. Dean ~~had~~ 1,062 pennies ~~in his bank~~. Shawn had 889. How many more pennies ~~did~~ Dean have than ~~Shawn~~?

John

does

John

A 173

B 223

C 227

D 283

Jerry collects rocks. Jerry keeps his rocks in 7 box

6. ~~Jerry keeps his rock collection~~ in 7 ~~boxes~~. Each box weighs about 6 or 7 pounds. How much does Jerry's whole rock collection weigh?

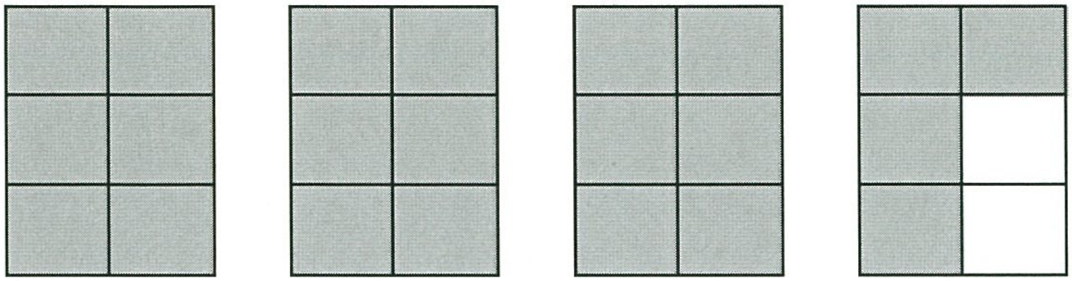
A between 50 and 60 pounds

B between 40 and 50 pounds

C between 30 and 40 pounds

D between 20 and 30 pounds

1. Which mixed number represents the shaded parts of the model?



A $3\frac{2}{6}$

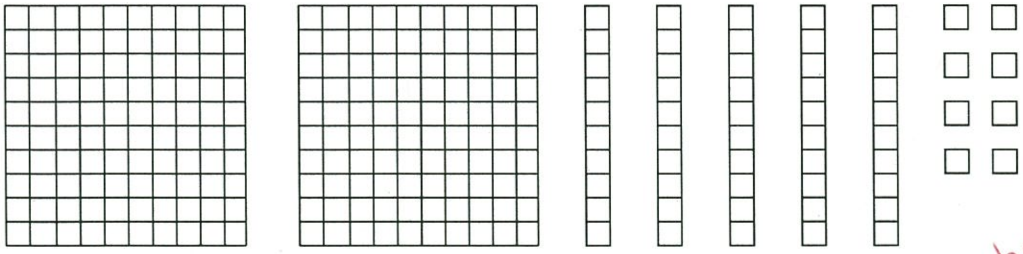
B $3\frac{4}{6}$

C $4\frac{2}{6}$

D $4\frac{4}{6}$

- Introduce Model or label it.
- Q. Is model part of curriculum
- add "shown below" at end
- another word for "model"? Boxes cells

2. Which number is 100 more than the model shown below?



A 158

B 258

C 358

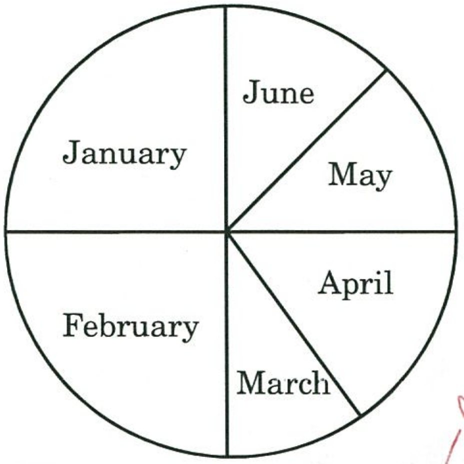
D 385

Is there better placement for the stem?

Car store
Mr. Jones

30. A dealership sold 200 cars in a six-month period. The circle graph below displays the distribution of sales by month.

Distribution of Car Sales



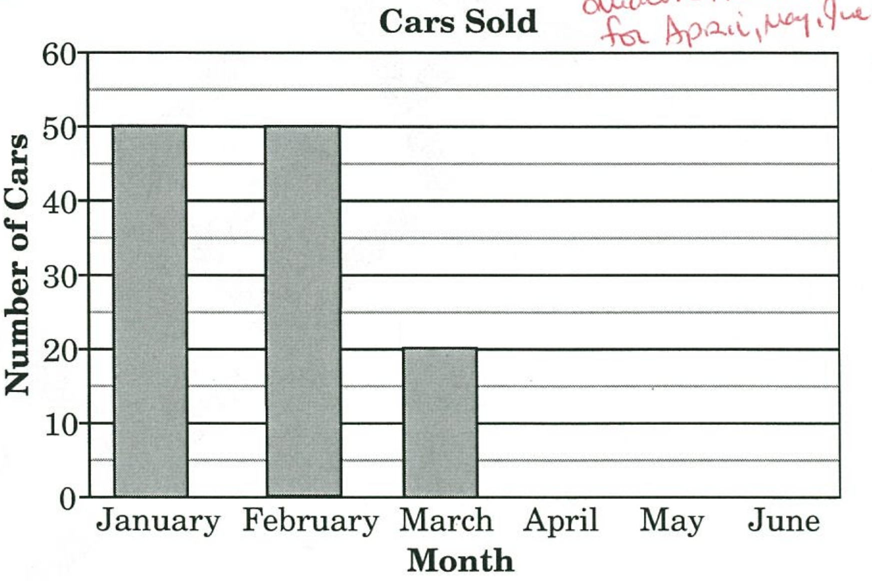
Like the graphs

Could break up sentences

The sales manager at the dealership created the bar graph below to show the number of cars sold each month during the six-month period. (~~The bars for April, May, and June have~~ not yet been drawn.)

Mr. Jones has not drawn the bars for April, May, June.

passive is difficult make it active



The dealership sold the same number of cars in June as in May. How many cars did it sell in April?

A 20

B 25

C 30

D 35

Test Development Process

How Our Teachers Write and Review Test Items

North Carolina teachers are very involved in the development of the End-of-Grade (EOG) Assessments, End-of-Course (EOC) Assessments, and the NC Final Exams beginning with the item writing process as explained below:

- North Carolina professional educators from across the state who have current classroom experience are recruited and trained as item writers and developers for state tests.
- Diversity among the item writers and their knowledge of the current state-adopted content standards are addressed during recruitment.
- The use of classroom teachers from across the state ensures that instructional validity is maintained.

North Carolina teachers are also recruited for reviewing the written test items.

- Each item reviewer receives training in item writing and reviewing test items.
- Based on the comments from the reviewers, items are revised and/or rewritten, item-objective matches are reexamined and changed where necessary, and introductions and diagrams for passages are refined.
- Analyses occur to verify there is alignment of the items to the curriculum.
- Additional items are developed as necessary to ensure sufficiency of the item pool.
- Test development staff members, as well as curriculum specialists, review each item.
- Representation for students with special needs is included in the review.
- This process continues until a specified number of test items are written to each objective, edited, reviewed, edited again, and finalized.

If a teacher is interested in training to become an item writer or reviewer for the North Carolina Testing Program, he/she can visit https://center.ncsu.edu/nc/x_courseNav/index.php?id=21 and take the appropriate subject area “A” level Content Standards Overview course and the “B” level Test Development Basics course in the Moodle system. Once the online training courses are completed, the teacher will be directed to go to an online interest form at http://goo.gl/forms/wXv4Imh0ko. Here the teacher can register to let the North Carolina Testing Program know he/she is interested in writing or reviewing. Teachers who submit interest forms will be contacted when item writing or reviewing is needed in their subject area.

For an in-depth explanation of the test development process see State Board policy GCS-A-013 or reference http://www.ncpublicschools.org/accountability/testing/shared/testdevprocess.

Technology Enhanced Item (TEI)
Usability Study Evaluator Questions

INDIVIDUAL STUDENT OBSERVATIONS	
STUDENT NAME:	*(CIRCLE ONE)* GENERAL / EXTEND2

Directions

1. Were the directions for each item type clear to the student?

   □ Yes □ No (explain)

____

____

2. On average, how much time did the student need to read directions before knowing how to answer the questions?

   □ 1 min or less □ 1 to 2 mins. □ 2 mins. or more

3. For each TE item, did the student know exactly how to indicate his/her answer choice?

   □ Yes □ No (explain)

____

____

Use

4. Did each TE item work correctly for the student?

   □ Yes □ No (explain)

____

____

5. Was it clear to the student that the computer registered his/her answer choice?

   □ Yes □ No (explain)

____

____

6. Was the student able to locate information on the screen as she/he needed it?

□ Yes □ No (explain)

____

____

7. Did the use of a scroll bar or slider bar diminish *usability* of the TE items?

□ No □ Yes (explain)

____

____

Accessibility

8. Did the use of a scroll bar or slider bar diminish *accessibility* of the TE items?

□ No □ Yes (explain)

____

____

9. Which online system accommodation features (e.g., color schemes, screen magnification, audio players, etc.) were used by the student?

____

____

10. Did you observe any access issues for this student?

□ No □ Yes (explain)

____

____

____

____

____

Reactions to New Item Types

11. How did the student react to the TE item types?

Programming

12. Did the TE items function correctly for the student?

□ Yes □ No (explain)

13. Were data/answers captured and stored correctly?

□ Yes □ No (explain)

14. Did the scoring work correctly?

□ Yes □ No (explain)

Summary Notes (Ask student if she has any comments.)

Technology Enhanced Item (TEI)
Usability Study Evaluator Questions

A Special Study of Innovative Assessment Items by the
North Carolina Department of Public Instruction and North Carolina State
University (TOPS) in Collaboration with Wake County Public Schools,
Fall 2011

Participating Schools:
Fuquay-Varina High
Fuquay-Varina Middle
Fuquay-Varina Elementary

Study Coordinator: Jerrie W. Brown, Sr. Educational Research and
Evaluation Consultant, North Carolina State University

Technology Enhanced Item (TEI)
Usability Study Evaluator Questions

SUMMARY OBSERVATIONS	
EVALUATOR NAME:	DATE:

Directions

1. Which students were confused by the directions of the item?

   General Ed. □ Extend 2 □

   ____

   ____

   ____

2. What changes to the directions for each item type (*Grid-Ins, Text Identify, String Replace, Sequence Order, Label Property Match*) do you recommend?

   ____

   ____

   ____

   ____

Use

3. For students with limited computer experience, do the TE items make sense (intuitive)?

   Yes □ No □

   ____

   ____

4. Did students have difficulty selecting their answer choices?

   Yes □ No □

5. For each TE item, were the students easily able to indicate their answer choices?

   Yes □ No □

6. In your opinion, are some item types susceptible to practice effects?

   Yes □ No □

7. Did the usability of the items vary across types of students (Extend2 versus General Ed.)?

   No □ Yes (explain) □

   ____

   ____

   ____

   ____

8. What changes do you recommend?

   ____

   ____

   ____

   ____

Accessibility

9. How did the online system accommodation features affect the usability of the TE items?

   ____

   ____

   ____

   ____

10. What recommendations can you make to minimize any access issues?

   ____

   ____

____

____

Programming

11. Did the multi-media present/work properly?

    Yes □ No (explain) □

    ____

    ____

    ____

12. What changes do you recommend?

    ____

    ____

    ____

    ____

Summary Recommendations

13. Should students be required to practice all TE item types prior to an operational assessment (to ensure that lack of familiarity with the TE item does not adversely affect their performance)?

    Yes □ No □

    ____

    ____

    ____

    ____

14. Given the amount of time required by some items, should the points awarded for a correct response be adjusted? (could be 0=wrong, 2 =right)

    Yes □ No □

    ____

    ____

15. What aspects of each item type (*Grid-Ins, Text Identify, String Replace, Sequence Order, Label Property Match*) minimized usability?

16. What aspects of each item type (*Grid-Ins, Text Identify, String Replace, Sequence Order, Label Property Match*) minimized accessibility?

17. What recommendations can you make to minimize such access issues and maximize usability?

Additional Comments:

Item Writing and Review for Bias and Sensitivity and Differential Item Functioning (DIF)

Including processes for EC, ESL, VI reviews

Defined

Item creation for the North Carolina Testing Program has an established history of inclusion of consideration for bias and sensitivity, and this has been considered as an integrated part of the development process prior to field testing. Vetting steps that specifically involve the EC/ESL/VI Specialists look for content that may present a bias or insensitivity issue such as contexts that might elicit an emotional response and inhibit students' ability to respond and contexts that students may be unfamiliar with for cultural or socio-economic reasons.

Participant Requirements

Teachers in North Carolina are the principal target population, but participants can be augmented with retired teachers and or those holding undergraduate degrees in the content area. The number of item writers and reviewers required during any item development period is determined by the need and the time allotted. All item writers and reviewers must be trained for bias and sensitivity.

Training Requirements

Item writers and reviewers must be trained on the standards and content being measured. All item writers and reviewers are subjected to extensive training on proper item design and they are also trained to consider bias and sensitivity of item content. Additionally, since the vetting process includes specific steps for EC, ESL, and VI check, training is required for these reviewers. Depending on the event and the experience of the group that is being asked to write and review, training may be best applied in a face-to-face session. However, the majority of training is designed to be delivered in self-directed online training modules.

Process and Timeline

Item writing can begin any time a change in standards has been initiated for any content that is required to be measured with a standardized test administration. See the flowcharts in the appendices for the process of writing and review that items must go through in order to be considered candidates for inclusion on either stand-alone field tests or as embedded experimental items on operational tests. Quantities and type of items per targeted standard and the time frame set by leadership of when operational tests are to exist helps determine the timeline for when items must be ready and how many item writers and reviewers are needed.

DIF Review

Defined

Per step 14 in the official SBE approved Test Development Process Flow Chart (http://www.ncpublicschools.org/docs/accountability/latestflowchart.pdf) bias reviews occur after items have been field tested and have data that supports further inspection of the items for bias or insensitivity. This is processed in steps within the online test development system (TDS) that are titled DIF Review.

The methodology used for the North Carolina Testing Program to identify items that show differential item functioning (DIF, sometimes called "statistical bias", is a concept that is different from the non-technical notion of "bias") is the Mantel-Haensel Delta-DIF method.

Calculating Statistical Bias using Mantel-Haensel Delta-DIF Method

Since the method depends on sample size, there is no single number or range of numbers that identifies an item as having moderate or more significant levels of DIF. Rather, the statistical methodology takes the sample size into account and determines whether an item should be rated as A, B, or C, according to whether it displays no significant DIF (A level), significant but still low level of DIF (B level), or more pronounced DIF (C level). A minimum number of 300 per subgroup is necessary in order to produce DIF values that are stable and do not exaggerate the counts of DIF in the B and C levels.

