The North Carolina Testing Program

Technical Report

2012–2015

Mathematics Assessments

End-of-Grade 3–8 and End-of-Course Math I

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

TD
[Email address]

Prepared by:

Kinge Mbella, Ph.D.

Min Zhu, Ph. D.

Thakur Karkee, Ph. D.

Hope Lung, Section Chief, Test Development

North Carolina Department of Public Instruction

March 2016

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 re-evaluating 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 Common Cores State Standards in Mathematics and Reading/English Language Arts (adopted by the SBE in June 2010) and the NC Essential Standards (adopted by the SBE in February 2010) were implemented, and the 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. Like 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	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 Mathematics EOG and EOC Assessments

This technical manual addresses that the End-of-Grade (EOG) assessments of Math in grades 3 through 8 are aligned to the NC Standard Course of Study (NC*SCS*) that measures NC students' mathematics skills. The standards are assessed again in high school with the Math I End-of-Course (EOC) assessment. End-of-grade and end-of-course assessments are only administered to students in English and Braille. Other native language translation versions are not yet available.

Each operational base form of the EOG Math assessment has between 44 and 50 operational items. In grades 3 and 4, all the items are multiple-choice. The EOG grade 5 assessment has 38 multiple-choice and 6 gridded response items for a total of 44 operational items. For EOG grades 6 through 8, the item breakdown is 41 multiple-choice and 9 gridded response items for a total of 50 operational items. EOC Math I assessments has 49[a] items, of which 39 are multiple-choice and 10 gridded response items. *Table 1.2* shows the complete summary of total operational items by item type and maximum possible observable score. In addition to the total number of operational items each EOG form has 10 field test items embedded within each form. EOC Math I has 11 field test items embedded in each form. These field test items embedded within the operational setting are used to replenish the item bank to build new forms as required.

The EOG assessments were available in Paper format only in 2012–13. Beginning in the 2014–15 school year, the EOG grade 7 was also available as a computer-based, fixed-form administration. EOC Math I assessment was designed as a computer-based fixed form assessment with paper-based fixed forms available for schools and individual students. Each computer fix form mode is the exact duplicate of a corresponding paper form.

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 instructional days of the semester. For students enrolled in yearlong courses, EOG and EOC assessments must be administered within the final ten instructional days of the school year. Students have up to four hours to complete each assessment.

[a] The original test blueprint was designed to have 50 items but during item analysis 1 item did not meet the psychometric criteria and an item was dropped from each form

Table 1.2 Number of Items and Maximum Possible Score by Item Type

Grade	MC Item		GR Item	
	Number of Items	MSP per Item	Number of Items	MSP per Item
Grade 3	44	1		
Grade 4	44	1		
Grade 5	38	1	6	1
Grade 6	41	1	9	1
Grade 7	41	1	9	1
Grade 8	41	1	9	1
Math I	39	1	10	1

Note: MC=Multiple-Choice; GR=Gridded Response; MSP=Maximum Score Possible

1. 3 Report Summary

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

Chapter 2 presents an overview of the validation framework embedded throughout the design and development of the EOG and EOC assessment. 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 document. First section provides a brief introduction of validity and an outline of key validity evidences as documented in this report. The second sections discusses the main proposed uses of scores from EOG and EOC assessments.

Chapter 3 describes the 22-step test development outline adopted by 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 NCDPI to assure the administration 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 disability, English Language Learners are able to take EOG and EOC assessments.

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

Chapter 7 describes the analyses of operational data after the first operational administration of EOG and EOC in 2012–13. The chapter begins with a description of the random spiraling process used to administer three parallel forms across North Carolina. This chapter 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 EOG and EOC. 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 Math 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 2015 administration cycles. 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. Section two presents samples and a summary description of the various standardized reports created by NCDPI and available to LEA to share assessments 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 structure of EOG and EOC assessments. Evidence 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 scores. The final sections of the chapter document validity evidence based on content summarized from the alignment study, evidence based on relation to other variables summarized from the EOG/EOC Quantile® Framework linking study and the last part 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 assessment. 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 proposed uses of scores from EOG and EOC assessments.

2. 1 Summary Validation Framework for Math

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*.

> *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 (p11).*

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, NCDPI will be gathering, evaluating, and documenting relevant evidences validating the proposed uses of test scores. From the test developer 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 developers of EOG and EOC, NCDPI has adopted a validation framework consistent with that prescribed in the *Standards*. Under this framework, 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 becomes 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 establish the technical quality of EOG and EOC Math 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 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. *Table 2.1* presents an outline of the validation framework with relevant components as documented in this report.

Table 2.1 NCDPI Validation Framework for Math EOG and EOC Assessments

Sources of Validity Evidence	References	Data
Evidence based on intended uses	Chapters 2 and 9	Score Reports
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	
Evidence based on internal structure and reliability	Chapter 10	Cronbach's 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 5 and 10	Test accommodation
Evidence based on appropriate reporting	Chapter 9	
Evidence based on relations to other variables	Chapter 10	Quantile Framework Linking study

2. 2 Uses of NC Math EOG/EOC Assessments

The North Carolina State Test Program (NCSTP) designs, develops, and administers customized high quality assessments in grades 3–8 and high school, which are aligned to College- and Career-Readiness standards for Mathematics adopted by the North Carolina State Board of Education in June 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 NCSBE to classify students based on their progress and readiness as defined by NC*SCS* College- and Career-Readiness standards.

- Assessment results are also 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. The educator effective model currently used in NC expects teachers (standard 6) and school executives (standard 8) will contribute to the academic success of students. Test scores from EOG and EOC assessments, Career and Technical Education Post-Assessments, and the Measures of Student Learning are used in a statewide value added growth model to provide ratings for these respective standards measuring the relative contribution of teachers and educators. In the second component, school performance grades—scores from EOG and EOC assessments—are used as indicators in the school report card in 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% of students' final grade in Math I has to come from their EOC assessment score. It is worth mentioning that the EOG in grades 3–8 is 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 levels. The North Carolina *Testing Code of Ethics* (see Appendix 2-A Testing Code of Ethics) dictates that educators use test scores and reports appropriately. This means that educators recognize that a test score is only one piece of information and must be interpreted as intended. This is at the core of validity and is reiterated throughout the *Standards* that it is the intended interpretation[s] of test scores which are valid, not the test itself.

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 available to test coordinators at the district level is used to generate a variety of score reports to assist with score interpretations: class roster reports, score

frequency reports, achievement level frequency reports, and goal summary reports. To help with interpretations of these various reports, the NCDPI also publishes on its website an interpretive guide for the various score reports intended to help educators and decision makers at the classroom, school, and district levels understand the content and uses of these reports. These guides are also 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 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 with the *Standards*, this chapter documents steps implemented by 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. Figure 3.1 shows the sequence of events 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 North Carolina State Assessment development process including test specifications, items and form formats, alignment studies, test administrations for alternate assessments, and students with disabilities and English Language Learners (ELL), standard setting, reporting, and uses of data for measuring growth can also be found in the technical brief (NCDPI, 2014) on the NCDPI web page.

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*SCS*. To

accommodate the shortened timeline, NCDPI made three modifications to the NCSBE assessment development flow chart *Table 3.1*:

I. The NCDPI waived the full-scale "item tryout" component (Steps 3–8) and implemented a smaller scale usability study for the newly developed innovative gridded response 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[b] Develop Test Specifications (Blueprint)	Step 9[c] 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	

[b]Activities done only at implementation of new curriculum

[c] Activities involving NC teachers

3.1 Content Standards and Curriculum Connectors

As stated in Chapter 1 (see *Table 1.1*), the NCSBE adopted the revised NC*SCS* in June 2010. The revised NC*SCS* are aligned to the Common Core state standards (CCSS). Operational test forms aligned to the NC*SCS* for ELA and Mathematics were administered in 2012–13 testing administration (READY initiative). Testing of North Carolina students' skills relative to the standards and objectives in the NC*SCS* 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), depth of knowledge and overall alignment of assessments to content standards.

3.1.1 Revised Bloom Taxonomy (RBT) and Depth of Knowledge (DOK)

As part of pre-item development training for the new EOG and EOC assessments, NCSTP with collaboration from NCDPI 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 NCSTP test measurement specialist (TMS), 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. Karin Hess (researcher at Center for Assessment) developed a 4-by-6 table containing Webb's DOK levels across the top and RBT process dimension across the side see *Table 3.2*. During the workshop, participants received training and started to classify NC*SCS* using Hess's matrix.