The current operational strategy is to reduce or eliminate the need for DIF Review by choosing not to use any item that has any significant degree of differential item functioning (C level DIF). In the rare case where an item is needed to fill test form design parameters and no A level DIF item exists, then an item in B (first choice) or C (last resort) DIF is put through an additional bias review process that content specialists coordinate.

The current subgroup analyses conducted are: Male/Female, White/Black, White/Hispanic, Urban/Rural, EDS/non-EDS.

This is the same system that the National Assessment of Educational Progress uses. For each analysis of DIF, there is a focal group and a reference group. For example in the male-female analysis, the focal group is females and the reference group is males. A plus (+) or minus (-) sign is used to indicate the direction of DIF. For example, if an item has a B- rating for the male-female analysis that means that the item slightly disfavors (minus sign) females (or slightly favors males). There may be many reasons for a B rating, and such a rating is by no means regarded as a reason to forbid the item to be on a test.

Below are some relevant links that describe the DIF methodology and related topics. The last link shows that NAEP sometimes does use items that have been flagged as having certain levels of DIF (click the individual links for the tests in the various NAEP content areas), provided that those items receive approval following the bias panel review and the subsequent content review. Ultimately, in NAEP's process, the final decision of whether to use an item is made by human beings based on all available info. It is not an automated decision produced purely by computer analyses.

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_proced.aspx
- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dif_categ.aspx

- https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_avoidviolat_results.aspx

Participant Requirements

DIF Review participants collectively must model the dimensions that are subject to the DIF parameters which match the Bias Review Panel participants. Since the volume of items that typically get flagged for non-A level values in the analysis that need to go through DIF Review is very small, the number of participants can likewise be a minimum set of five or six.

Training Requirements

DIF Review participants are required to go through the same training provided to the item writers and reviews and the Bias Review panel participants.

Review Process and Timeline

Tests are administered both fall and spring and the DIF analyses is done after the spring administration on combined data (fall and spring).

February through May:

- DIF reviews of DIF flagged items from the Fall

June through September:

- DIF reviews of DIF flagged items from the Spring

October through February:

- Spring base forms are assembled and embedded items are placed

DIF Review Questions

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
   No
   Yes - Explain

2. Does the item contain any local references that are not a part of the statewide curriculum?
   No
   Yes - Explain

3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
   No
   Yes - Explain

4. Does the item contain any demeaning or offensive materials?
   No
   Yes - Explain

5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
   No
   Yes - Explain

6. Does the item assume that all students come from the same socioeconomic background?
(e.g., a suburban home with two-car garage)
   No
   Yes - Explain

7. Does the artwork adequately reflect the diversity of the student population?
   Yes
   N/A
   No - Explain

8. Is there any source of bias detected in this item?
   No
   Yes - Explain

Additional Comments:

Sample Bias and Sensitivity Training Materials

Instructions for Review

What is the purpose of this review?

After items are field tested, statistics are gathered on each item based on examinees' responses. Sometimes, the statistics indicate the possibility of Construct-Irrelevant Variance – "noise" in the item that prevents us from knowing something about the student's abilities and is measuring something else instead. Your part in this review is to judge whether the content of the item is in fact measuring something about the student other than his or her ability or knowledge in the content area that the question was intended to measure.

How were these items identified for review?

Through a statistical technique called "Differential Item Functioning" (DIF). After controlling for students' ability, are there differences in performance on the item between groups? If an item behaves differently statistically for one group of examinees than it does for another group of examinees, it is flagged for review.

The content of the items was not considered during the statistical analysis. So, these items were flagged for review because we need to determine if there is anything about these items that may be a source of bias.

What is bias?

TRUE Bias is when

- An item measures membership in a group more than it measures a content objective.
- An item contains information or ideas that are unique to the culture of one group AND this information or idea is not part of the course of study (North Carolina Essential Standards or North Carolina Common Core Standards).
- The item cannot be answered by a person who does not possess some certain background knowledge.

Sensitivity is another issue that could occur in an item. Sensitivity issues occur when

- An item contains information or ideas that some people will find objectionable or raise strong emotions AND this information or idea is not part of the course of study.
- Assumptions are made within the item that all examinees come from the same background.

Bias is NOT

- Just having a boy's name or a girl's name in the item
- Just mentioning a part of the state, country, or world
- Just mentioning an activity that is variably familiar to certain groups (e.g., vacations, using a bank)
- Just mentioning a "boy" activity (e.g., sports) or a "girl" activity (e.g., cooking) Think about: Jackee Joyner-Kersee or Babe Zaharias; Emeril or The Cajun Chef

DIF versus Bias

There is, then, a distinction between DIF and bias. DIF is a statistical technique whereas bias is a qualitative judgment. It is important to know the extent to which an item on a test performs differently for different students. DIF analyses examine the relationship between the score on an item and group membership, while controlling for ability, to determine if an item may be behaving differently for a particular group. While the presence or absence of true bias is a qualitative decision, based on the content of the item and the curriculum context within which it appears, DIF can be used to quantitatively identify items that should be subjected to further scrutiny.

Guidelines for Bias Review

All groups of society should be portrayed accurately and fairly without reference to stereotypes or traditional roles regarding gender, age, race, ethnicity, religion, physical ability, or geographic setting. Presentations of cultural or ethnic differences should neither explicitly nor implicitly rely on stereotypes nor make moral judgments. All group members should be portrayed as exhibiting a full range of emotions, occupations, activities, and roles across the range of community settings and socioeconomic classes. No one group should be characterized by any particular attribute or demographic characteristic.

The characterization of any group should not be at the expense of that group. Jargon, slang, and demeaning characterizations should not be used, and reference to ethnicity, marital status, or gender should only be made when it is relevant to the context. For example, gender neutral terms should be used whenever possible.

In writing items, an item-writer, in an attempt to make an item more interesting, may introduce some local example about which only local people have knowledge. This may (or may not) give an edge to local people and introduce an element of bias into the test. This does not mean, however, that no local references should be made if such local references are a part of the curriculum (in North Carolina history, for example). The test of bias is this: Is this reference to a cultural activity or geographic location something that is taught as part of the curriculum? If not, it should be examined carefully for potential bias.

Name of Reviewer: ______________________________ Date: _________

When reviewing testing materials for bias, consider the following:

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
2. Does the item contain any local references that are not a part of the statewide curriculum?
3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
4. Does the item contain any demeaning or offensive materials?
5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
6. Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
7. Does the artwork adequately reflect the diversity of the student population?
8. Other comments
9. No source of bias detected in the item

Test Development Process
Item, Selection and Form Development

North Carolina Testing Program

Published December 2015

North Carolina Department of Public Instruction
Accountability Services Division

Table of Contents

Item Development Process

Prior to **Step 1**, the standards to be measured must be defined. The test development process begins after new content standards are adopted by the North Carolina State Board of Education. All item writers and reviewers are required to complete North Carolina developed online-training modules available through the NC Education site. The training includes a general course on item writing guidelines, including lessons on sensitivity and bias concerns. The writers and reviewers must also complete subject-specific courses on the Essential Standards or North Carolina *Standard Course of Study*.

Step 1: Item Created
Test items are written by North Carolina-trained item writers, including North Carolina teachers and/or curriculum specialists, and Content Specialists at Technical Outreach for Public Schools at North Carolina State University. All items are submitted through an online test development system. The item writer assigns the item:

- a Clarifying Objective/Standard
- a secondary Clarifying Objective/Standard (when appropriate)
- a Depth-of-Knowledge (DOK) rating (if applicable)
- a knowledge type and cognitive category (if applicable)
- category (when appropriate)

The item writer is also responsible for citing sources for any stimulus material to an item.

Step 2: Item Evaluation
Content Specialists review the item for accuracy of content, appropriateness of vocabulary (both subject-specific and general), overall readability, adherence to item writing guidelines, and sensitivity and bias concerns. All content specialists (subject and the Exceptional Children/English as a Second Language/Visually Impaired (EC/ESL/VI) specialist) look for contexts that might elicit an emotional response and inhibit students' ability to respond as well as contexts that students may be unfamiliar with for cultural or socio-economic reasons. The specialists review the item's assigned:

    - Clarifying Objective/Standard
    - secondary Clarifying Objective/Standard (if applicable)
    - DOK rating (if applicable)
    - Key/appropriate foils
    - difficulty rating
    - category (if applicable)
    - knowledge type and cognitive category (if applicable)

- If the content of the item is not accurate or does not match an objective/standard, or if the DOK of the item is not appropriate, the item is revised or deleted.
- If necessary, the specialist should edit the stem and foils of the items for clarity and adherence to established item writing guidelines.
- If there are necessary revisions outside the technical scope of the specialist (such as artwork, graphs, or edits to English/Language Arts (ELA selections), the item is moved to **Step 3** for edits by Production staff.
- If the item contains stimulus material, the item is moved to **Step 3** for copyright checks by Copyright staff.

Once the item is accepted, the item is sent to **Step 4** (Teacher Content Review).
The item is sent to teacher review once the content specialist has spent the needed time on revising the item as necessary.

Step 3: Production Edits/Copyright Checks
Items needing revisions outside the technical scope of the Content Specialist (such as artwork, graphs, and ELA selections) are revised by Production staff. Items with stimulus materials are reviewed by Copyright staff for copyright concerns and proper citation. Once the item is revised by Production or reviewed for copyrights, it is moved to **Step 2** for another review by a Content Specialist.

Step 4: Teacher Content Review
Teacher content item reviewers are required to undergo the same training as item writers. Two North Carolina-trained item reviewers look for any quality issues or bias/sensitivity issues and suggest improvements, if necessary. These trained reviewers evaluate the item in terms of:

- alignment to grade-level content standard
- content of item: accurate content, one and only one correct answer, appropriate and plausible context
- the stem is clearly written
- plausible but incorrect distractors
- item design conforms to North Carolina item writing guidelines
- appropriate language for the academic content area and age of students
- bias or sensitivity concerns

Step 5: Reconcile Teacher Content Reviews
A Content Specialist carefully reviews all comments/suggestions from the content reviewers and makes any appropriate revisions. The Content Specialist may choose one of the following options:

- Send the item to **Step 6** (Production) if there are revisions required that are outside the technical scope of the Content Specialist.
- Send the item to **Step 7** (NCDPI-Curriculum and Instruction and EC/ESL/VI) if the item is ready for the next stage of review.
- Send it back to **Step 4** (teacher review) if major revisions are made.
- Delete the item.

Step 6: Production Edits
Items needing revisions outside the technical scope of the Content Specialist (such as artwork, graphs, and ELA selections) are revised by Production staff. Once the item is revised by Production staff, it is sent back to **Step 5** for review by a Content Specialist.

Step 7A: NCDPI-Curriculum and Instruction Review
A North Carolina Department of Public Instruction (NCDPI)-Curriculum and Instruction Specialist reviews the item and assigns a Clarifying Objective (Essential Standards) or a Standard (NC *Standard Course of Study*). The reviewer evaluates the item in terms of:

- alignment to grade-level content standard
- one and only one correct answer
- the assigned Cognitive Process and Knowledge Type (Essential Standards) or Depth of Knowledge (NC *Standard Course of Study*)
- bias, insensitivity, or accessibility issues
- overall item quality

The NCDPI-Curriculum and Instruction reviewer rates the item as acceptable, acceptable with revisions, or unacceptable. The review can also include additional comments. In the additional comments, the reviewer can also request that the item be returned to this step by the Test and Measurement Specialist when he or she reviews the item.

Step 7B: Exceptional Children (EC), English as a Second Language (ESL), and Visually Impaired (VI) Review
The EC/ESL/VI Specialists reviews the item for accessibility concerns for the exceptional children, English as a Second Language, and Visually Impaired student populations. This review addresses concerns due to bias or insensitivity issues, such as contexts that may elicit an emotional response, inhibit a student's ability to respond, or may be unfamiliar to a student for cultural or socio-economic reasons. Each item is evaluated in terms of:

- stem is a clear and complete question
- straightforward foils
- no repetitive words
- grammar of stem agrees with foils
- alignment to grade-level expectation
- overall content and readability
- review modifying words
- make suggestions to add or remove bold print and italics
- review for idioms and two-word verbs that may provide inhibit accessibility for ESL students
- accessibility of graphics (and ability to Braille graphics) for students for visually impaired students

Step 7C: Literacy Review (Portfolio Item Review only)
For Grade 3 Portfolio Items, a Literacy specialist evaluates each item for grade-level appropriateness.

Step 8: Reconcile Step 7 Reviews
A Content Specialist reviews comments/suggestions from the NCDPI-Curriculum and Instruction and EC/ESL/VI reviewers (and the Literacy reviewer for Grade 3 Portfolio), and makes any necessary revisions. The Content Specialist should indicate in the comments if any comments/suggestions from the reviewers were not approved and incorporated. The Content Specialist may choose one of the following options:

- Send the item to **Step 9** (Production) if there are revisions required that are outside the technical scope of the Content Specialist.
- Send the item to **Step 10** (Test Measurement Specialist Review) for review.
- Send it back to **Step 4** (Teacher Review) if major revisions are made.
- Delete the item.

Step 9: Production Edits
Items needing revisions outside the technical scope of the Content Specialist (such as artwork, graphs, and ELA selections) are revised by Production staff. Once the item is revised by Production staff, it is sent back to **Step 8** for another review by a Content Specialist.

Step 10: NCDPI-Test Measurement Specialist Review
A NCDPI-Test Measurement Specialist (TMS) reviews for overall item quality. The TMS also checks that quality control measures have been followed by reading the comments from all previous reviews and verifying that the comments have been addressed by the Content Specialists. The TMS evaluates the item for:

- alignment to grade-level content standard and vocabulary
- verification of one and only one correct answer
- assigned Cognitive Process and Knowledge Type (Essential Standards) or Depth of Knowledge (North Carolina *Standard Course of Study*)
- bias, insensitivity, or accessibility issues
- overall item quality

The TMS has four options when submitting the review:

- If the TMS approves the item as is, the item proceeds to **Step 13** (Grammar Review).
- If the TMS indicates edits are needed, the item proceeds to **Step 11** for review by a Content Specialist.
- If NCDPI-Curriculum and Instruction staff indicated they would like to see the item again, the TMS can move the item back to **Step 7** for reconciliation.
- The TMS can also choose to delete the item.

Step 11: Reconcile TMS Review, Grammar Review, or Security Review
A Content Specialist reviews comments/suggestions from the Test Measurement Specialist from **Step 10**, Editing staff from **Step 13** (Grammar Review), or Production staff from **Step 14** (Security Review) and makes any necessary revisions. The Content Specialist should indicate in the comments if any comments/suggestions from the reviewers were not approved and incorporated. The Content Specialist may choose one of the following options:

- Send the item to **Step 12** (Production) if there are revisions required that are outside the technical scope of the Content Specialist.
- Send the item to **Step 13** (Grammar Review).
- Send it back to earlier stages of review if major revisions are made.
- Delete the item.

Step 12: Production Edits
Items needing revisions outside the technical scope of the Content Specialist (such as artwork, graphs, and ELA selections) are revised by Production staff. Once the item is revised by Production staff, it is sent back to **Step 11** for review by a Content Specialist.

Step 13: Grammar Review
Editing staff reviews the item for grammatical issues. If the item had previously been sent back to S**tep 11** by Editing, the editor should check that the suggested revisions were addressed.

- If the editor suggests revisions to the item, the item will move back to **Step 11** for review by a Content Specialist.
- If the editor approves the item as is, the item proceeds to **Step 14** (Security Check).

Step 14: Security Check
Production staff checks to make sure no duplicate copy of the item exists in the test development databases. If there is a duplicate copy of the item or a requested revision was not made, then the item is flagged and sent back to **Step 11**.

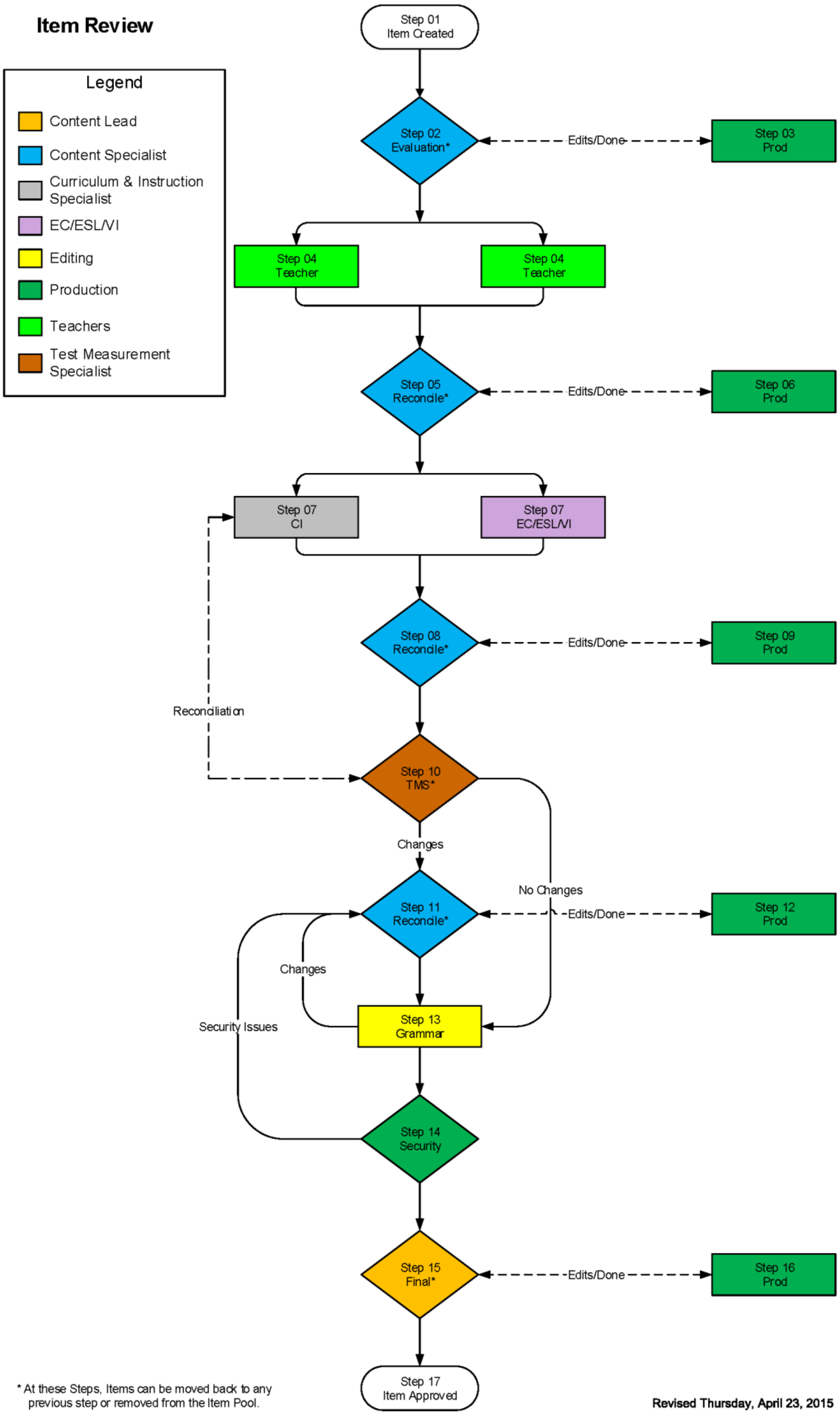
Step 15: Final Approval
The Content Lead reviews the item comment history to ensure all comments have been addressed and makes any final necessary revisions. . The Content Lead may choose one of the following options:

- Send the item to **Step 16** (Production) if there are revisions required that are outside the technical scope of the Content Lead.
- Approve the item and move it to **Step 17** (Item Approved).
- Send it back to **Step 2** if major revisions are made.
- Delete the item.

Step 16: Production Edits
Items needing revisions outside the technical scope of the Content Lead (such as artwork, graphs, and ELA selections) are revised by Production staff. Once the item is revised by Production staff, it is sent back to **Step 15** for review by the Content Lead.

Step 17: Item Approved
The item is now ready for placement on a form.

Item Review



* At these Steps, Items can be moved back to any previous step or removed from the Item Pool.

Revised Thursday, April 23, 2015

Selection Review Process

Prior to Step 1, the English Language Arts Content Specialist searches for appropriate selections for each assigned grade using criteria from Test Development staff, NCDPI-Curriculum and Instruction staff, and the North Carolina *Standard Course of Study.* The ELA Content Specialist also reviews the selections for any bias and sensitivity concerns.

———————————————Offline———————————————

Step 1: Folder Created

The Content Specialist creates a folder (color-coded by genre) for the selection. A Selection Form Submission slip is completed with the necessary copyright information (Content Specialist's name, date, title, author, source, excerpts, photographs, etc., as well as copyright date and ISBN, if applicable and the selection's readability score), and is attached to the inside of the folder. Any suggested edits are noted on the selection. A selection routing sheet is attached (includes grade level and title of selection) to the outside of the folder.

Step 2: Copyright Approval & Title/Author Search

Editing staff:

- determine if the selection is public domain, gratis, or copyrighted (if copyrighted, determine whether the publisher may be used or if there is a problem, such as excessive expense).
- search all selection databases to determine if the selection is already in use.

Step 3: Content Approval

The Content Lead evaluates the selection in terms of:

- alignment to grade-level expectations
- content and length of the selection
- readability of the selection
- bias or sensitivity concerns
- issues brought up by copyright review

Based on review, the Content Lead can:

- approve the selection as is
- approve the selection with edits or additions (including edits to or addition of artwork); the Content Lead sends a new copy to the Copyright Staff so they can seek permission from the publisher if copyrighted
- delete the selection

Step 4: Exceptional Children (EC), English as a Second Language (ESL), and Visually Impaired (VI) Review

The EC/ESL/VI reviewer evaluates the selection for accessibility concerns for EC, ESL, and VI students in terms of:

- concerns due to bias or insensitivity issues, such as contexts that might elicit an emotional response and inhibit students' ability to respond and contexts that students may be unfamiliar with for cultural or socio-economic reasons
- accessibility of graphics for students with or without vision
- appropriateness for Brailling
- prior knowledge required to understand the selection
- unfamiliar vocabulary that cannot be understood from the surrounding context

Based on review, the EC/ESL/VI reviewer can recommend:

- use the selection
- use the selection with suggested edits
- not use the selection

Step 5: Test Measurement Specialist Review

The Test Measurement Specialist (TMS) evaluates the selection in terms of:

- alignment to grade-level expectations
- content and length of the selection
- readability of the selection
- bias or sensitivity concerns

The TMS also evaluates:

- any bias or sensitivity concerns raised by the EC/ESL/VI reviewer
- edits made by content at **Steps 1 and 3**, or edits suggested in the **Step 4** review

If the TMS rejects the selection, it is deleted from the pool. If the TMS approves the selection, then it moves to **Step 6**.

Step 6: Prepare for online

Any issues noted in EC/ESL/VI and TMS reviews are reconciled by a Content Specialist, and selection is sent to production to enter into the online test development system. *NOTE:* If any edits or additions are made to the selection (including edits to or addition of artwork), the Content Specialist sends a new copy to the Copyright Staff so they can seek permission from the publisher if copyrighted.

————————————In Online Test Development System————————————

Step 1: Selection Created
Production staff enters the selection into the test development system.

Step 2: Compare Original
Editing staff compares the original copy of the selection to what has been entered into the test development system and indicates any necessary corrections. The corrections may arise from discrepancies between the TDS and the original or from correctable errors in the original, such as grammatical errors, misspellings, or archaic/foreign spelling of words.

Step 3: Creation Reconcile
A Content Specialist resolves corrections indicated in **Step 2**. The Specialist indicates in the comments if any comments/suggestions from Editing staff were not approved and incorporated.

Step 4: Creation Edits
Production makes requested changes and selection is sent back to **Step 3** for a Content Specialist to confirm requested changes have been made.

Step 5: NCDPI-Curriculum and Instruction Review
A Curriculum and Instruction Specialist reviews the selection. The reviewer evaluates the selection in terms of:

- alignment to grade-level expectations
- content and length of the selection
- readability of the selection
- bias or sensitivity concerns

The Curriculum and Instruction Specialist rates the selection as acceptable, acceptable with revisions, or unacceptable. The Specialist can also include additional comments.

Step 6: Test Measurement Specialist Review
The TMS does a final review on the selection and reviews all comments from the Curriculum and Instruction Specialist. The TMS either approves the selection (with comments regarding revisions, if any) or deletes the selection from the pool.

Step 7: Reconcile Curriculum and Instruction Review and Test and Measurement Specialist Review
A Content Specialist reviews any comments/changes requested by Curriculum and Instruction or by the Test and Measurement Specialist, and sends changes to **Step 8** (Production) to be made if necessary. Once any changes are made, the selection is sent to **Step 9**.

NOTE: If any edits or additions are made to the selection (including edits to or addition of artwork), the Content Specialist sends a new copy to the Copyright Staff so permission may be sought from the publisher if copyrighted.

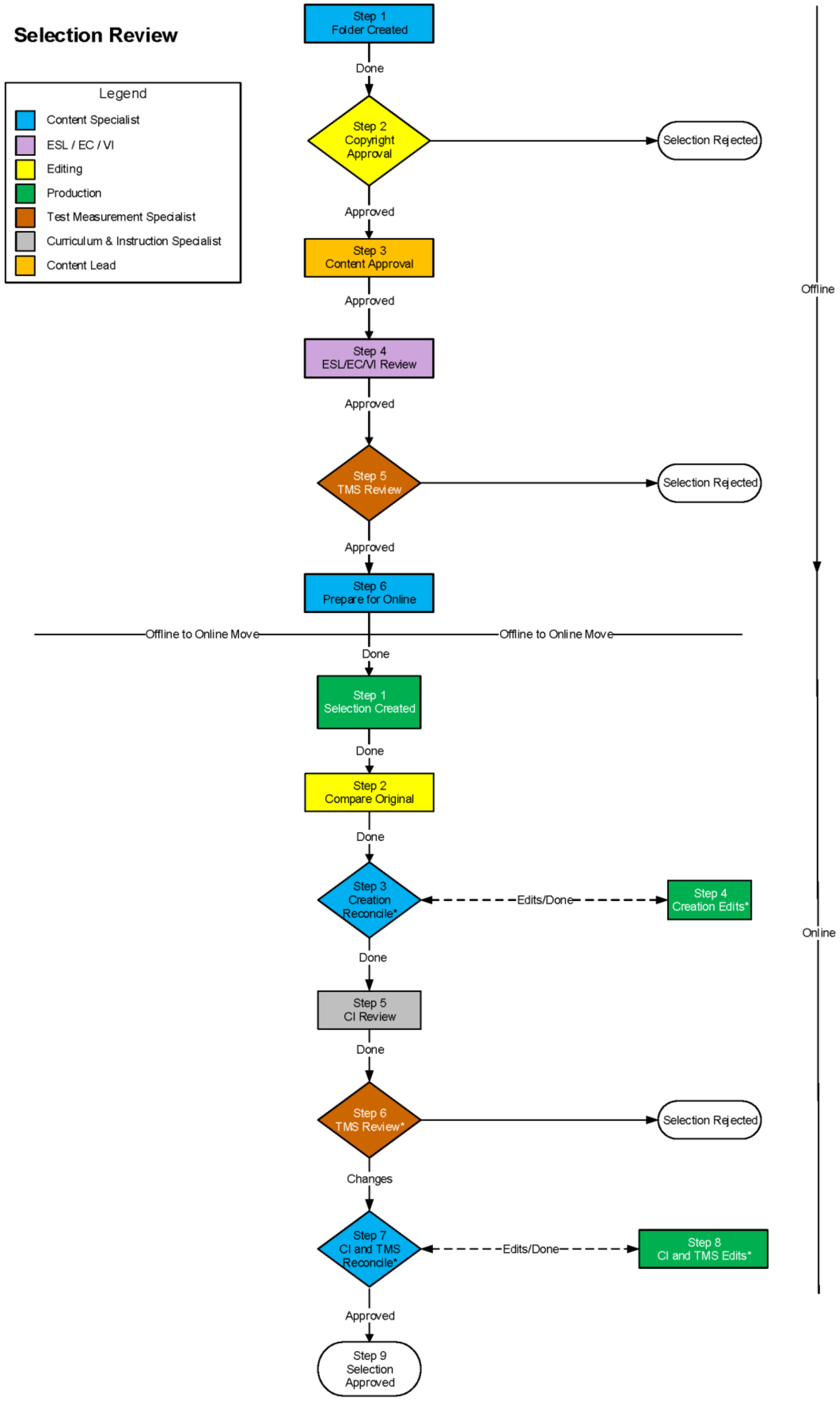
Step 8: Production Edits

Production makes requested changes and selection is sent back to **Step 7** for a Content Specialist to confirm requested changes have been made.

Step 9: Selection Approved

Selection is now ready to have items written.

Selection Review



* At these Steps, Selections can be moved back to any previous step or removed from the Selection Pool.

Operational Base Form Review Process

Prior to Step 1, a Psychometrician chooses the test items for the initial placement of the preliminary base form, taking key balance into consideration.

Step 1: Ordered Item Numbers Supplied
A psychometrician creates the form, and uploads a file listing the Item IDs to populate the form. The form is sent to **Step 3** for form review. Forms can come back to this step from **Step 3** with suggestions for replacements, or from **Step 4** with suggestions for replacements or revisions (either the content of the item or for key issues). The Psychometrician can replace items or incorporate revisions. The Psychometrician sends the form to **Step 2** (Production Edits) for revisions to artwork, graphs, or ELA selections. After any revision, the Psychometrician sends the form back to **Step 3**.