Table 3.2 Hess' Cognitive Rigor Matrix with Curricular Examples

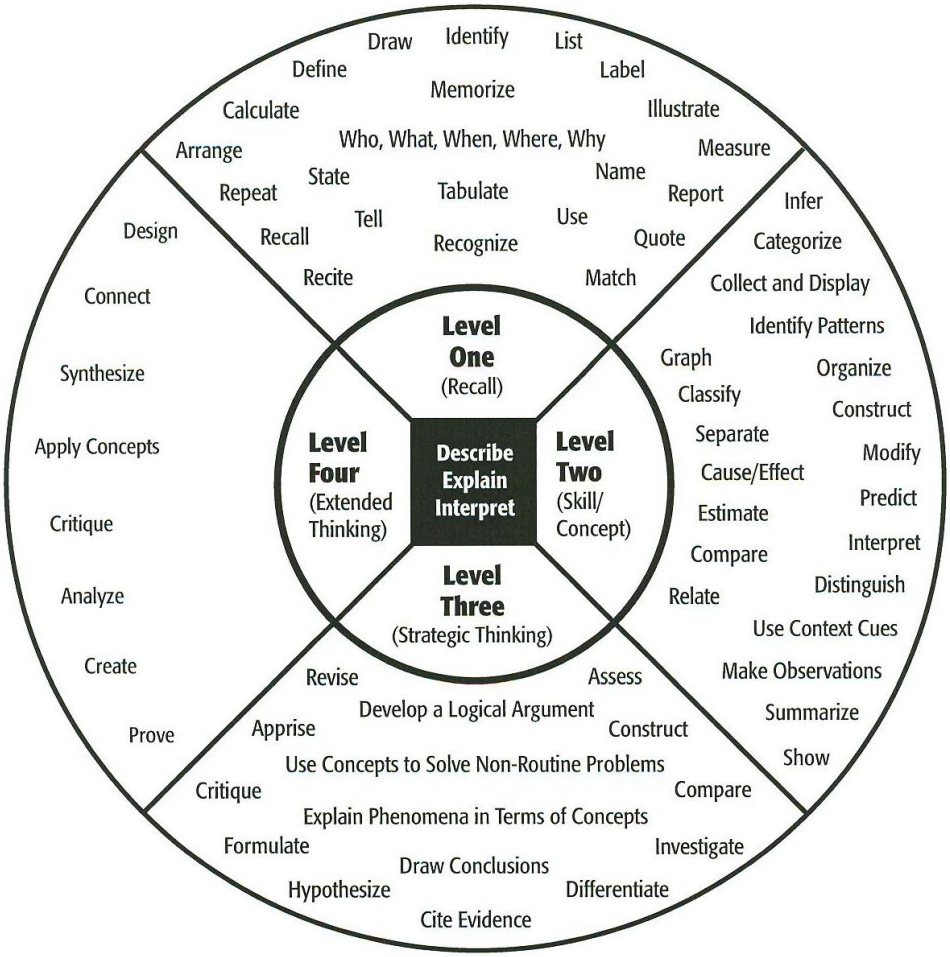
Bloom's Revised Taxonomy of Cognitive Process Dimensions	**Webb's Depth-of-Knowledge (DOK) Levels**			
	Level 1 **Recall & Reproduction**	**Level 2** **Skills & Concepts**	**Level 3** **Strategic Thinking/Reasoning**	**Level 4** **Extended Thinking**
Remember Retrieve knowledge from long-term memory, 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	○ 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

On July 26, 2010 NCDPI organized a one day face-to-face training session on Webb's Alignment. Norm Webb was invited to facilitate the training on alignment and DOK. During the first 4 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 was built on the July 8 workshop in which participants were able to classify standards using the Hess matrix. During the July 26 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.

Figure 3.1 Webb Alignment Tool

Depth of Knowledge (DOK) Levels



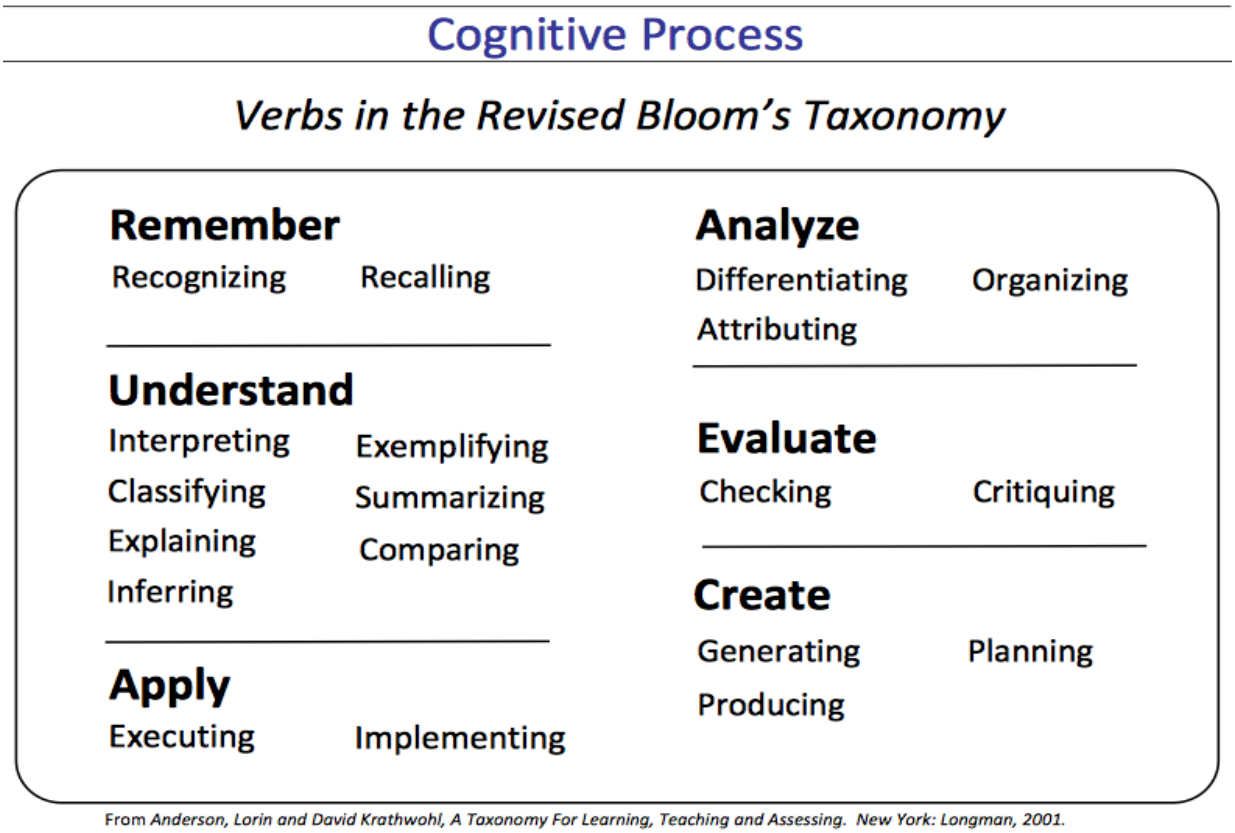
Level One Activities	Level Two Activities	Level Three Activities	Level Four Activities
Recall elements and details of story structure, such as sequence of events, character, plot and setting.	Identify and summarize the major events in a narrative.	Support ideas with details and examples.	Conduct a project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

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 on 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



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[d] for the NCSTP were developed in accordance with the standards and objectives specified in the NCSCS. 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 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 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* and *Table 3.4* show the adopted content domain specification for Math EOG grades 3–8 and EOC Math I assessments.

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

Table 3.3: Content Standards and Weight Distributions EOG Math Grades 3–5

Domain/Standards	Grade 3	Grade 4	Grade 5
Operations and Algebraic Thinking	30–35%	12–17%	5–10%
Number and Operations in Base Ten	5–10%	22–27%	22–27%
Number and Operations–Fractions	20–25%	27–32%	47–52%
Measurement and Data	22–27%	12–17%	10–15%
Geometry	10-15%	12-17%	2-7%
Total	**100%**	**100%**	**100%**

Table 3.4: Content Standards and Weight Distributions EOG Math Grades 6–8 and EOC Math I

Domain/Standards	Grade 6	Grade7	Grade 8	Math I
Ratios and Proportional Relationships	12– 17%	22– 27%	NA	NA
The Number System	27– 32%	7– 12%	2– 7%	NA
Expressions and Equations	27– 32%	22 – 27%	27– 32%	NA
Functions	NA	NA	22–27%	35– 40%
Geometry	12– 17%	22– 27%	20– 25%	10– 15%
Statistics and Probability	7– 12%	12– 17%	15– 20%	15– 20%
Number and Quantity	NA	NA	NA	5– 10%
Algebra	NA	NA	NA	25– 31%
Total	**100%**	**100%**	**100%**	**100%**

The Math NC*SCS* consists of a set of content domains/standards for each grade. The sampling of standards and corresponding weights across grades are shown in *Table 3.3* and *Table 3.4*. The NC*SCS* for Grades 3 through 5 were written to include content from Operations and Algebraic Thinking, Number and Operations in Base Ten, Number and Operations–Fractions, Measurement and Data, and Geometry. Based on the content specification, about 64% of grade 6 math content is on the Number System and Expression and Equations. For grade 7, the core content areas are Ratios and Proportional Relationships, Expressions and Equations, and Geometry. In grade 8, core content areas include Expression and Equations, Functions, and Geometry. In Math I, the focus is on Functions and Algebra. 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. However, at the domain level and in terms of the relative emphasis of the standards coverage, all test blueprints conform to the content domain specification *see* Appendix 3-B Math Test Specifications & Blueprints.

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

Prior to 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 participant to this work included personnel from NCDPI Accountability division (that also included test development section), 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 used for item writers and reviewers. Lessons learned from this training were used to re-evaluate how items for the new assessments were developed following the plain English framework which emphasize clarity without altering the construct being assessed. In general, the goal was to develop items that assess the construct without adding in 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 student's 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 student's true ability level. 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 which these factors may exist while still maintaining appropriate measurement of the construct for the entirety of the student population. The concept of Universal Design originated in architecture with the goal to provide 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 students with disabilities, and students with limited English proficiency. These core principles are emphasized in the item writer training courses designed by NCDPI and required to be taken by all potential item writers/reviewers. The complete workshop materials including the workshop agenda is 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 NC*SCS* was addressed during recruitment. The use of North Carolina educators to develop items strengthened the instructional and face validity of the items. Teachers and educators were recruited as needed. To be included in the item writer or reviewer pool, potential teachers and educators from North Carolina were asked to 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.

The "A" level subject course covers two main topics. The first section presents an overview tutorial unpacking the NC*SCS* standards for the specific content area. This is intended to broaden understanding of the content standards and the areas of interest. The second section of the tutorial provides trainees with an overview of Webb's Depth of Knowledge (DOK) and Webb's alignment model adopted by the NCDPI as a tool to help them develop test questions that closely agree with the NCSCS standards.