Step 2: Production Edits
Revisions to operational items such as artwork, graphs, and ELA selections are made by Production staff. If any revisions are made, the form is sent back to **Step 1** for review by a Psychometrician.

Step 3: Form Review
A Content Specialist reviews:

- the items on the form for content alignment and quality of content, and
- the form for conflicts or repetition of content.

If any items are replaced due to concerns regarding conflicts or repetition of content among items, or for quality concerns, the Content Specialist sends the form back to **Step 1** with comments for the psychometrician. Otherwise, the form is sent to **Step 4** for Test Measurement Specialist Review.

Step 4: Test Measurement Specialist Review/Key Balance
This review step is conducted to ensure that the form is ready for Outside Content Key Check (i.e., the form is ready to send to printer).

- This review covers both item and form level quality.
- The Test and Measurement Specialist (TMS) reviews each item, including any comments. Suggestions for revisions to items are made as needed.
- After reviewing the quality of each item, the form is evaluated in terms of cueing, repetition, content coverage, and balance across Depths of Knowledge or Knowledge Types/Cognitive Processes.
- The key balance of the form is checked. If the key balance needs adjusting, these suggestions are made by the TMS and submitted to the Test Development Section Chief who has to approve/disapprove and the form is returned to **Step 1**.

After reviewing each item, the TMS can add form-level comments and suggested improvements, and can:

- send the form back to **Step 1** with suggestions for replacements or revisions,
- move the form to **Step 5** (Reconcile), or
- delete the form from the pool.

Step 5: Reconcile
At this step, the form is sent for Outside Content Key Check. The Content Specialist reviews the form comments to ensure any suggested replacements or revisions have been addressed, and that any approved replacements or revisions have been made correctly. If any replacements or revisions need adjusting, the Content Specialist moves the form back to **Step 1** with comments. Otherwise, the form moves to **Step 6** (Outside Content Key Check).

Step 6: Outside Content Specialist Key Check
An Outside Content Specialist reviews the form by answering each item and providing any comments and/or suggestions. This review is done on-site.

Step 7: Reconcile Outside Content Review
A Content Specialist checks the keyed response from the Outside Content Review against the key for each item, and reviews all comments and/or suggestions from the Outside Content Expert. Any key disagreements are reconciled, and any comments and/or suggestions from the Outside Content Specialist are addressed.

Step 8: Psychometric Review/Key Balance
A Psychometrician:

- reviews comments/suggestions from the Outside Content Specialist and from Editing staff, with consultation with the TMS and Content Specialists.
- checks key agreement with the Outside Content Specialist and resolves any disagreements through consultation with the TMS and Content Specialists.
- makes any approved revisions, or indicates revisions for Production staff to make, and sends the form to **Step 9** (Production Edits).
- re-uploads the form if any items are replaced.

Step 9: Production Edits
Revisions to items outside the technical scope of the Psychometrician (items such as artwork, graphs, and ELA selections) are made by Production staff. Once the revisions are made, the form is sent back to **Step 8** for review by a Psychometrician.

Step 10: Grammar Review
Two editors independently review the form for grammatical and/or formatting issues, providing comments and/or suggestions as needed.

Step 11: Content Lead Review/Finalize Form

A Content Lead reviews the base form and reviews all comments from editing staff and addresses any suggestions. The Content Lead reviews the form comment history to ensure all comments have been addressed. After reviewing the form, the Content Lead either:

- approves the form, and moves it to **Step 12** (Item Placement). The form is cloned when the Content Lead approves the form, so all the needed versions of the base form will be at **Step 12** for item placement.
- moves the form back to **Step 8** if any edits to operational items need review.

Step 12: Item Placement

A Content Specialist places approved items in the embedding slots. The Content Specialist needs to check:

- the placed items match the layout files for the version of the base form
- the quality of items embedded for experimental use
- the items do not cue operational items or other embedded items
- the keys of the embedded items do not create an unbalanced key for the overall form
- as a group, the items' difficulty and Depth of Knowledge or Knowledge Type/Cognitive Process are consistent with the surrounding base form.

After placing the items, the Content Specialist may choose one of the following options:

- Send the form to **Step 13** (Production Edits) for revisions to artwork, graphs, or ELA selections.
- Send the form to **Step 14** (Cueing Check).
- Delete the form.

Step 13: Production Edits

Revisions to embedded experimental items such as artwork, graphs, and ELA selections are made by Production staff. Once the revisions are made, the form is sent back to **Step 12** for review by a Content Specialist.

Step 14: Cueing Check

The Content Specialist and TMS review the entire form to check that the embedded items do not create cueing or repetition issues, and that the embedded items' quality is acceptable. The TMS also should make sure the key balance is adequate. After the review, the Content Specialist can replace or revise embedded items based on the review. Then the Content Specialist moves the form to **Step 15** for Outside Content/Grammar check.

Step 15: Outside Content Specialist Key Check and Grammar Check

An Outside Content Specialist and Editing staff member each review the embedded items. The Outside Content Specialist reviews the embedded items by working and answering each item and providing any comments or suggestions as needed; Editing staff reviews the items for any grammatical and/or formatting issues, providing comments and/or suggestions as needed.

Step 16: Reconcile
A Content Specialist checks the keyed response from the Outside Content Review against the key for each item, and reviews all comments and/or suggestions from the Outside Content Expert. Any key disagreements are reconciled, and any comments and/or suggestions from the Outside Content Expert are addressed. The Content Specialist also reviews suggestions from Editing Staff, and makes any necessary revisions.
If any items require substantial revisions, the item should be replaced, and the form sent back to **Step 15**.

The Content Specialist can:

- send the form to **Step 17** (Production Edits) for revisions to artwork, graphs, or ELA selections,
- send the form to **Step 18** (TMS Final Review), or
- delete the form.

Step 17: Production Edits
Revisions to embedded experimental items such as artwork, graphs, and ELA selections are made by Production staff. Once the revisions are made, the form is sent back to **Step 16** for review by a Content Specialist.

Step 18: Test Measurement Specialist Final Review
The TMS reviews the form, considering the comments from the **Step 15** reviews to ensure all comments have been addressed properly. The key balance of the form is checked. The TMS makes any needed edits to items. Then the TMS sends the form to **Step 20** (Final Grammar).

Step 19: Production Edits
Revisions to operational items such as artwork, graphs, and ELA selections are made by Production staff. Once the revisions are made, the form is sent back to **Step 18** for review by the TMS.

Step 20: Final Grammar Review
An Editor reviews the entire form for grammatical and/or formatting issues, providing comments and/or suggestions as needed.

Step 21: Final Manager Review
A Content Manager reviews comments/suggestions from the Final Grammar Review or **Step 24** (Compare) and makes any necessary revisions to embedded items. The Manager checks the form for overall quality and reviews the form comment history to ensure all comments have been addressed.

After reviewing the form, the Content Manager may choose one of the following options:

- Approve the form and send it to **Step 23** (Audio Approval) if the form will be administered online,
- Approve the form and send it to **Step 24** (Compare) if the form will be administered on paper,
- Send the form to **Step 20** (Psychometrician) if there are suggested revisions to operational items for the Psychometrician to consider.
- Send the form to **Step 22** (Production Edits) for revisions to artwork, graphs, or ELA selections.
- Reject the form.

Step 22: Production Edits
Revisions to embedded experimental items such as artwork, graphs, and ELA selections are made by Production staff. Once the revisions are made, the form is sent back to **Step 21** for review by a Content Manager.

Step 23: Audio Approval
A Content Specialist reviews the audio for each item and either approves the audio or indicates it needs correction. After all items' audio have been approved, the form is sent to **Step 24** (PDF/Online Check).

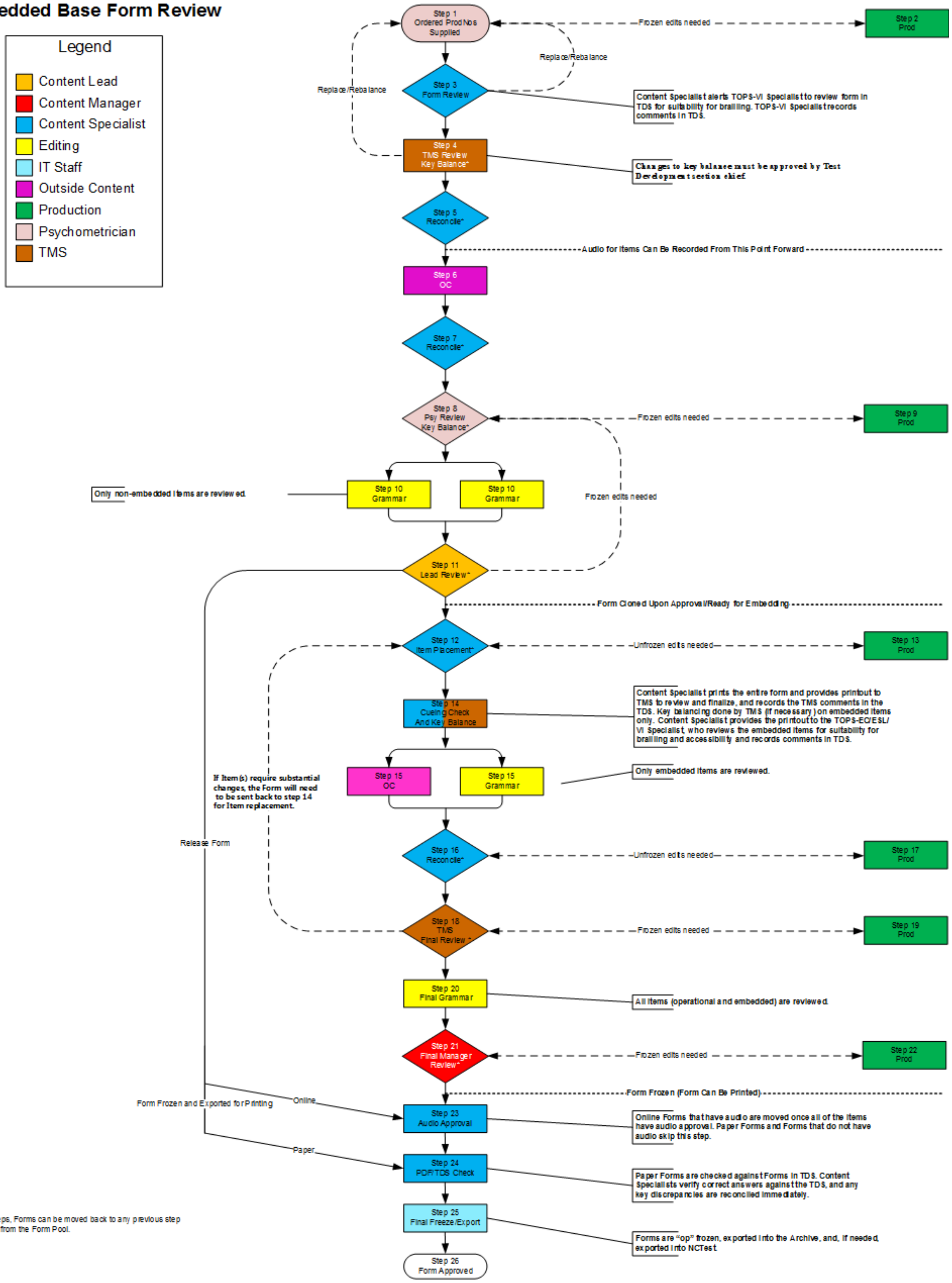
Step 24: PDF/Online Check
At this step, Production staff exports the form as a document and formats the document per formatting guidelines. The form is placed in a folder with a signoff sheet.

- Two Editors review the form for formatting concerns as well as any grammatical issues.
- A Content Specialist reviews the form for content and evaluates any comments and or suggestions from Editing reviews. If there are any edits to embedded items to execute in the online test development system, the Content Specialist indicates with each item what edits are approved and sends the form back to **Step 21**. Any suggestions that are rejected should be noted in the form comments.
  Any suggested edits to operational items that Content staff feel warrant consideration are directed to the TMS and Psychometrician for consideration.
- A Content Manager makes any approved edits in the online test development system and sends the form to **Step 23** for online forms or **Step 24** for paper forms.
- After production staff makes corrections to the paper copy, the file is converted to a PDF and printed. The printed copy undergoes the same review as bullets 1–3 above.
- After the PDF of the form is approved, the form is sent to **Step 25** (Final Freeze/Export). If the forms are also offered online, the online forms will be sent to **Step 25**.

Step 25: Final Export
The form, all items, and any selections are operationally locked to prevent any revisions. This is to ensure that the published versions of the form, items, and selections are preserved electronically. Any online forms undergo checks in a variety of platforms to ensure that each item's content displays correctly, and audio files for non-ELA subjects read correctly.

Step 26: Form Approved
The form is approved for administration.

EOC/EOG
Embedded Base Form Review


Legend
🟧 Content Lead
🟥 Content Manager
🟦 Content Specialist
🟨 Editing
IT Staff
🟪 Outside Content
🟩 Production
Psychometrician
🟫 TMS
Step 1
Ordered Prod Nos
Supplied
Frozen edits needed
Step 2
Prod
Replace/Rebalance
Replace/Rebalance
Step 3
Form Review
Content Specialist alerts TOPS-VI Specialist to review form in TDS for suitability for brailling. TOPS-VI Specialist records comments in TDS.
Step 4
TMS Review
Key Balance*
Changes to key balance must be approved by Test Development section chief.
Step 5
Reconcile*
Audio for Items Can Be Recorded From This Point Forward
Step 6
OC
Step 7
Reconcile*
Step 8
Psy Review
Key Balance*
Frozen edits needed
Step 9
Prod
Only non-embedded items are reviewed.
Step 10
Grammar
Step 10
Grammar
Frozen edits needed
Step 11
Lead Review*
Form Cloned Upon Approval/Ready for Embedding
Step 12
Item Placement*
Unfrozen edits needed
Step 13
Prod
Step 14
Cueing Check
And Key Balance
Content Specialist prints the entire form and provides printout to TMS to review and finalize, and records the TMS comments in the TDS. Key balancing done by TMS (if necessary) on embedded items only. Content Specialist provides the printout to the TOPS-EC/ESL/VI Specialist, who reviews the embedded items for suitability for brailling and accessibility and records comments in TDS.
If Item(s) require substantial changes, the Form will need to be sent back to step 14 for Item replacement.
Step 15
OC
Step 15
Grammar
Only embedded items are reviewed.
Release Form
Step 16
Reconcile*
Unfrozen edits needed
Step 17
Prod
Step 18
TMS
Final Review *
Frozen edits needed
Step 19
Prod
Step 20
Final Grammar
All items (operational and embedded) are reviewed.
Step 21
Final Manager
Review*
Frozen edits needed
Step 22
Prod
Form Frozen (Form Can Be Printed)
Form Frozen and Exported for Printing
Online
Paper
Step 23
Audio Approval
Online Forms that have audio are moved once all of the items have audio approval. Paper Forms and Forms that do not have audio skip this step.
Step 24
PDF/TDS Check
Paper Forms are checked against Forms in TDS. Content Specialists verify correct answers against the TDS, and any key discrepancies are reconciled immediately.
Step 25
Final Freeze/Export
Forms are "op" frozen, exported into the Archive, and, if needed, exported into NCTest.
Step 26
Form Approved

* At these Steps, Forms can be moved back to any previous step or removed from the Form Pool.