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
6. Next Steps

Once the online training courses are completed, the teacher is 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 items. Teachers who submit interest forms will be contacted when item writing or reviewing is needed in their subject area. For complete description of 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 Gridded-Response Items

As part of the Accountability and Curriculum Reform Effort (ACRE) initiative and the redesign of the end-of-grade and end-of-course assessments in 2011, the NCDPI conducted a usability study on new item types with the goal to make assessments more authentic and engaging to students. The usability study for math was on Gridded Response (GR) items. The evaluation criteria centered on aspects of accessibility, user-friendliness, and authenticity of construct measured. While the new item type hold promise to improve student engagement, appeal of the assessment and to minimize possibility of guessing, it does 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 less computer skills are not disadvantaged. A usability study allowed test developers to observe students interacting with the new items and provided valuable feedback on the improvement, design and selection of GR items.

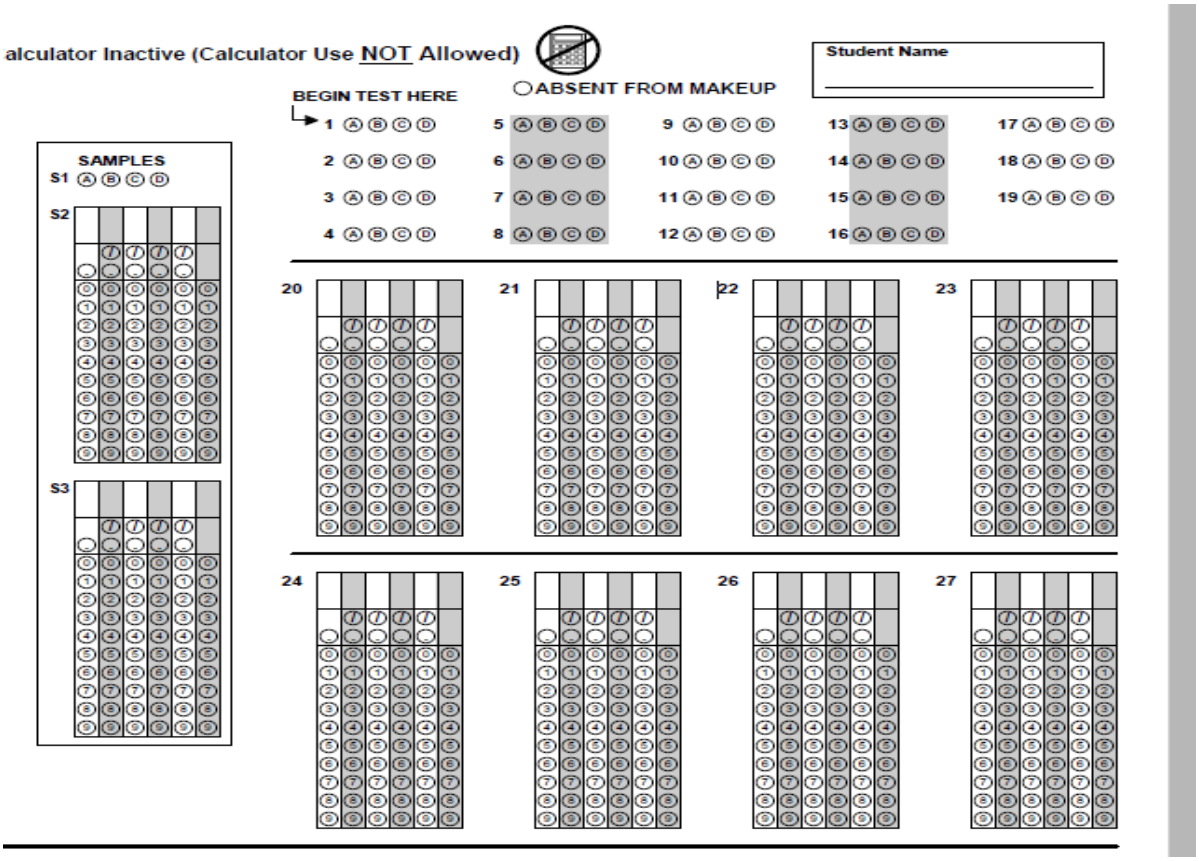
Figure 3.3 shows snapshot of the GR item and sample response sheet that is used for Math Grade 5, 6, 7, 8, and Math I. Students are instructed to read the stem then enter their answers into the text box provided for computer forms or using the grid shown in Figure 3.4. Only numbers 0 to 9 and symbols ., - or / are allowed in the answer.

Figure 3.3 Gridded Response Item Example

Questions 17 through 22 require you to write your answers in the boxes provided on your answer sheet. Write only one number or symbol in each box and fill in the circle in each column that matches what you have printed. Fill in only one circle in each column.

17 The fifth grade has 152 students. Each student has 18 pencils. How many pencils do the students have altogether?

Figure 3.4 Sample Gridded Response Answer Sheet.



The usability study for the computer based GR items in Math was conducted with 6th grade math students. The goal was to design GR items with an intuitive and easy-to-use interface. With this goal in mind, the NCDPI purposefully selected volunteer schools that had a low computer-student ratio for the study, since such schools were more likely to have students with relatively less exposure to computers. For Grade 6 Math, a total of 4 students from Fuquay-

Varina Middle School in Wake County took part in the usability study. During the two day window, evaluators from the NCDPI met with selected students at their schools with laptops pre-loaded with assessment software.

Each student worked on six GR items with one evaluator for up to one hour in a meeting room in which the evaluator recorded the session and interacted with the student using a defined protocol. During the session, the evaluator explained the purpose of the study, set a relaxed tone, and encouraged the student to talk openly about each item that was presented to him/her on the computer. Since the purpose of the usability study was to evaluate the user-friendliness of the item interface, the content of the questions was not challenging for the student, and no scores were reported. *Table 3.5* shows the usability study process in detail.

Table 3.5 Usability Study 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 item 2– *4*	• *Repeat protocol with question 2–4*	7–10
6. Conclusion	• Present survey questions. • Replay recording of interaction and ask the student what he/she was thinking during certain parts of the interaction. • Thank the student for their feedback and participation.	5–15
	TOTAL	35– 60

At the end of each session, evaluators went over a set of survey questions with each student. Evaluators also completed a second evaluator survey at the end of the study. The complete survey instrument is presented in Appendix 3-E TEUS Survey Questions_2011**.**

Four students completed the Math Grade 6 usability study. It took an average of about eight minutes for students to complete the six GR items. Overall results were positive, and students in general reacted fine to the GR items on the computer. From the perspective of the students, below are summaries from the interviews.

- After reading the directions, did students know how to show their answers?

  The survey results showed the item type was challenging for Grade 6 students. All four students who participated spent fewer than two minutes on the directions before they started working on the items. One student did not understand what the directions (i.e. "only 0 to 9, etc., allowed") meant. After reading the directions, two students (50%) still could not figure out how to enter a mixed number answer. Two students didn't realize there was a calculator/highlighter on the screen that they could use. However, the scroll bar did not create much of a problem for the students to answer the questions.

- Was anything confusing or unclear about these questions?

  On the Math 6 test, various technical issues were reported when students answered the questions. Three out of four students reported that items could not be recorded correctly; one student reported the item disappeared after highlighting; and two students were not sure if scoring was done correctly. When students were confused with these issues, facilitators intervened and provided help.

  Despite the technical problems, students in general reacted to the items positively. One student said the GR items were not very different from what she was used to, and one indicated that he liked GR items.

3.3.4 Item Difficulty

For the purposes of guiding item writers to provide a variety of items, item writers 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 the item pool to ensure 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.

3.3.5 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 with high construct irrelevance variance and will underestimate or overestimate students' achievement. 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/Curriculum and Instruction section[e]. These groups independently reviewed proposed items through NC's online item writing system, the Test Development System (TDS), and classified them by the NC*SCS* and Depth of Knowledge (DOK) levels. Any items with discrepant classifications were prevented from continuing through item development until the discrepancy was resolved.

[e]The NCDPI/test development created an alignment plan in 2010 prior to 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.

3.3.6 Item Format

The Math grades 3–4 assessments consist of four-foil (distractor) multiple-choice items. EOG grade 5 assessment has 38 multiple-choice and 6 gridded response items, for a total of 44 operational items. For Math grades 6 through 8, the item breakdown is 41 multiple-choice and 9 gridded response items, for a total of 50 operational items. EOC Math I assessments has 49[f] items, of which 39 are traditional four-foil multiple-choice and 10 gridded response items. Each form is separated into a calculator inactive and a calculator active section. For examples and description of gridded response items see *Figure 3.3*

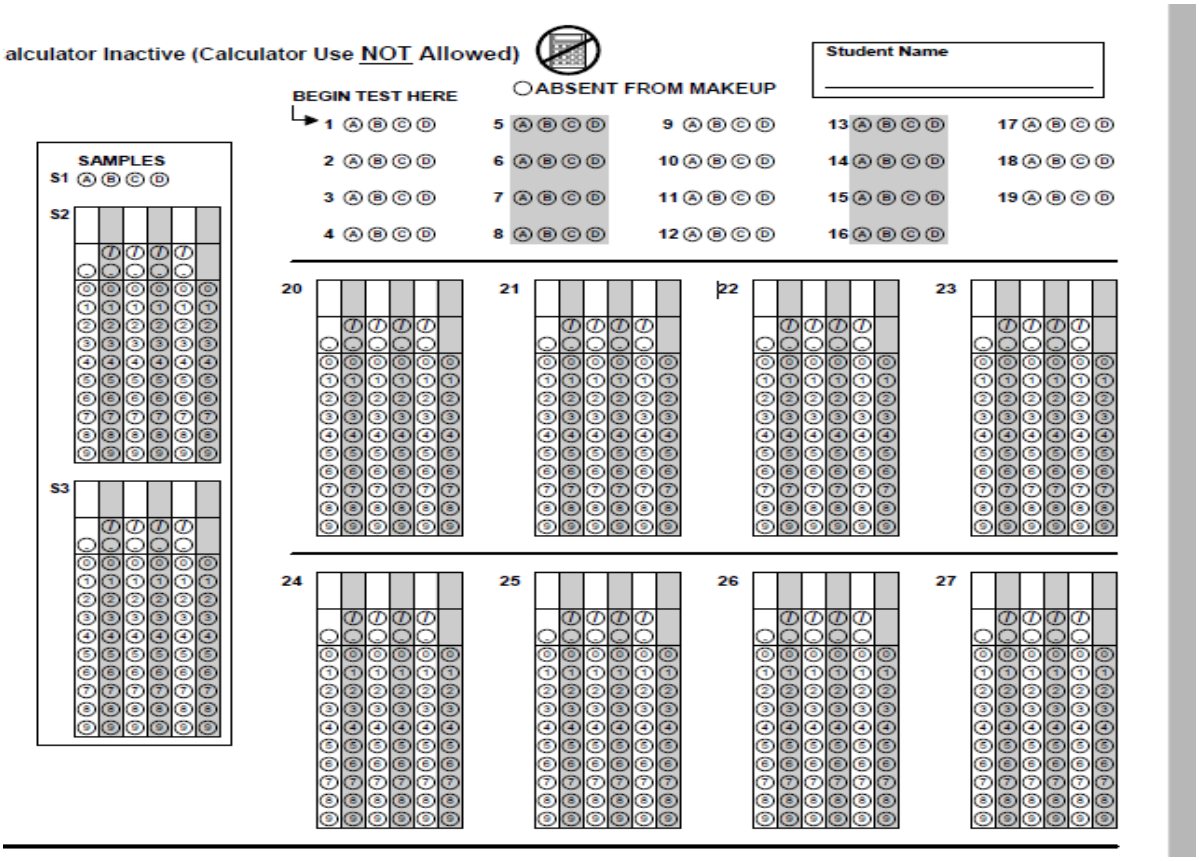
[f] The original test blueprint was designed to have 50 items but during item analysis 1 item did not meet the psychometric criteria and the item was dropped from each form.

Figure 3.3 Gridded Response Item Example

Questions 17 through 22 require you to write your answers in the boxes provided on your answer sheet. Write only one number or symbol in each box and fill in the circle in each column that matches what you have printed. Fill in only one circle in each column.

17 The fifth grade has 152 students. Each student has 18 pencils. How many pencils do the students have altogether?

Figure 3.4 Sample Gridded Response Answer Sheet.



3.4 Step 9. Item Review for Field Testing

To ensure that items developed were aligned to the NC*SCS* standards, each item went through a detailed review process prior to being placed on a field test. AERA/APA/NCME 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 was recruited to review all items. Once items had gone through educator review, test development staff members with input from curriculum specialists also reviewed every item. Items were further reviewed by educators and/or staff familiar with the needs of students with disabilities and English Language Learners (ELL).

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)

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

3.5 Steps 10–11: Assembling and Reviewing Field Test Forms

Items for each grade level were assembled into field test forms[g] 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 **3.6*** 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.

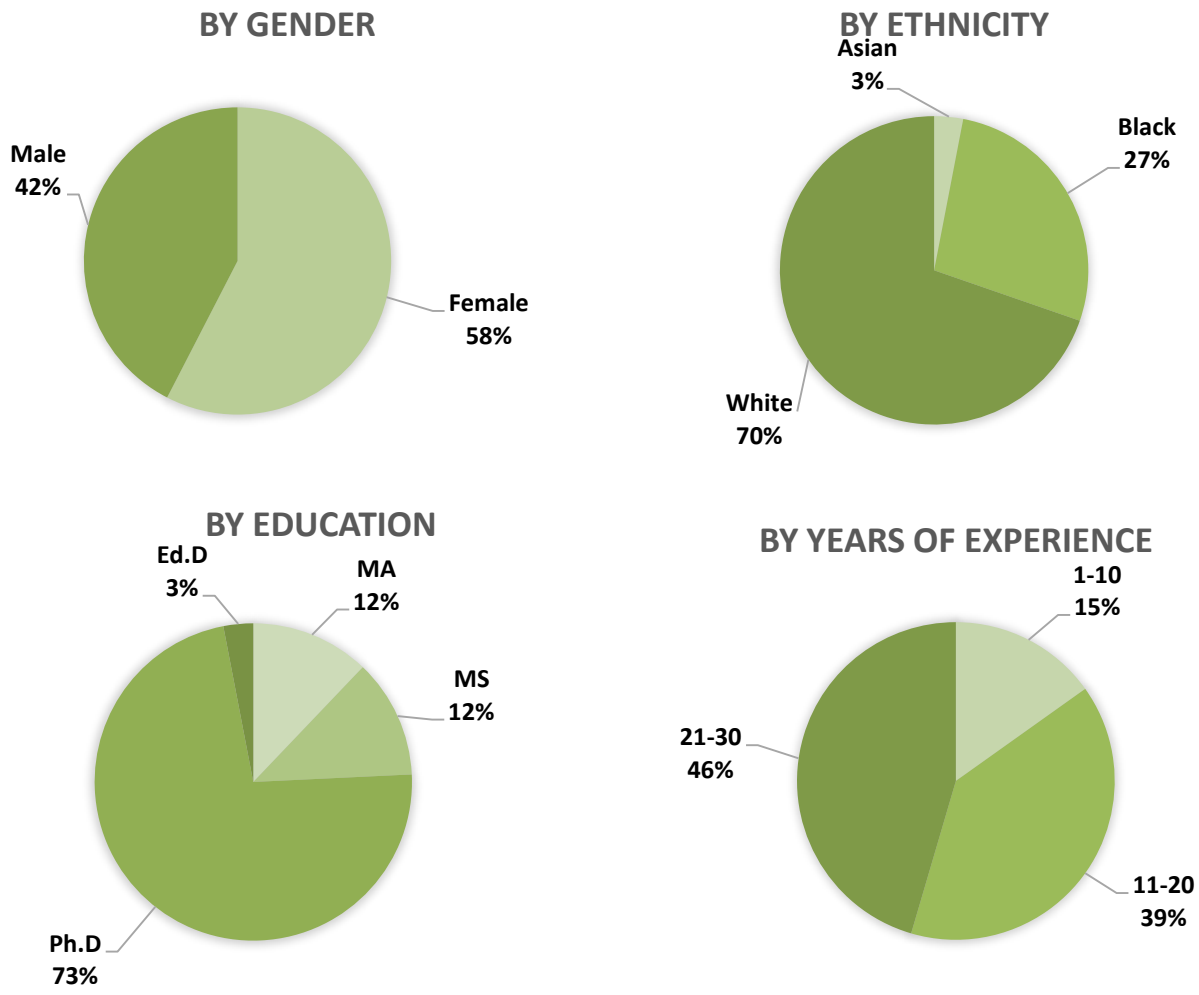
Table 3.6 Number of Items Field Tested for EOG Math and EOC Math I

[g] See complete form assembly process described in chapter 5

Grade Level	Number of Forms	Items Per Form	Total Items Field Tested
Grade 3	10	50	500
Grade 4	10	50	500
Grade 5	10	50	500
Grade 6	10	50	500
Grade 7	10	50	500
Grade 8	10	50	500
Math I	10	50	500

Prior to the field test administration, following steps similar to operational form review, outside content reviewers reviewed the assembled field test forms for clarity, correctness, potential bias or sensitivity and 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. Reading, Math, and Science) have served as external form reviewers during this cycle of EOG and EOC assessments. Descriptive summaries of their demographic and educational backgrounds are shown in the pie charts inFigure 3.5. These experts provided an independent outside evaluation of the forms. All the form reviews were done using the NCSU-TOPS online test development system (TDS). All comments were recorded and reviewed, and any issues were addressed before the forms were administered.

Figure 3.5 Demographic Information for Outside Form Reviewers.



Chapter 4 Field-Test Administration and Operational Form Construction

The NC Math stand-alone field test was administered in Spring 2012. 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, this 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[h]

Sampling for 2011–12 stand-alone field testing of the North Carolina Math assessment was accomplished using stratified random sampling at school level, with the goal being to select a representative sample made up of about 20% of students at every grade from 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
- English Language Learners
- Previous year's test scores

[h] 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.

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).

Table 4.1 shows a comparison of the proportions of students selected for the stand-alone field test compared to the population. The desired sampling rate was set at 15% from each grade level. After attrition, the effective sample across the grade levels ranged from 19,400 for Grade 5 to 22,798 for Math I. 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 Math I where proportion of white students in the sample was about 5% more than in the population, and black students was about 4% less in the sample. In terms of special population categories, the field test samples are representative of the population distribution for ELL and EDS students. The proportion of SWD between the sample and population at the respective grade levels is not as similar as the other variable, with an average of 3.8% difference in proportions. But 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.