APPENDIX 4-C TIF AND CSE PLOTS

Figure 1. EOG Math Grade 3 Test Information with Associated Standard Errors

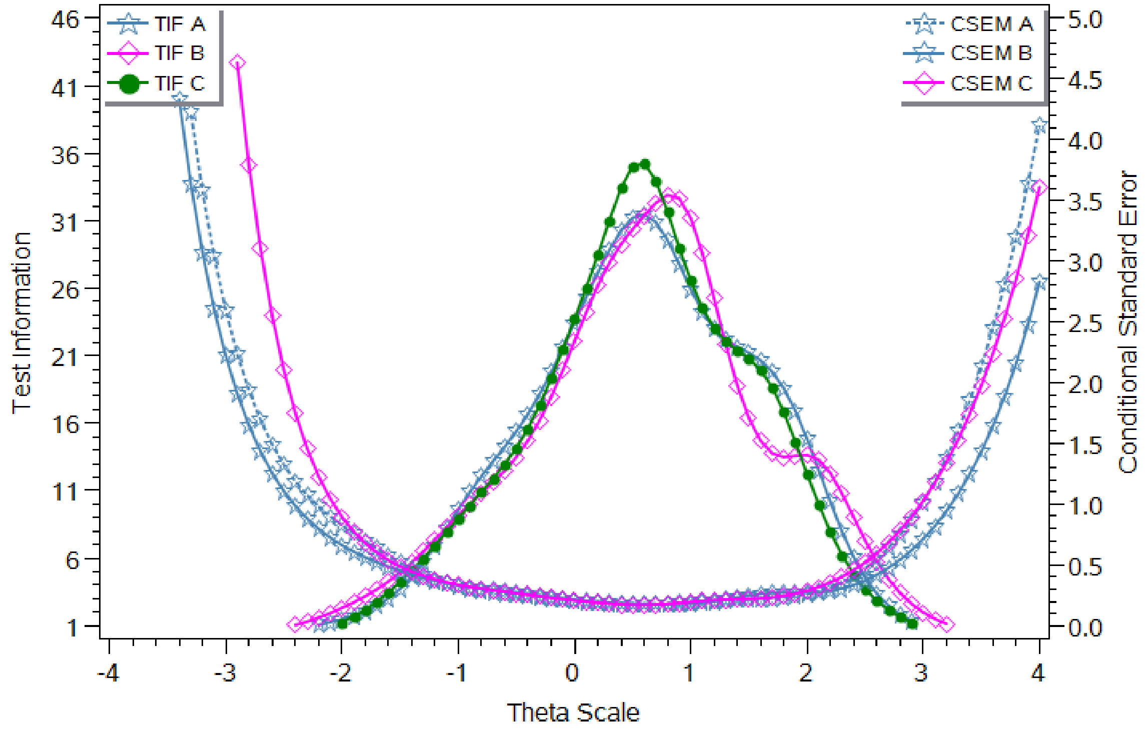


Figure 2. EOG Math Grade 4 Test Information with Associated Standard Errors

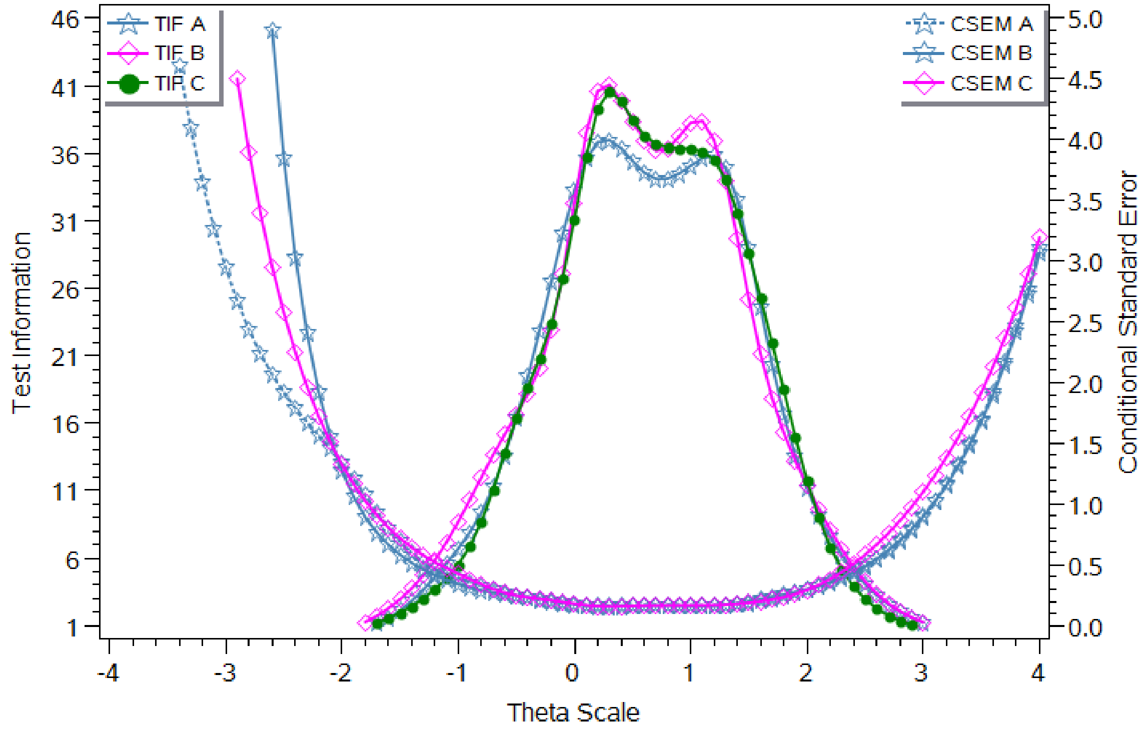


Figure 3. EOG Math Grade 5 Test Information with Associated Standard Errors

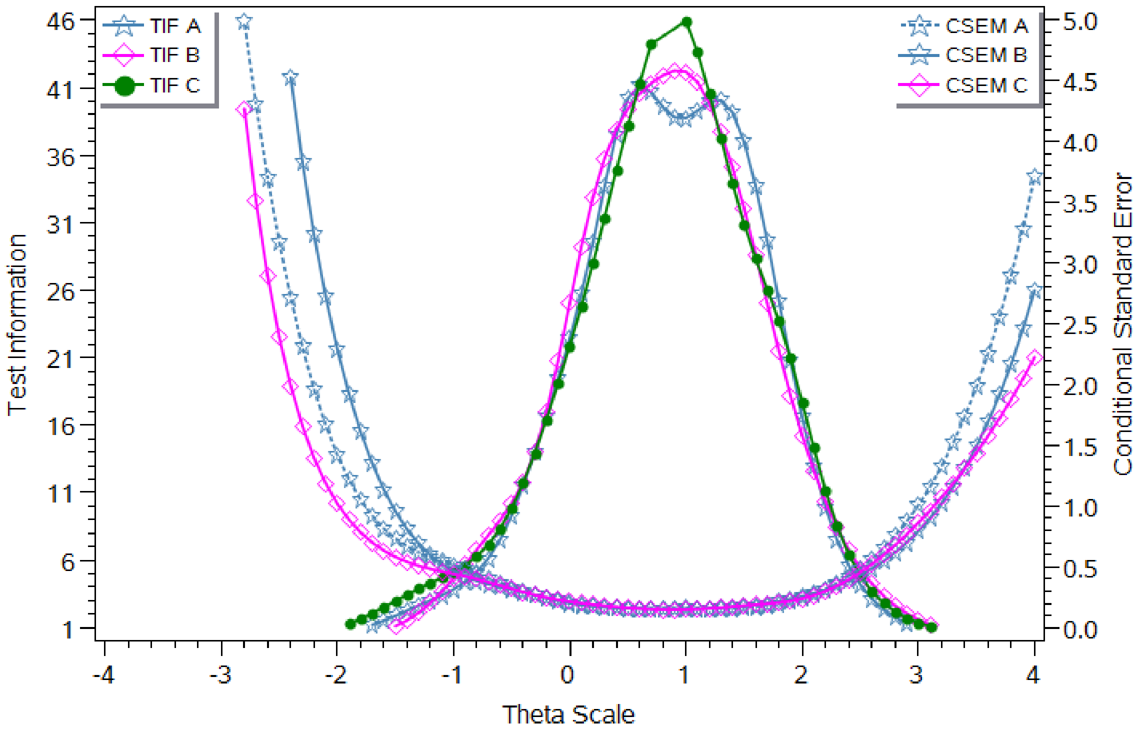


Figure 4. EOG Math Grade 6 Test Information with Associated Standard Errors

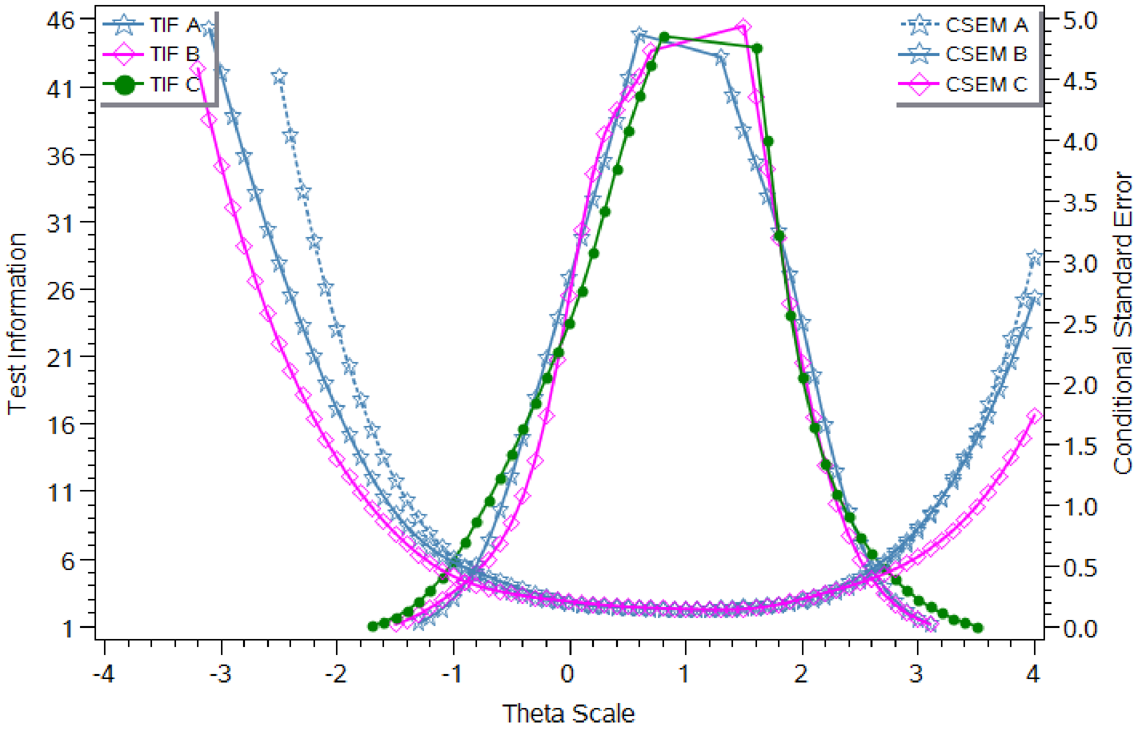


Figure 5. EOG Math Grade 7 Test Information with Associated Standard Errors

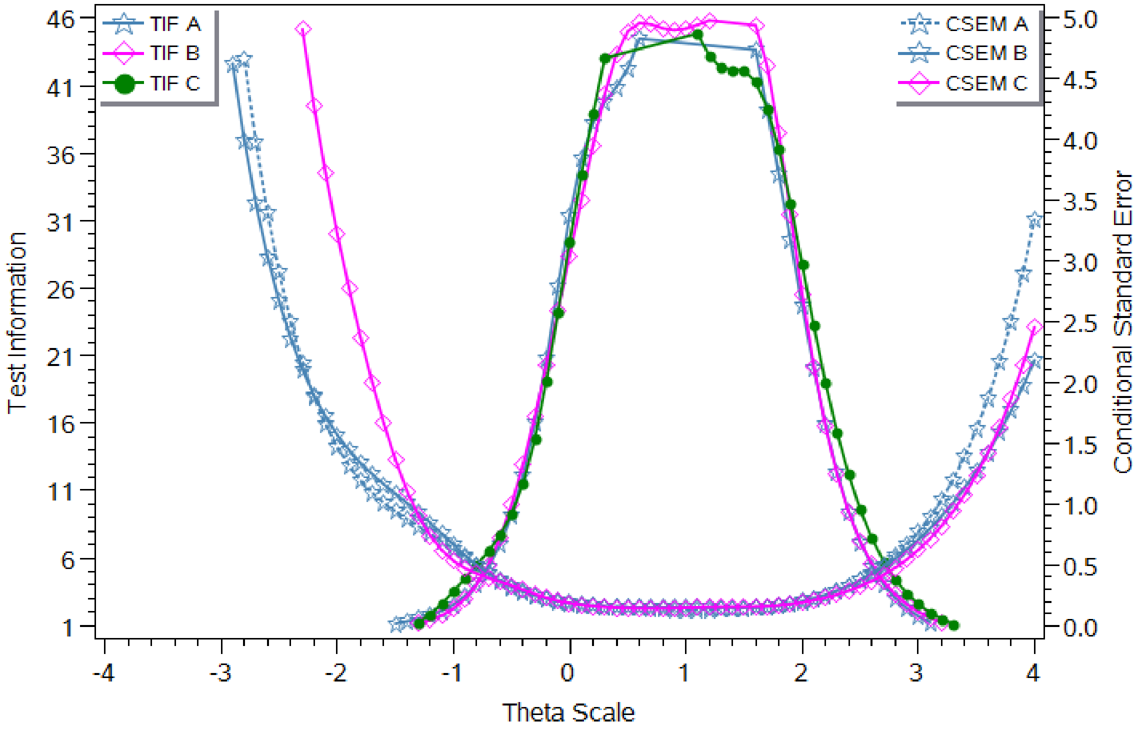


Figure 6. EOG Math Grade 8 Test Information with Associated Standard Errors

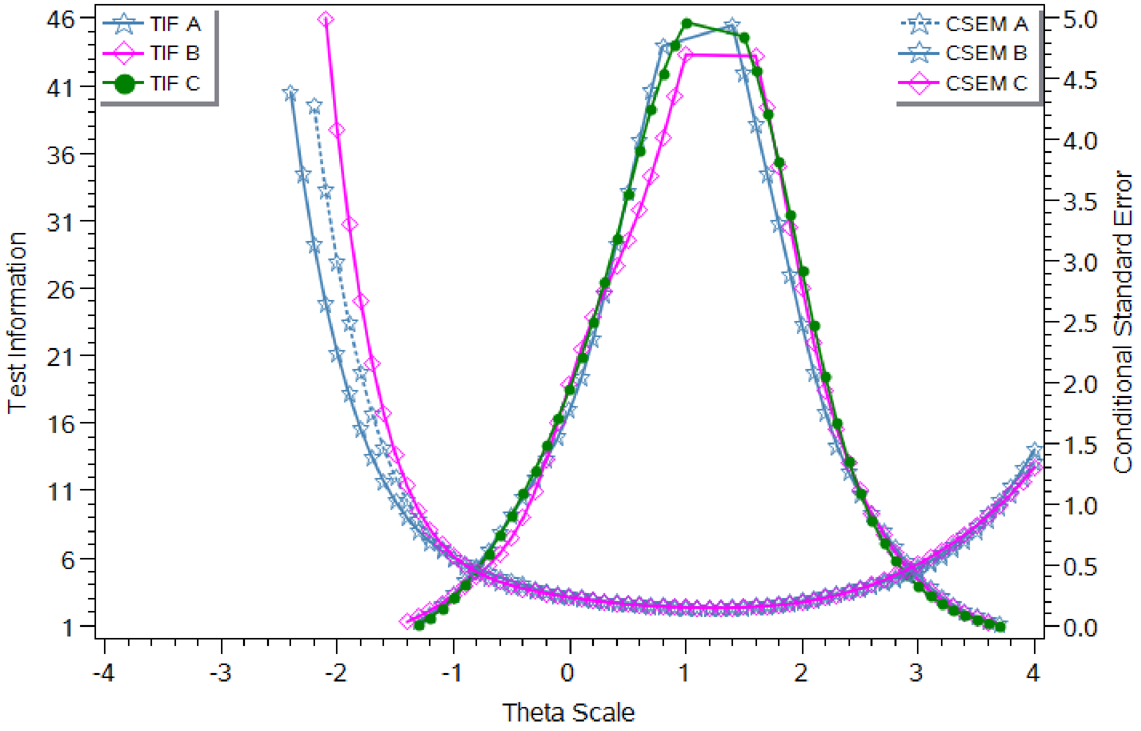
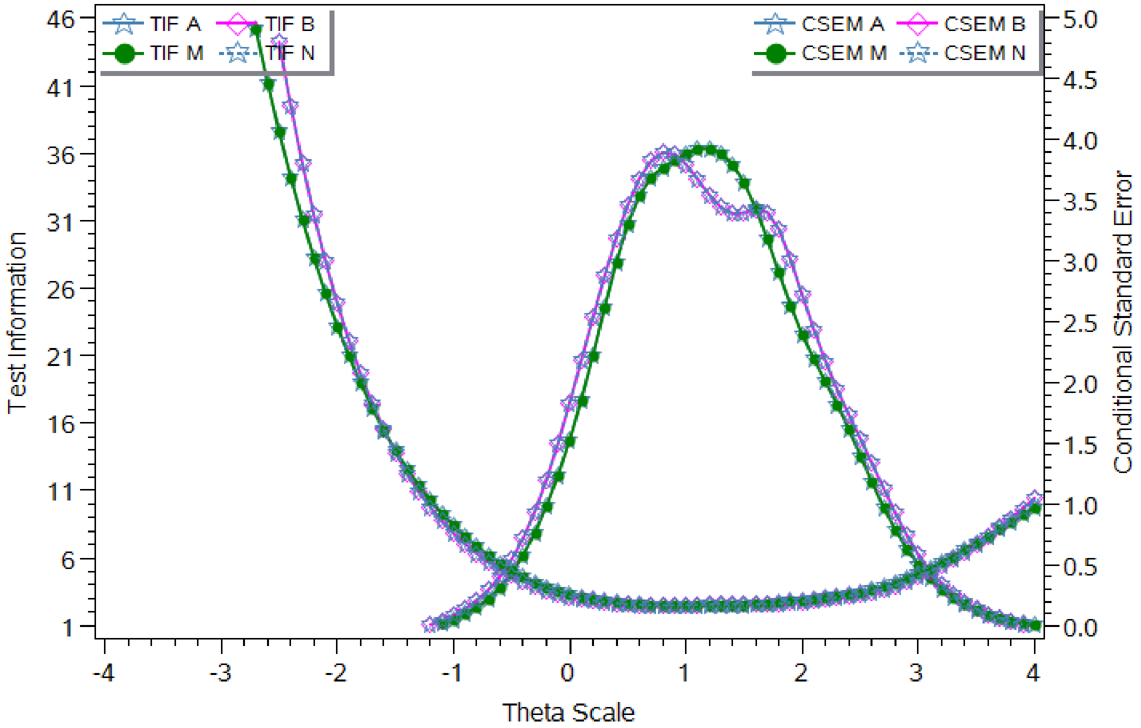


Figure 7. EOC Math I Test Information with Associated Standard Errors



North Carolina Testing Program

Standard Setting
for the
End-of-Course and End-of-Grade Assessments

Executive Summary

J. Michael Clark III, Ph.D.
Stephen T. Murphy, Ph.D.

August 24, 2013

E ecutive Summary

Introduction

Nine committees of North Carolina educators convened to make cut score recommendations for the End-of-Grade (EOG) assessments for Grades 3-8 Mathematics, Grades 3-8 Reading, and Grades 5 and 8 Science and for the End-of-Course (EOC) assessments for Biology, English II, and Mathematics I. A total of 164 North Carolina educators convened in Chapel Hill, North Carolina between July 22 and July 26, 2013, using the item mapping method to make content-oriented recommendations for cut scores. A brief summary of the outcomes of this workshop are provided in this executive summary, and a more detailed account of the workshop is provided in the full standard setting technical report.

Panelists

All panelists were asked to provide voluntary demographic information. A brief summary of panelist characteristics is provided in this executive summary. Complete panelist demographics are provided in the full standard setting technical report.

The panelists years of experience as educators are summarized in Table 1. As illustrated by this table, participants in this standard setting had a wide range of teaching experience.

Table Panelist E perience

Panel	N	Years in urrent Position					
							NR
Mathematics 3-5	20	1	4	8	2	4	1
Mathematics 6-8	16	2	3	4	5	2	0
Reading 3-5	18	1	3	5	1	8	0
Reading 6-8	19	2	2	6	6	3	0
Science 5	16	1	5	5	5	0	0
Science 8	17	3	6	5	1	2	0
Biology	20	2	5	6	4	3	0
English II	17	3	5	5	2	1	1
Mathematics I	21	4	3	5	2	7	0

Note: NR = No Response.

The panelists professional backgrounds are summarized in Table 2 and Table 3. As will be described in greater detail in a subsequent section of this executive summary, panelists summarized in Table 2 made cut score recommendations for three grade levels within a particular subject area. Individuals reported as teaching in lower, middle, or upper grades are reported in the context of their committee. For example, a lower-grade panelist in the Mathematics 3-5 panel teaches Grade 3 Mathematics, while a lower-grade panelist in the Reading 6-8 panel teaches Grade 6 Reading. Panelists who reported teaching more than one grade level within the subject area are listed under the multiple grades column, and panelists who primarily teach a grade level outside of the panel s range (e.g., a Grade 2 teacher who participated in the Mathematics 3-5 panel) are listed in the off-grade column. Finally, other groups of educators are summarized in the remaining columns of this table. As shown in this table, all grade levels were represented on these panels, and a variety of professional backgrounds was represented on these panels.

Table Panelist Pro essional Bac ground Three rade Panels

Panel	LO	MID	UP	MUL	O	SED	SPE	OA	NS	OT
Mathematics 3-5	3	6	5	2	1	0	2	1	0	0
Mathematics 6-8	7	3	3	1	0	1	1	0	0	0
Reading 3-5	3	1	4	3	1	0	4	0	1	1
Reading 6-8	4	5	3	2	0	3	0	0	0	2

Note: LOW = lower grade, MID = middle grade, UP = upper grade, MUL = multiple grades, OFF = off-grade, SED = special education, SPE = specialist, COA = coach, GNS = grade level not specified, OTH = other.

Panelists summarized in Table 3 recommended cut scores for a single grade and/or subject. Panelists listed in the on-grade column actively teach in the grade/subject for which standards are being set. Panelists summarized in the off-grade column teach in a related subject area, but at a different grade level. Other types of professional backgrounds are summarized to the right of these columns in the table. As shown in this table, the majority of each panel was comprised of individuals who teach the grade/subject of interest, but each showed diversity in panelists professional backgrounds as well.

Table Panelist Pro essional Bac ground Single rade Panels

Panel	ON	O	SED	SPE	OA	ED	OT	RET	NR
Science 5	7	4	0	1	0	0	4	0	0
Science 8	11	4	1	0	0	0	0	0	1
Biology	17	0	0	1	0	1	0	1	0
English II	11	2	1	0	0	2	1	0	0
Mathematics I	15	2	0	1	1	0	1	0	1

Note: ON = on-grade, OFF = off-grade, SED = special education, SPE = specialist, COA = coach, HED = higher education, OTH = other, RET = retired, NR = no response.

Table 4 contains a summary of panelists gender and ethnicity, and Table 5 summarizes panelists geographic regions within the state. As these two tables illustrate, panels generally were representatively diverse in terms of gender, ethnicity, and geographic region.