Table 4.1 Demographic Summary for Math Field Test 2012 Sample Participants

Math		N	Gender		Ethnicity							Special Subgroup		
			% Female	% Male	% Asian	% Black	% Hispanic	% American Indian	% Multiracial	% Native Hawaiian/Pacific Islander	% White	% ELL	% SWD	% EDS
Grade 3	Population	126,302	48.74	51.26	2.65	25.81	15.27	1.43	4.06	0.08	50.70	10.90	13.02	58.36
	Sample	21,516	49.42	50.58	2.34	25.26	15.89	1.30	4.38	0.10	50.74	11.21	9.78	57.65
Grade 4	Population	125,079	48.73	51.27	2.58	26.36	14.90	1.38	3.86	0.09	50.84	8.57	13.85	58.27
	Sample	19,570	49.41	50.59	3.07	25.57	15.83	1.27	3.88	0.09	50.29	9.23	10.36	57.83
Grade 5	Population	126,871	48.70	51.30	2.50	26.83	13.99	1.43	3.74	0.09	51.42	6.31	13.81	57.44
	Sample	19,428	50.20	49.80	2.20	27.04	13.43	1.45	3.67	0.05	52.15	5.85	9.53	56.94
Grade 6	Population	125,167	48.56	51.44	2.46	27.32	13.13	1.57	3.63	0.09	51.79	5.25	13.26	56.52
	Sample	20,469	49.96	50.04	2.56	25.74	14.01	2.49	3.57	0.07	51.57	5.41	8.74	55.79
Grade 7	Population	123,120	48.74	51.26	2.39	27.75	12.44	1.50	3.56	0.10	52.26	5.35	13.11	55.48
	Sample	20,091	49.31	50.69	2.48	26.90	12.48	1.43	3.50	0.08	53.13	5.16	9.10	56.01
Grade 8	Population	121,569	48.47	51.53	2.37	27.50	11.80	1.61	3.59	0.10	53.03	4.95	12.65	53.92
	Sample	20,334	48.89	51.11	2.67	26.85	12.28	1.69	3.32	0.11	53.09	4.98	8.97	51.75
Math I	Population	134,368	47.12	52.88	2.37	29.74	11.85	1.57	3.46	0.10	50.91	5.96	13.29	53.09
	Sample	22,798	49.46	50.54	2.31	25.78	11.33	0.86	3.59	0.10	56.04	4.77	9.61	48.84

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 Item Analysis Summary From Field Test

Classical item analyses of the field test items were conducted in SAS and included evaluation of item p-value and biserial correlation statistics to determine if items met NCDPI item quality criteria. Item p-value summarizes the proportion of examinees answering each item correctly and is 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 answered correctly. The general NCDPI rule is to keep items with a p-value range of 0.15 to 0.85.

The biserial and point-biserial correlation coefficients are special cases of Pearson correlation coefficient and describes the relationship between a dichotomous variable and a continuous or multi-step variable. Biserial coefficients provides evidence of how well each item on a test form correlates with the total test score. It can also be used as an estimate of item discrimination, or in other words, a measure of how well an item differentiates between 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 done only under exceptional cases and with thorough vetting from the content experts and psychometricians. Items with negative biserial correlations are not retained for use on the operational assessment. *Table 4.2* shows summary-descriptive classical statistics from a field test item pool.

Table 4.2 CTT Field Test 2012 Item Pool Descriptive Statistics for EOG Math 3–8 and EOC Math I

EOG and EOC Math	Number of Items		P-Value				Biserial Correlation			
	Multiple Choice	Gridded Response	Mean	SD	Min	Max	Mean	SD	Min	Max
Grade 3	500		0.54	0.21	0.06	0.96	0.45	0.15	-0.03	0.78
Grade 4	500		0.53	0.19	0.11	0.97	0.48	0.15	-0.08	0.80
Grade 5	420	80	0.47	0.20	0.00	0.91	0.49	0.15	-0.03	0.85
Grade 6	420	80	0.45	0.19	0.03	0.90	0.46	0.16	-0.01	0.83
Grade 7	420	80	0.43	0.20	0.02	0.95	0.48	0.17	0.04	0.91
Grade 8	420	80	0.40	0.20	0.00	0.92	0.43	0.18	-0.12	0.95
Math I	420	80	0.31	0.15	0.00	0.80	0.29	0.20	-0.24	0.85

4.2.2 Item Response Theory (IRT) Summary from Field Test

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 and the two-parameter logistic (2PL) model for the gridded response items. Equation (4-1) presents the mathematical representation for the 3PL, where:

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

where $P_i(\theta)$ is the probability that a randomly chosen examinee 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 major difference between a 3PL model and a 2PL model is that the 2PL model does not directly account for a chance-score parameter. The 2PL model can be expressed as a special

case of the 3PL model with $c_i = 0$ (see Equation (4-2)). For gridded response items, students are required to provide their answers by entering numbers from 0 to 9 and/or symbols ., - or / rather than to select an answer from several choices. The chance to get an item right by guessing would be almost zero.

$$P_i(\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]} \quad \textbf{(4-2)}$$

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)]. 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 parameters statistics from field test item pool.

Table 4.3 IRT Field Test 2012 Item Pool Descriptive Statistics for EOG Math 3–8 and EOC Math I

EOG and EOC Math	Number of Items		Slope(a)				Threshold(b)				Asymptote(g)			
	MC	GR	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Grade 3	500		1.66	0.67	0.09	5.56	0.21	1.22	-5.50	3.22	0.20	0.08	0.04	0.52
Grade 4	500		1.80	0.79	0.15	7.09	0.36	1.12	-3.09	7.72	0.20	0.08	0.04	0.63
Grade 5	420	80	1.87	0.78	0.17	6.16	0.55	1.02	-2.94	4.87	0.18	0.11	0.00	0.63
Grade 6	420	80	1.86	0.80	0.30	7.26	0.73	1.04	-2.25	5.83	0.18	0.11	0.00	0.55
Grade 7	420	80	1.96	0.82	0.14	8.89	0.74	1.04	-5.77	4.23	0.18	0.10	0.00	0.49
Grade 8	420	80	1.79	0.85	-2.29	8.64	0.99	1.06	-3.12	4.34	0.18	0.10	0.00	0.46
Math I	420	80	1.65	1.07	-1.69	7.82	1.70	1.54	-4.35	8.91	0.17	0.10	0.00	0.49

4.2.3 Differential Item Functioning

As the developers 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 AERA/APA/NCME *Standards* (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.

$$\boldsymbol{\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}} \qquad \textbf{(4-3)}$$

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:

$$\beta_{MH} = \log_e(\alpha_{MH})$$

(4-4)

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

$$|D| = -2.35\beta_{MH}$$

(4-5)

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). An 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, rurals, EDS)

Table 4.4 shows field test pool 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 Math EOG grade 5, females performed somewhat better on 258 items compared to males of similar ability, and males performed somewhat better on 221 items compared to females of similar ability. Seven items showed marginal DIF in favor of females, and nine showed marginal DIF in favor of males. A total of five items showed significant DIF, two in favor of females and three in favor of males. The rest of the table is interpreted in a similar fashion. NCDPI rule is to remove all items with DIF flag of "C" from the item bank, and "B" items are sent for further review and only placed on operational form upon a positive review from the bias panel or if a replacement item is not readily available for that content domain. Across all grades, the most "C" DIF items were flagged for "White" versus "Black" category.

Based on recommendations from NCDPI National Technical Advisory Committee (NCTA) the NCDPI has now included two new DIF categories in its DIF evaluation. The first is a school base Urban-versus-Rural category, with urban set as reference groups. Schools in the state are classified as "City," "Suburban," "Town," "Urban," or "Rural" based on assignment criteria defined by the federal department of education. The second DIF category added is an Economically Disadvantaged Students category (EdS). EdS classification is based on whether the student is eligible for school meals as defined by the national nutrition program. Students who are eligible for meal programs make up the focal group, and non-eligible students serve as the reference group.

Table 4.4 Mantel-Haenszel Delta DIF Summary for Math 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-
Grade 3	249	240	6	5				239	199	22	32	4	4		225	226	20	24	2	3
Grade 4	239	245	7	9				243	231	12	13	1			224	235	21	18		2
Grade 5	258	221	7	9	2	3	1	227	206	18	31	6	11	1	241	213	15	18	3	9
Grade 6	238	225	13	19	1	4		240	215	15	21	2	7		241	205	17	25	2	10
Grade 7	238	217	16	14	4	11		239	216	9	25	2	9		231	222	19	21	2	5
Grade 8	245	214	15	18		8	4	237	205	13	27	4	10	6	229	225	15	16	4	5
Math I	217	265	8	5	3	1		236	218	13	21	6	5		250	221	9	15	2	2

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:

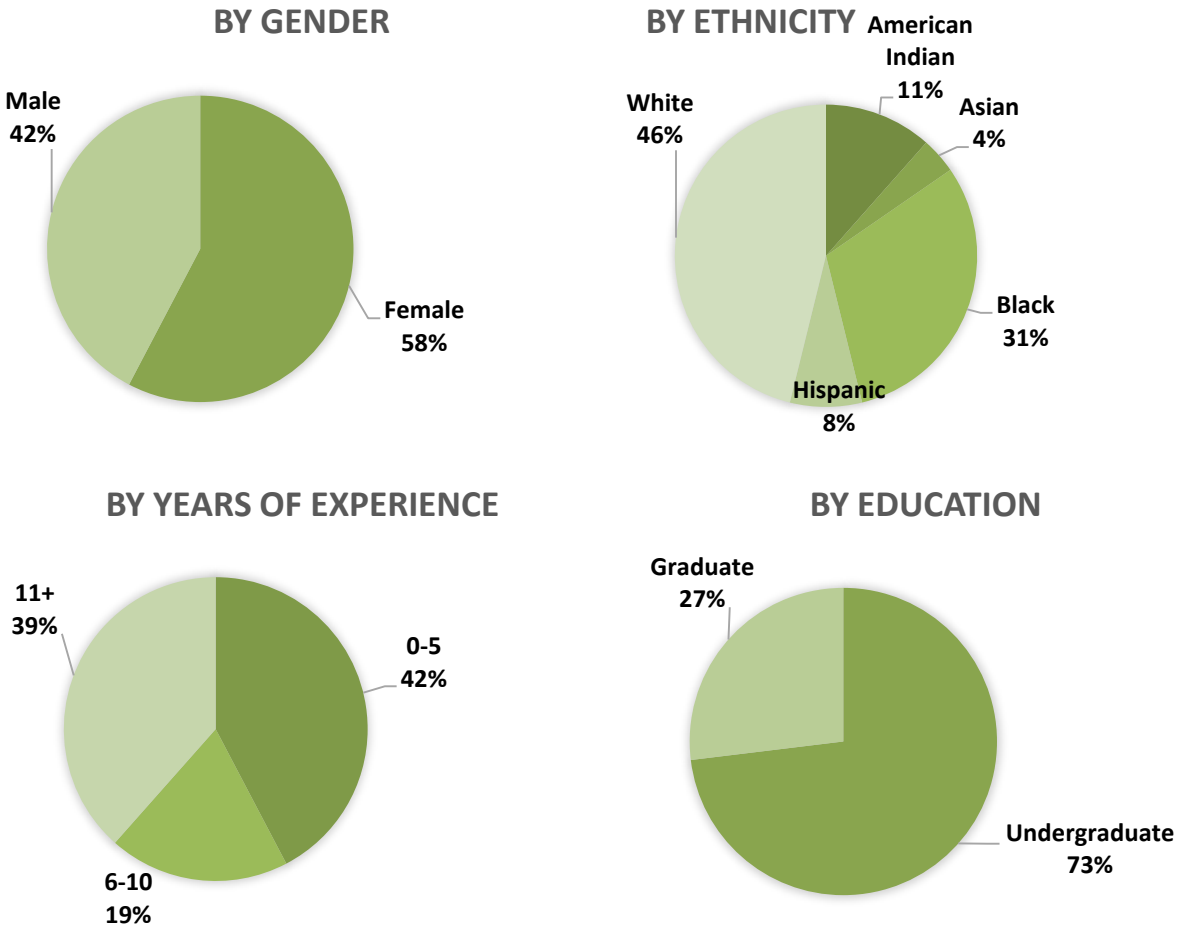
Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What 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.

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 2011– 2015 cycle, NCDPI recruited a total of 26 reviewers to serve on the bias review panel. Their demographic information is illustrated in *Figure 4.1*.

Figure 4.1 Demographic Information for Bias Review Panels from 2011–2014



Prior to 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 banked. 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?
- Is there other bias or are there sensitivity concerns?

The online review platform requires that if there is any indication that the reviewer suspects an item is associated with a bias, sensitivity, or accessibility issue then he/she explicitly documents his/her concern.

Following the review of all flagged items by the panel, a final determination must be made whether to retain or delete any of these items from the operational item pool. Items that were flagged 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 go through additional review by content test specialists at NCDPI and NCSU-TOPS. These experts included,

at a minimum, the Test Measurement Specialist, Psychometrician, and Lead Content Specialist at NCSU-TOPS. These items are only included on operational forms if no other viable alternative is available in the item bank, and all experts agree 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 adopted is to exempt from the operational pool all DIF "C-" flagged items.

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 stand-alone field test to set reasonable expectation of how long it will take students to complete each assessment. The EOG and EOC assessments were not designed to be power tests, but for practical reasons NCDPI intended to use data to set reasonable timing guidelines, which will 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 stand-alone field test, students' start and end time data were recorded. Summary data of how long it took students to complete each assessment is shown in *Table 4.5* The table includes data for Math EOG and EOC assessments administered under regular conditions—that is, no accommodations of extended time and multiple test sessions. For all grade levels except Grade 5, 75% of students completed the assessments within the 2-hour (120 Minutes) window, 99% of students in the sample took about three hours twenty minutes (200 minutes) to complete the assessment with exception in Grade 5 (230 minutes). EOG grade 5 is the first time students encountered gridded response items, and as evident it took on average about 15 to 20 minutes longer for students complete the assessment (109.3 minutes). Also, for 75% of grade 5 students, it took two hours fifteen minutes (134 minutes) to complete the test, and for 99% it took almost four hours (230 minutes) to complete.

Table 4.5 Math EOG and EOC Recorded Test Duration from Field Test 2012

			Summary		Percentile				
EOG/EOC	**N**	**Number of Items**	**Mean**	**SD**	**25th**	**Median**	**75th**	**95th**	**99th**
Grade 3	19,828	50	80.95	34.47	57	75	100	145	190
Grade 4	18,396	50	87.26	37.02	62	82	108	155	200
Grade 5	17,923	50	109.3	41.21	81	105	134	185	230
Grade 6	18,599	50	95.65	32.54	75	93	115	150	190
Grade 7	18,336	50	94.82	32.96	73	91	115	150	200
Grade 8	18,789	50	92.96	32.87	70	90	112	150	199
Math I	21,557	50	87.52	38.84	60	83	111	155	201

4.5 Step 15. Operational Test Construction

The field test plan was designed to generate enough items to construct three equivalent forms for EOG Math grades 3–8 and two equivalent forms for EOC Math I. The use of multiple forms at each grade levels 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 Content State standards 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 malpractice at the classroom level (students copying). For the EOC Math I, 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. For each grade level, one additional form was also created from the remainder of items left in the pools and published as a release form on the NCDPI website. The release forms were available to teachers, students and all interested stakeholders so they could be familiarize with the new assessment prior to operational administration.

4.5.1 Criteria for Item Inclusion in Operational Pool

Standard 3.2 of the *Standards* states:

> *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 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 (Reserved):
  - 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 choice 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 80% of the math items in grades 3–8 were retained or kept as reserve for use on the operational test. However, for EOC Math I, only 61% of items field tested met the "Keep" criteria. This was the main reason why only two base forms were created for Math I in 2012–13. These items that met the psychometric criteria provided a sufficient item pool for the construction of three parallel forms in Grades 3 through 8 and two parallel forms for Math I.

Table 4.6 Field Test 2012 Item Pool Summary for Math

Grade Level	Psychometric Evaluation Summary					
	Keep		Reserve		DELETE	
	N	Row %	N	Row %	N	Row %
Grade 3	304	61	112	22	84	17
Grade 4	338	68	106	21	56	11
Grade 5	336	67	95	19	69	14
Grade 6	324	65	103	21	73	15
Grade 7	341	68	90	18	69	14
Grade 8	306	61	106	21	88	18
Math I	182	36	112	22	206	41
Total	2,131	61	724	21	645	18

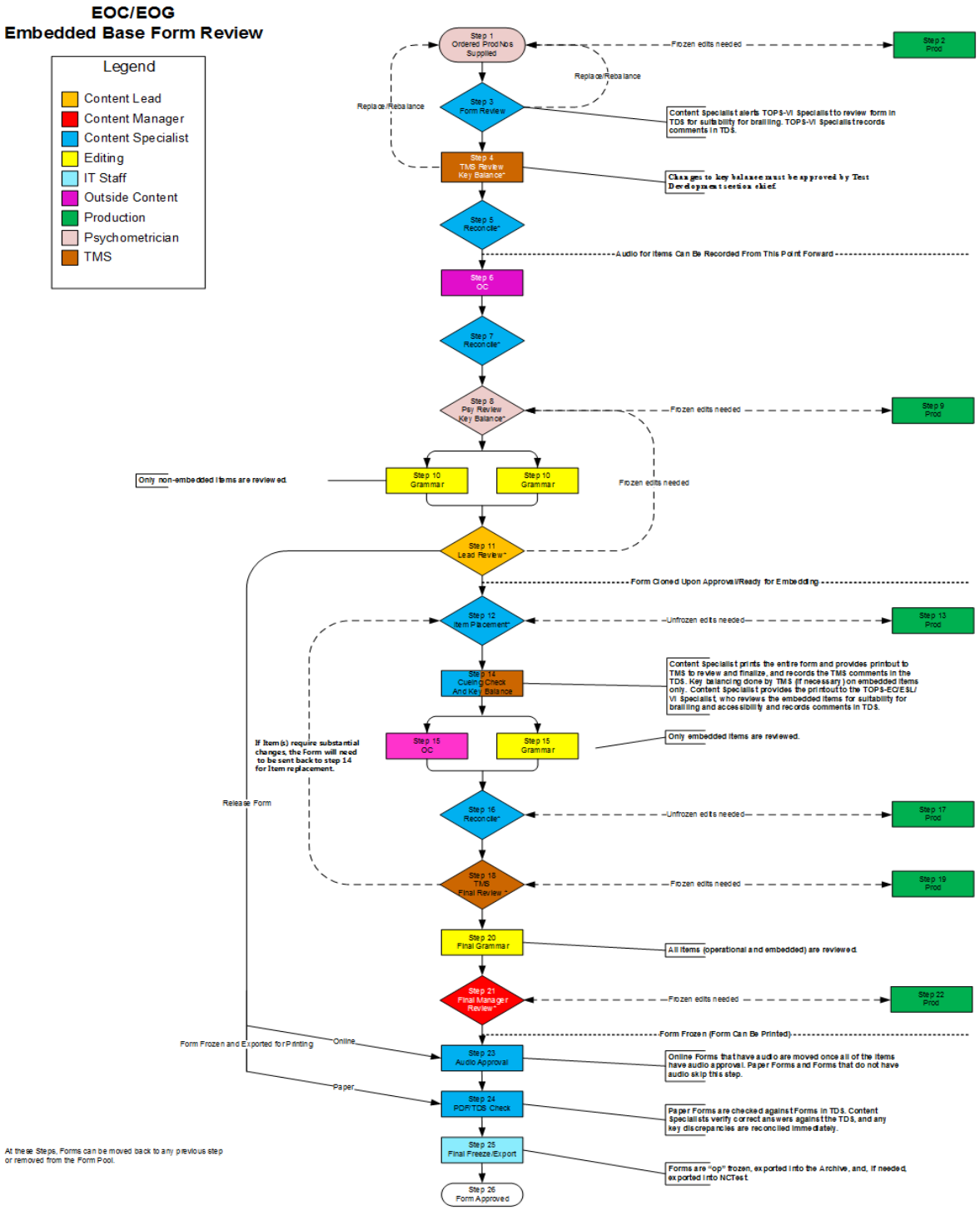
4.5.2 Operational Form Assembly

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**, **Psychometricians** build 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 ELA 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 form clears step 3, the form 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 a significant number of 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 occur when the base form with only operational items is cloned to specified numbers of versions, then field test items are selected, reviewed, and added onto each form version. Once all field test items have been approved, the form is reviewed once more by the TMS at step 18, grammar at step 20, and content manager at step 21. If there are no issues, the form is frozen and no future changes are allowed. Steps 23 through 26 are production steps where computer-based versions are produced, audio is recorded for read aloud, large prints and braille forms are created for accommodations, and final PDFs are published and printed for paper-based forms. A 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



4.5.3 Psychometric Targets based on Classical Test Theory

In setting expected form difficulty, NCDPI recognized that all item statistics were based on stand-alone field tests in 2011 when the newly adopted content standards in Math 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.