Table Panelist ender and Ethnicity

	ender			Ethnicity						
Panel		M	NR	AA	AS	I	NA		MU	NR
Mathematics 3-5	18	2	0	7	0	0	0	12	0	1
Mathematics 6-8	11	5	0	3	0	1	0	12	0	0
Reading 3-5	17	1	0	7	1	1	1	6	2	0
Reading 6-8	18	1	0	4	0	0	1	14	0	0
Science 5	16	0	0	4	0	0	0	12	0	0
Science 8	13	4	0	0	1	1	1	13	1	0
Biology	17	3	0	1	0	1	0	18	0	0
English II	14	3	0	1	0	2	0	14	0	0
Mathematics I	20	1	0	3	0	1	0	17	0	0

Note: F = female, M = male, NR = no response, AA = African American, AS = Asian, HI = Hispanic, NA = Native American, WH = white, MU = multiple responses.

Table   Panelist   eographic Region

Panel		N	NE	N	S	SE	S		MU	NR
Mathematics 3-5	4	1	0	1	4	4	5	1	0	0
Mathematics 6-8	1	2	1	1	2	3	4	2	0	0
Reading 3-5	2	1	1	0	4	3	4	2	0	1
Reading 6-8	0	1	1	4	2	5	5	0	1	0
Science 5	4	2	0	0	2	1	5	2	0	0
Science 8	5	1	1	1	4	2	2	1	0	0
Biology	3	4	1	3	1	2	5	0	0	1
English II	4	0	1	3	4	2	2	1	0	0
Mathematics I	6	2	0	3	4	0	6	0	0	0

Note: C = central, NC = north central, NE = northeastern, NW = northwestern, SC = south central, SE = southeastern, SW = southwestern, W = western, NR = no response.

Method and Procedure

A total of nine panels set standards for 17 grades and subjects. Panels were divided into two groups. Panelists setting standards for Mathematics or Reading for grades 3-8 each worked on three adjacent grade levels (3-5 or 6-8). These panels are referred to in this executive summary as three-grade panels. For the remaining grades and subjects   Grades 5 and 8 Science, Biology, English II, and Mathematics I   panelists set standards for a single grade/subject. These are referred to as single-grade panels. Although all nine panels used a similar methodology for panelists to render their judgments, the scope of activities varied across panel types. The three-grade panels convened between July 22-26, 2013, while the single-grade panels convened between July 24-25, 2013. The agenda for the single-grade panels is provided in Appendix A, and the agenda for the three-grade panels is provided in Appendix B.

On the morning of Monday, July 22, prior to the standard setting workshop, training was held for table leaders for the three-grade panels. For the single-grade panels, table leader training was held during the morning of Wednesday, July 24. During this training session, table leaders were introduced to the standard setting facilitators, trained on their role in the standard setting process, and received a general introduction and instruction on the item mapping process. Following table leader training, representatives of the North Carolina Department of Public Instruction and Pearson presented an opening session to all panelists. The three-grade panel opening session occurred on July 22, and the single-grade opening session occurred on July 24. After the conclusion of the opening session, panelists dispersed to their breakout session meeting rooms. Each panel convened in a separate breakout session room to complete the required standard setting activities.

Following committee introductions, the three-grade panels spent the remainder of Monday, July 22 writing and discussing achievement level descriptors (ALDs), which serve as content-oriented statements describing expectations of student performance at each achievement level, for the three grade levels assigned to their panels. For the single-grade panels, a portion of July 24 was devoted to ALD writing for their single assigned assessment, and then the single-grade panels moved on to other standard setting activities that day.

Following ALD writing activities, panelists performed tasks to set standards for their assigned subject area and grade(s). Panelists began by writing just barely level descriptors: statements describing performance expectations for students who are *just barely* at the three cut points separating the four achievement levels. Next, panelists reviewed the ordered item book (OIB), which contains items from the previous administration s assessment as well as additional supplemental items selected from the item pool, ordered in ascending empirical difficulty as estimated from actual student performance, and presented such that each page of the booklet contains a single item.

The item mapping procedure (Lewis, Green, Mitzel, Baum, Patz, 1998 Mitzel, Lewis, Patz, Green, 2001) is the judgmental process that was used in this standard setting. According to this procedure, panelists are asked to identify the item in the ordered item book that is the last item that a student who is just barely at a given achievement level should be able to answer correctly more often than not. The locations for the items in the ordered item book were established using a guess-adjusted response probability of two-thirds (or 2/3), representing the point on the item characteristic curve at which the probability of a correct response is two-thirds of the way between the curve s lower asymptote and 1.0. Following item mapping methodology training by a Pearson breakout session facilitator and a practice round of judgment, panelists began the standard setting process. For the three-grade panels, standard setting activities began at the lower grade (i.e., grade 3 for the panels assigned to grades 3-5, grade 6 for panels assigned to grades 6-8). Panelists set three recommended cut scores, which separate student performance into four distinct achievement level categories.