Also, a concerted effort was made to construct a developmental scale with monotonically increasing difficulty (i.e., decreasing p-value) across the grade span for math. The rationale for this was that the material covered in each subsequent grade became more complex. After reviewing the results of the scaling effort, Pacific Metrics and NCDPI determined that the data from school year 2012–2013 did not support the use of a developmental scale. The NCDPI therefore, did not adopt a developmental scale for EOG Math. In 2013, the tests covered a number of new content standards and changed the grade levels when some contents are expected to be taught. One plausible reason for the behavior of the data is that curricular and instructional practices were still adjusting to the new Mathematics standards so that they were not yet taught in the same vertical complex manner as they were tested. *Table 7.2* shows expected p-value and actual p-value summaries of operational forms based on stand-alone field test and operational statistics.

4.5.4 Psychometric Targets based on IRT Parameters

Test Characteristic Curves (TCC) generated from IRT parameters calibrated from the stand-alone 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 TCC from parallel forms will perfectly overlay each other. Furthermore, assuming that content and blueprint specifications are met, well-aligned TCC 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 -∞ to +∞ can be represented with a nonlinear monotonically increasing curve called an item characteristic curve or ICC (Hambleton & Swaminathan, 1985). The ICC curves represent a summary figure, which can be used to evaluate the statistical properties for each item. Conclusions about difficulty, discrimination, and chance 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 Curves (TCC)**

In IRT, Test Characteristics Curves (TCC) are essential for form assembly and scaling. TCC 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 ICC for all items on the test (see equation (4-6). During form assembly, items with known parameters were 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), TCCs for parallel forms plotted on the same graph is an easy way to examine the relation of summed score with theta.

$$TCC = \sum_{k}^{I} \sum_{k=0}^{k-I} KT_{ik}\,(\theta) \qquad \textbf{(4-6)}$$

- **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 errors, 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, p117) highlighted, 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 test. 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. For more on this see Hambleton & Swaminathan (1985), and Thissen & Orlando (2001). Some features of TIF as noted in Hambleton & Swaminathan (1985, p104) are:

- 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.

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i\,(\theta)^2}{P_i\,(\theta) Q_i\,(\theta)} \tag{4-7}$$

  (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 4.9* display TCCs for parallel operational forms assembled based on field test item parameters for each grade level. 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 TIF & CSE Plots Based on Field Test Parameters-Math. 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 3 TCC Math Forms A, B, and C