The standard setting process consisted of three rounds of judgment. Panelists were provided with feedback data, which was intended to inform panelists decisions, to consider and discuss between each round. Following Round 1, panelists broke up into small groups of 5 to 7 and discussed their cut scores and associated rationales. Following small-group discussions, the entire panel shared their cut scores. For both discussions, panelists were instructed that reaching consensus was not the goal of these discussions, but rather, they should share their perspectives that led to their chosen cut scores.

In addition to the Round 1 cut score agreement data, panelists were shown external data to further inform their judgments in subsequent rounds of judgment. Panelists were provided with empirical item difficulty data showing the proportion of all test-takers from the spring 2013 administration who correctly answered each item (i.e., item *p*-values). The standard setting facilitator also shared with panelists the ACT Explore cut score, which was linked to the North Carolina assessment by NCDPI, representing the score point at which students are on-track to be college and career-ready. Finally, the facilitator shared with panelists the expected cut scores obtained by NCDPI from a recent survey of North Carolina educators. Following discussion of Round 1 cut scores and the provided feedback data, panelists proceeded to the second round of judgment.

Following Round 2, panelists received updated cut score agreement data and engaged in discussions in both small groups and across the entire panel. Additionally, panelists were shown a graphical display of student impact data. The impact data displayed the percentages of spring 2013 test-takers who would be classified into the four achievement levels based on the panel s median cut score recommendation. Impact was shown for the overall North Carolina test-taking population, and impact was also broken down by gender and ethnicity subgroups. Panelists were given an opportunity to discuss the appropriateness of their cut scores given the current impact data. Following discussion of the Round 2 feedback data, panelists proceeded to the third and final round of judgment.

Following Round 3, panelists were shown their final recommended cut scores, which were based on their median cut score judgments from this final round of ratings. Panelists were shown impact data, again illustrating overall impact as well as impact broken down by gender and ethnicity. After reviewing and discussing the Round 3 impact data, panelists completed an evaluation survey capturing their reactions to the final cut score recommendations and associated impact data. The results of the evaluation survey are documented in the full standard setting technical report.

The standard setting workshop activities concluded at this point for the single-grade committees. For the three-grade committees, the breakout session facilitator guided panelists through the same process for the middle and upper grades. Following the conclusion of standard setting activities, all panelists were dismissed with the exception of table leaders, who attended the vertical articulation session on Friday, July 26.

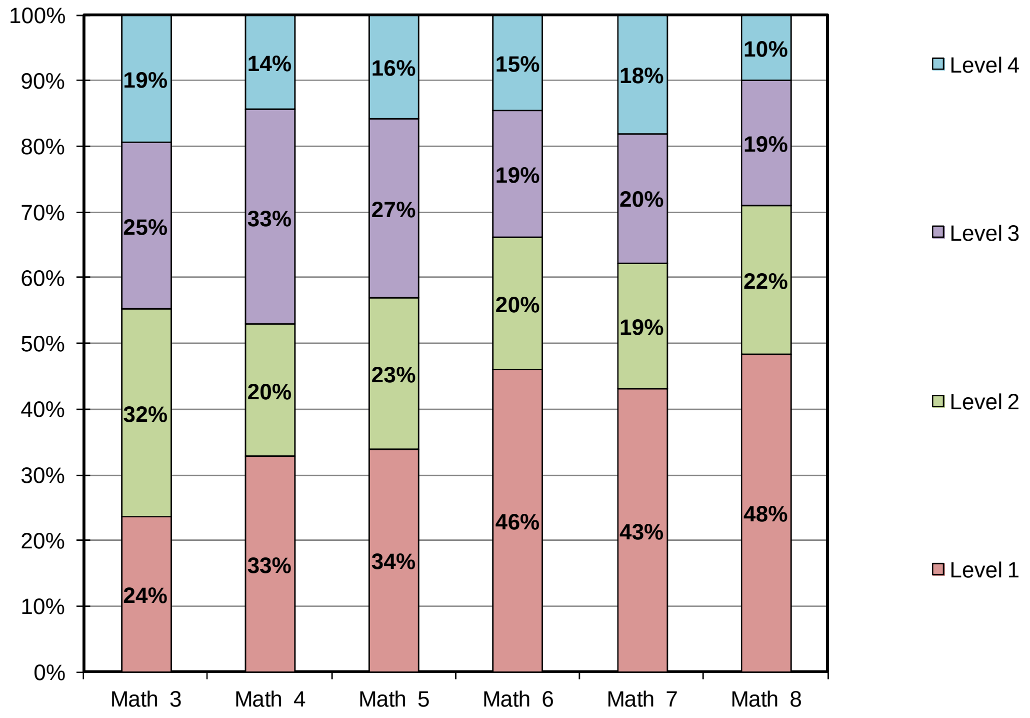
Table leaders from each committee convened in a single room to participate in the vertical articulation session. During this session, impact data were compared across grade levels within subject areas (e.g., Grades 3-8 Reading) and also across subjects. Panelists were asked to evaluate, from a policy perspective, the reasonableness of the committees content-oriented cut score recommendations and the impact of imposing these achievement expectations on student test scores. Panelists were guided through a process whereby they evaluated the reasonableness of impact for particular grades/subjects, both in isolation and in contrast to other grades and subject areas. Table leaders from each committee were present in the vertical articulation meeting, which allowed them an opportunity to share with the entire group their recollection of the process and discussions that occurred within their committees. Following group discussion, each participant on the vertical articulation panel considered the recommended cut scores and their impact data as well as other potential cut scores and the changes in impact data associated with other potential cut scores. Each member of the vertical articulation committee provided a unique recommendation to keep or change the final cut scores. Prior to rendering judgments, the lead facilitator impressed upon the vertical articulation panel that their holistic, policy-oriented cut score recommendations would supplement, not overwrite, the content-oriented cut recommendations provided by the standard setting panels and would provide the North Carolina State Board of Education with additional information to consider when deciding which cut scores to adopt.

Results

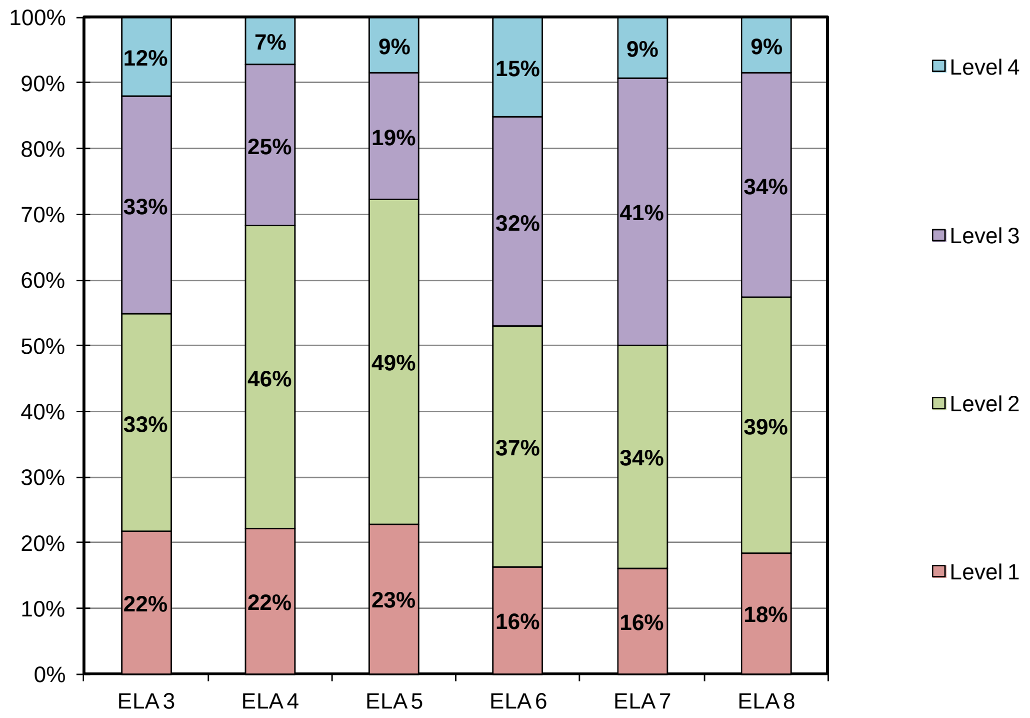
The standard setting panels final recommended cut scores, obtained prior to the vertical articulation session, are presented in Table 6. The reader should note that these cut scores are reported as page numbers within the ordered item book, not raw scores. NCDPI will translate these page cuts into the final reporting scale in a future study. The figures following Table 6 display impact data for the Mathematics, Reading, Science, and End-of-Course, respectively, based upon these cut score recommendations.

Table Pre Vertical Articulation Page uts

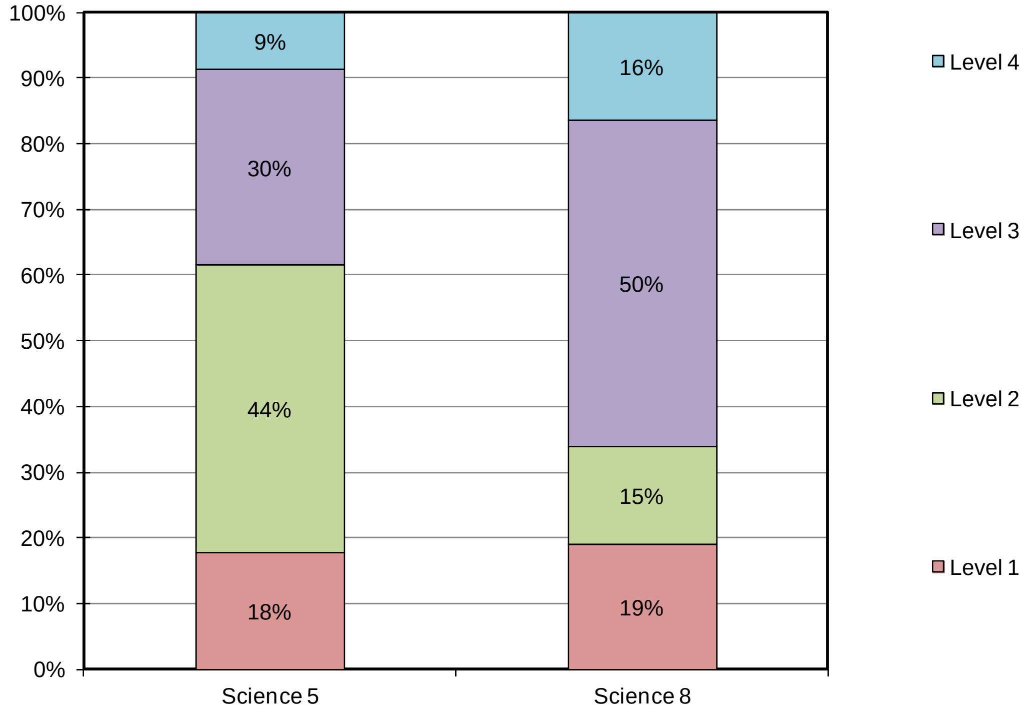
Assessment	Level	Level	Level
Mathematics 3	16	41	69
Mathematics 4	15	34	70
Mathematics 5	9	33	65
Mathematics 6	10	32	67
Mathematics 7	9	28	59
Mathematics 8	10	30	70
Reading 3	26	55	74
Reading 4	25	58	75
Reading 5	23	55	71
Reading 6	15	46	69
Reading 7	15	45	70
Reading 8	16	42	70
Science 5	12	45	69
Science 8	6	20	64
Biology	20	47	68
English II	9	34	79
Math I	9	29	60



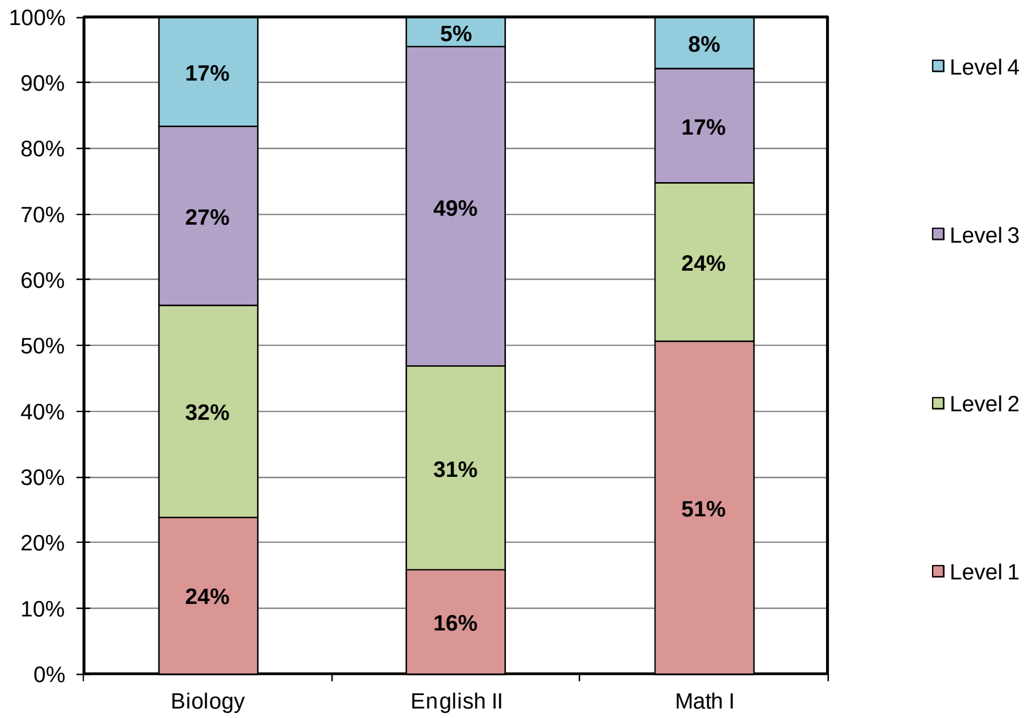
igure Pre Vertical Articulation Impact Data Mathematics



igure Pre Vertical Articulation Impact Data Reading



igure Pre Vertical Articulation Impact Data Science and

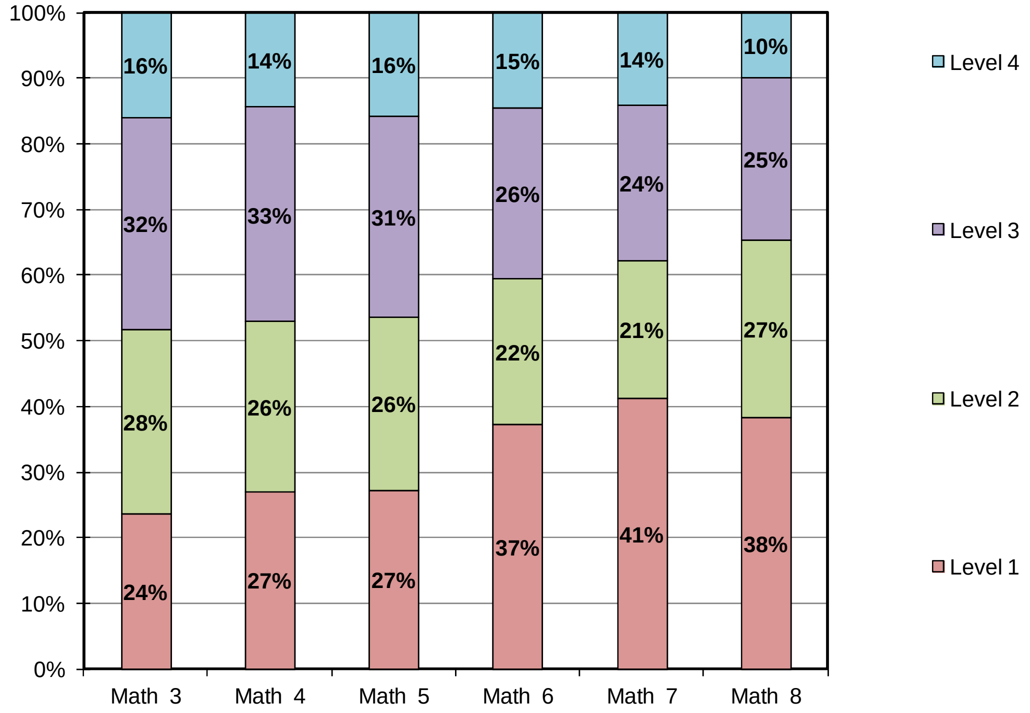


igure Pre Vertical Articulation Impact Data EO

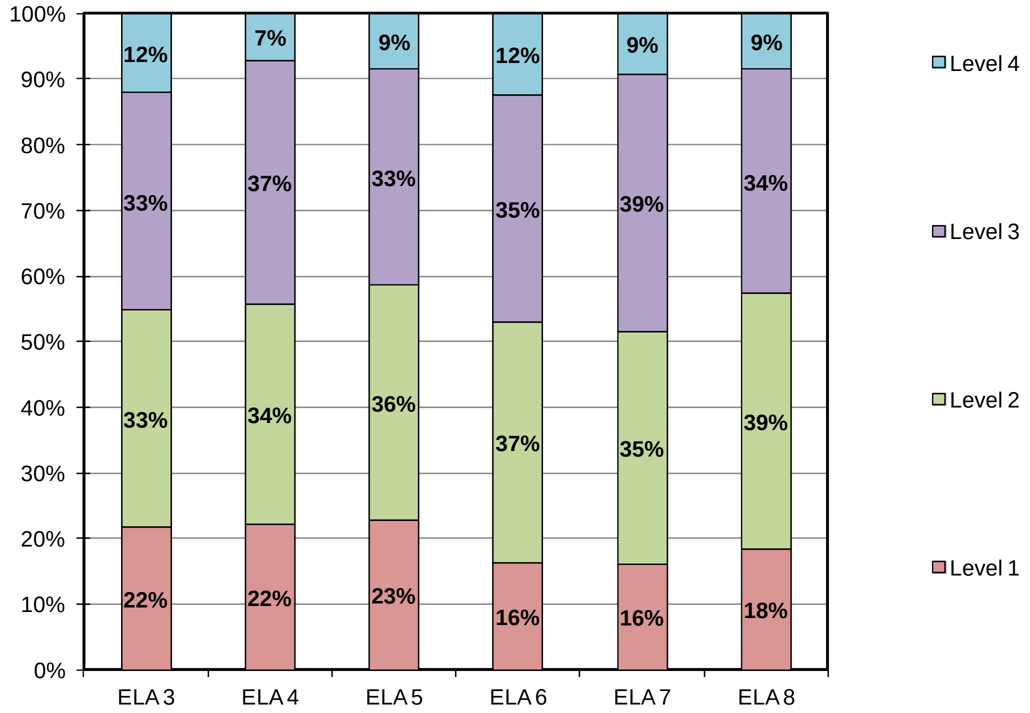
Cut scores obtained following the vertical articulation session are shown in Table 7, and impact data associated with these recommended cut scores are displayed in the subsequent figures.

Table Post Vertical Articulation Page uts

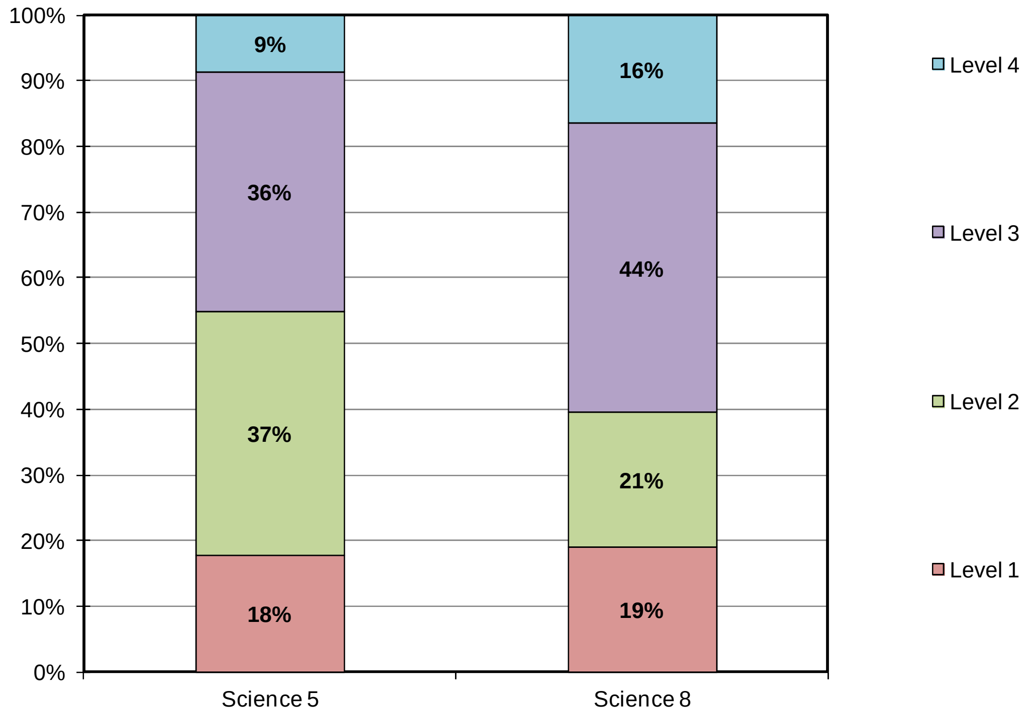
Assessment	Level	Level	Level
Mathematics 3	16	38	73
Mathematics 4	10	34	70
Mathematics 5	7	30	65
Mathematics 6	4	24	67
Mathematics 7	6	28	65
Mathematics 8	5	25	70
Reading 3	26	55	74
Reading 4	25	50	75
Reading 5	23	46	71
Reading 6	15	46	73
Reading 7	15	47	70
Reading 8	16	42	70
Science 5	12	40	69
Science 8	6	25	64
Biology	20	47	71
English II	9	36	79
Math I	2	20	60



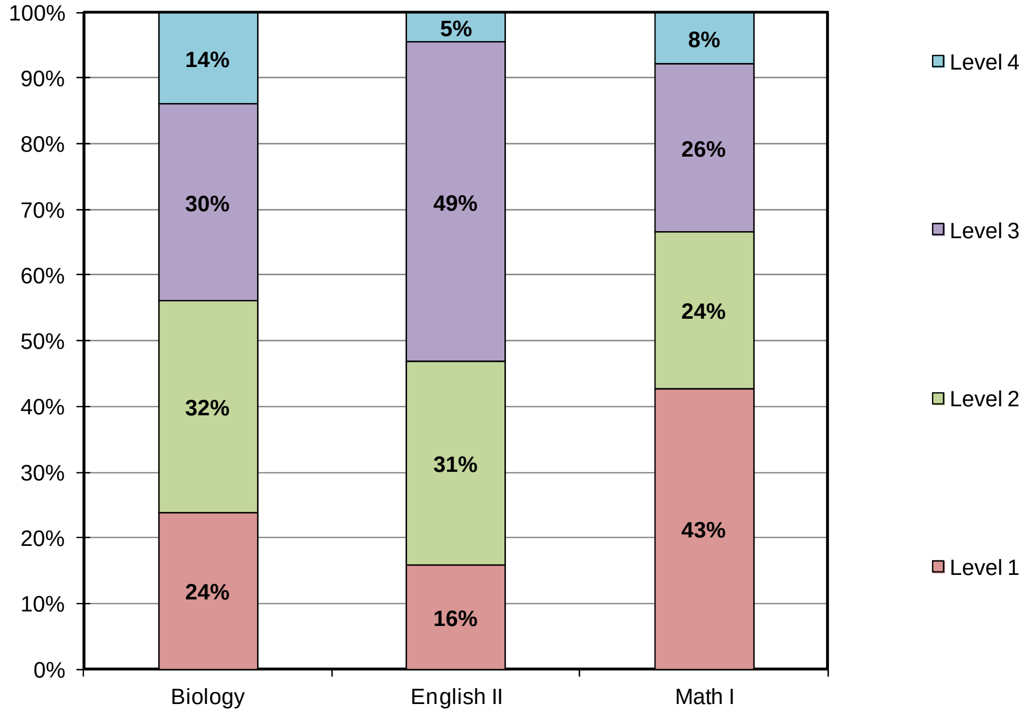
igure Post Vertical Articulation Impact Data Mathematics



igure Post Vertical Articulation Impact Data Reading



igure Post Vertical Articulation Impact Data Science and



igure Post Vertical Articulation Impact Data EO

References

Lewis, D. M., Green, D. R., Mitzel, H.C., Baum, . Patz, R.J. (1998). The Bookmark standard setting procedure: Methodology and recent implementations. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Mitzel, H. C., Lewis, D. M., Patz, R. J., Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

North arolina Testing Program
EO EO Standard Setting

Agenda: Single-Grade Panels

Day ednesday uly

Activity	Time
Table leader training (*Table leaders only*)	8:00 8:45 AM
Large group kick-off meeting	9:00 9:30 AM
Break	9:30 9:45 AM
Committee introductions	9:45 10:00 AM
Achievement level descriptor revision training	10:00 10:15 AM
Achievement level descriptor revisions	10:15 AM 12:15 PM
Lunch	12:15 1:00 PM
Just barely level descriptions	1:00 2:15 PM
Ordered item booklet review	2:15 3:15 PM
Break	3:15 3:30 PM
Standard setting training and practice round	3:30 4:15 PM
Round 1	4:15 5:30 PM

Day Thursday uly

Activity	Time
Round 1 feedback and discussion	8:00 9:15 AM
Round 2	9:15 10:15 AM
Break	10:15 10:30 AM
Write behavioral descriptions	10:30 11:15 AM
Round 2 feedback and discussion	11:15 AM 12:15 PM
Lunch	12:15 1:00 PM
Round 3	1:00 1:30 PM
Break/Collect secure materials	1:30 2:30 PM
Round 3 feedback and discussion	2:30 3:00 PM
Wrap-up and evaluations	3:00 3:15 PM

Day riday uly

Activity	Time
Vertical articulation (*Table leaders only*)	1:00 3:30 PM

North arolina Testing Program
EO EO Standard Setting

Agenda: Three-Grade Panels

Day Monday uly

Activity	Time
Table leader training (*Table leaders only*)	8:00 8:45 AM
Large group kick-off meeting	9:00 9:30 AM
Break	9:30 9:45 AM
Committee introductions	9:45 10:00 AM
Achievement level descriptor revision training	10:00 10:15 AM
Achievement level descriptor revisions LOWER GRADE	10:15 AM 12:15 PM
Lunch	12:15 1:15 PM
Achievement level descriptor revisions MIDDLE GRADE	1:15 3:15 PM
Break	3:15 3:30 PM
Achievement level descriptor revisions UPPER GRADE	3:30 5:30 PM
NCDPI/Pearson debrief meeting	5:45 6:15 PM

Day Tuesday uly

Activity	Time
Just barely level descriptions LOWER GRADE	8:00 9:15 AM
Ordered item booklet review LOWER GRADE	9:15 10:15 AM
Break	10:15 10:30 AM
Standard setting training and practice round	10:30 11:15 AM
Round 1 LOWER GRADE	11:15 AM 12:30 PM
Lunch	12:30 PM 1:30 PM
Write behavioral descriptions LOWER GRADE	1:30 2:15 PM
Round 1 feedback and discussion LOWER GRADE	2:15 3:30 PM
Break	3:30 3:45 PM
Round 2 LOWER GRADE	3:45 4:45 PM
NCDPI/Pearson debrief meeting	5:00 5:30 PM

Day ednesday uly

Activity	Time
Round 2 feedback and discussion LOWER GRADE	8:00 9:00 AM
Round 3 LOWER GRADE	9:00 9:30 AM
Just barely level descriptions MIDDLE GRADE	9:30 10:45 AM
Round 3 feedback and discussion LOWER GRADE	10:45 11:15 AM
Ordered item booklet review MIDDLE GRADE	11:15 AM 12:15 PM
Lunch	12:15 1:00 PM
Round 1 MIDDLE GRADE	1:00 2:15 PM
Write behavioral descriptions MIDDLE GRADE	2:15 3:00 PM
Break	3:00 3:15 PM
Round 1 feedback and discussion MIDDLE GRADE	3:15 4:30 PM
Round 2 MIDDLE GRADE	4:30 5:30 PM
NCDPI/Pearson debrief meeting	5:45 6:15 PM

Appendi B

Day Thursday uly

Activity	Time
Round 2 feedback and discussion MIDDLE GRADE	8:00 9:00 AM
Round 3 MIDDLE GRADE	9:00 9:30 AM
Just barely level descriptions UPPER GRADE	9:30 10:45 AM
Round 3 feedback and discussion MIDDLE GRADE	10:45 11:15 AM
Ordered item booklet review UPPER GRADE	11:15 AM 12:15 PM
Lunch	12:15 1:00 PM
Round 1 UPPER GRADE	1:00 2:15 PM
Write behavioral descriptions UPPER GRADE	2:15 3:00 PM
Break	3:00 3:15 PM
Round 1 feedback and discussion UPPER GRADE	3:15 4:30 PM
Round 2 UPPER GRADE	4:30 5:30 PM
NCDPI/Pearson debrief meeting	5:45 6:15 PM

Day riday uly

Activity	Time
Round 2 feedback and discussion UPPER GRADE	8:00 9:00 AM
Round 3 UPPER GRADE	9:00 9:30 AM
Break/Collect secure materials	9:30 10:30 AM
Round 3 feedback and discussion UPPER GRADE	10:30 11:00 AM
Wrap-up and evaluations	11:00 11:15 AM
Lunch	11:15 AM 1:00 PM
Vertical articulation (*Table leaders only*)	1:00 3:30 PM