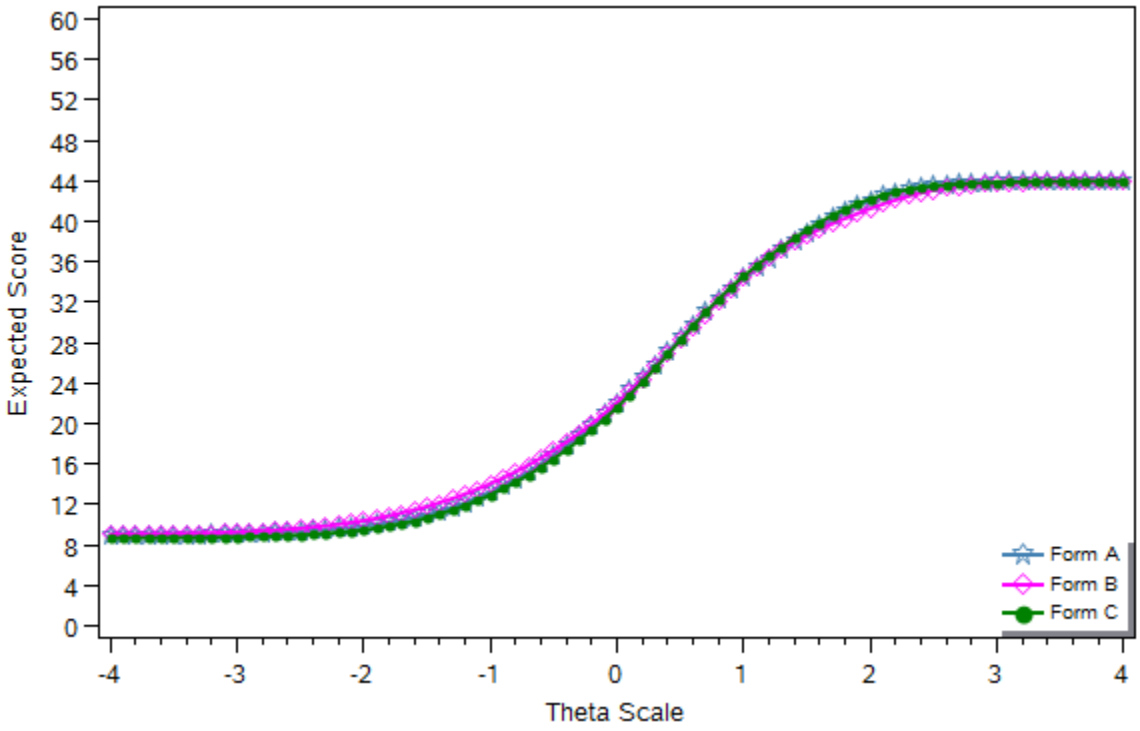


Figure 4.4 EOG Grade 4 TCC Math Forms A, B, and C

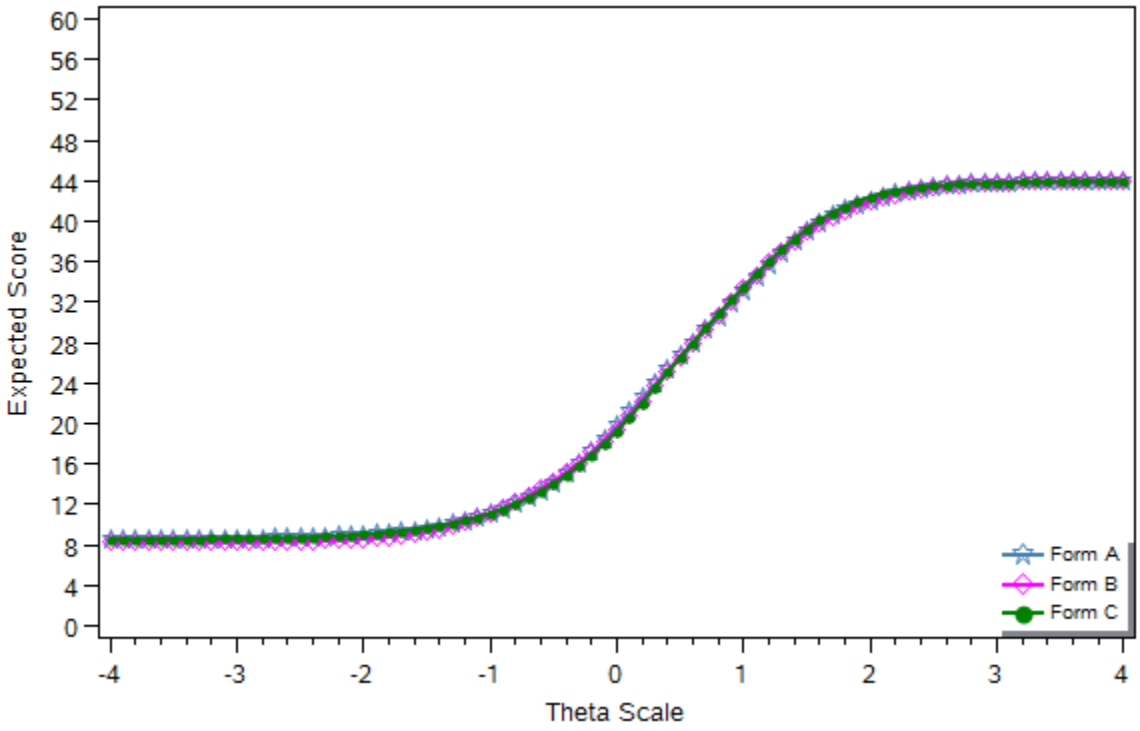


Figure 4.5 EOG Grade 5 TCC Math Forms A, B, and C

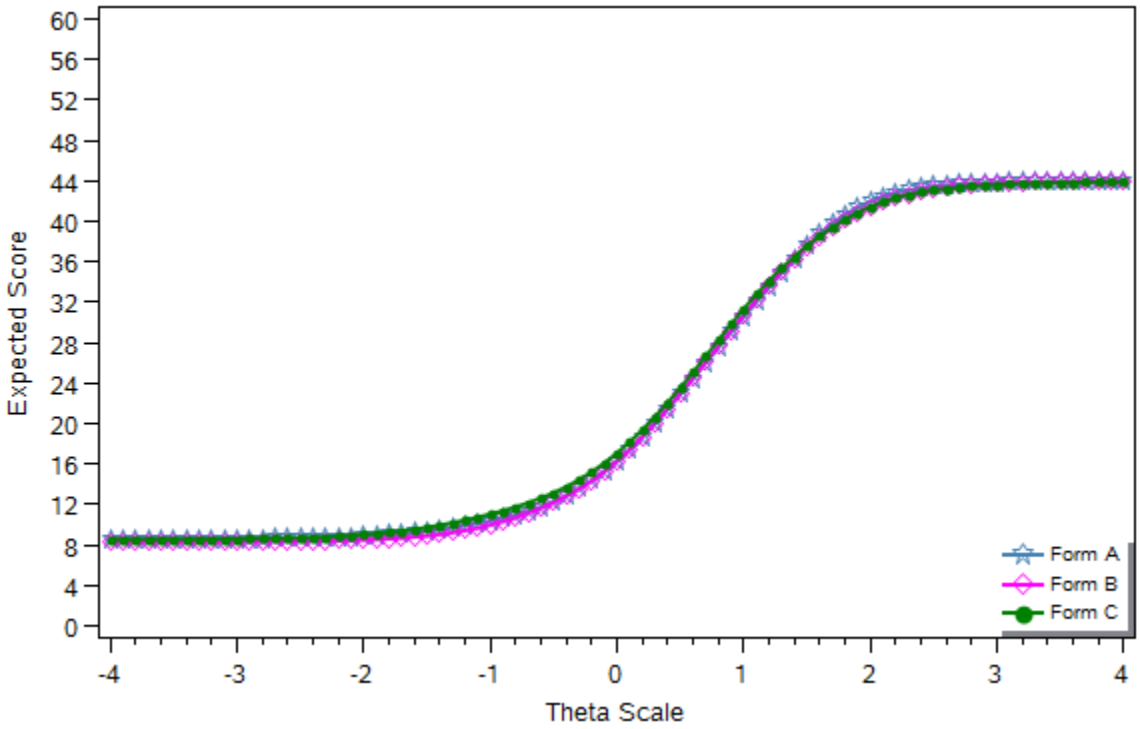


Figure 4.6 EOG Grade 6 TCC Math Forms A, B, and C

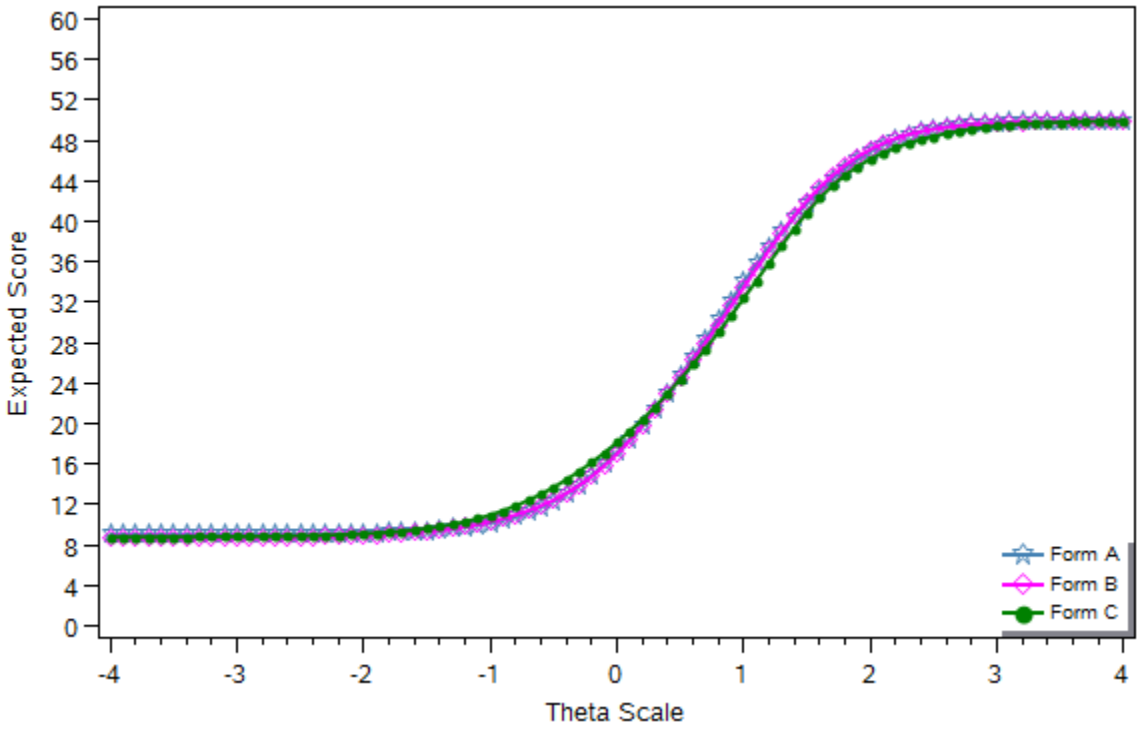


Figure 4.7 EOG Grade 7 TCC Math Forms A, B, and C

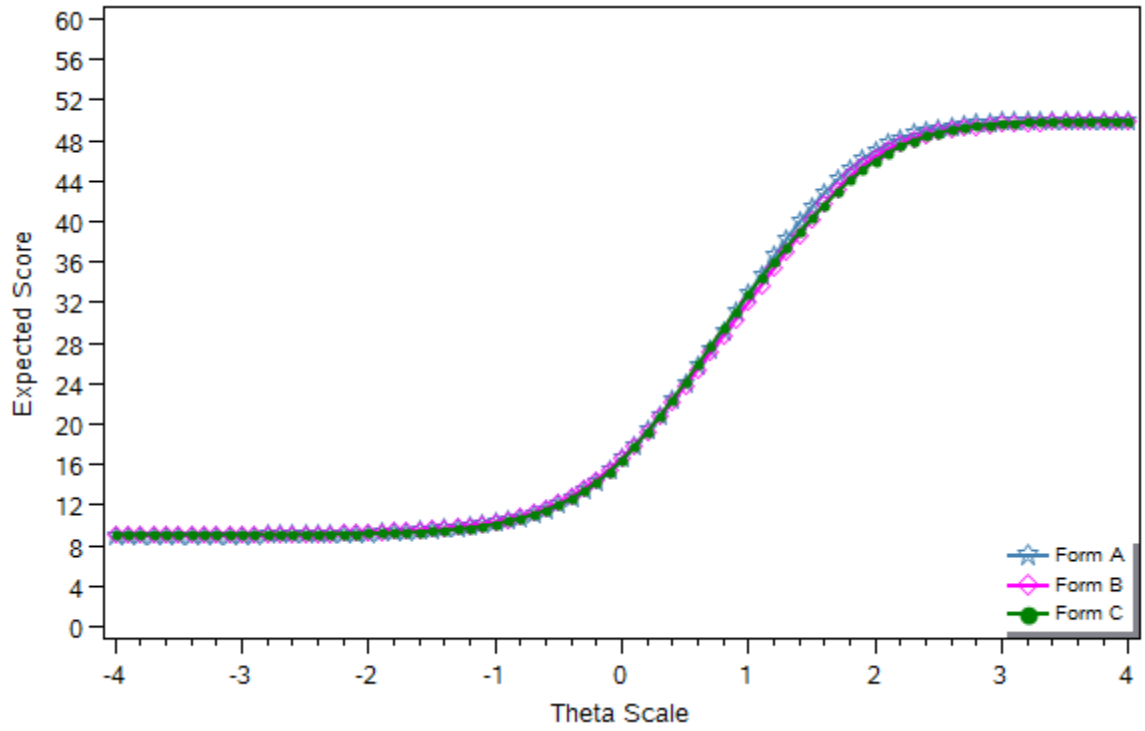


Figure 4.8 EOG Grade 8 TCC Math Forms A, B, and C

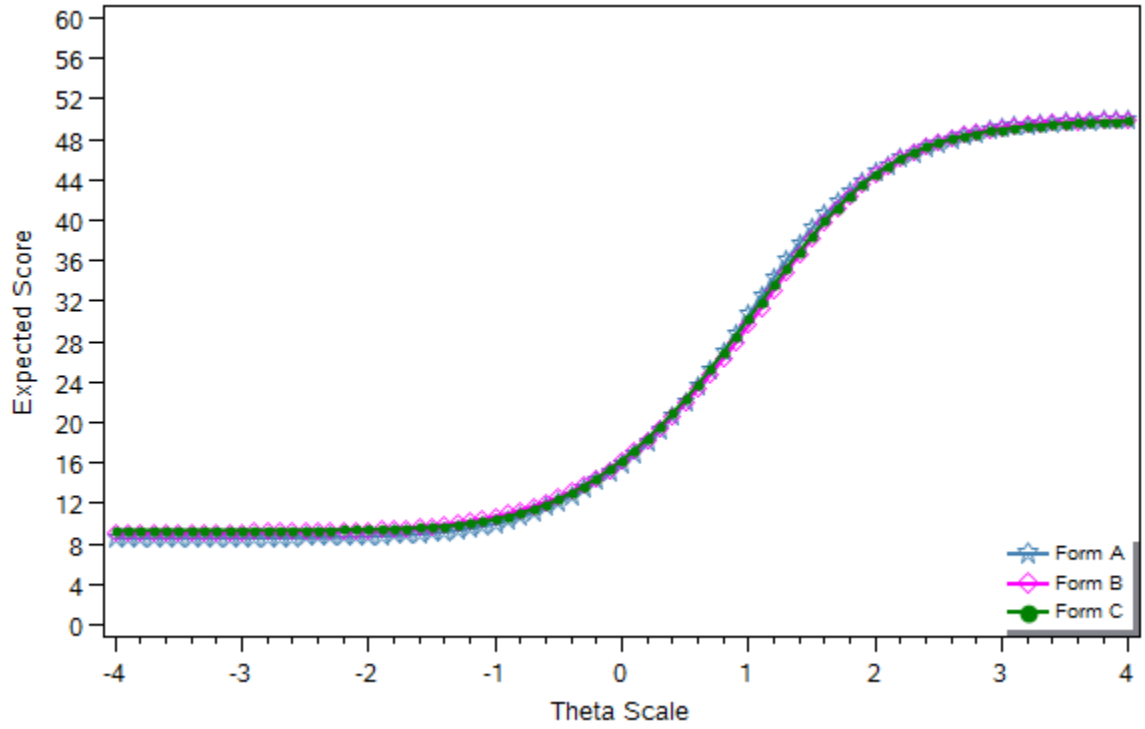
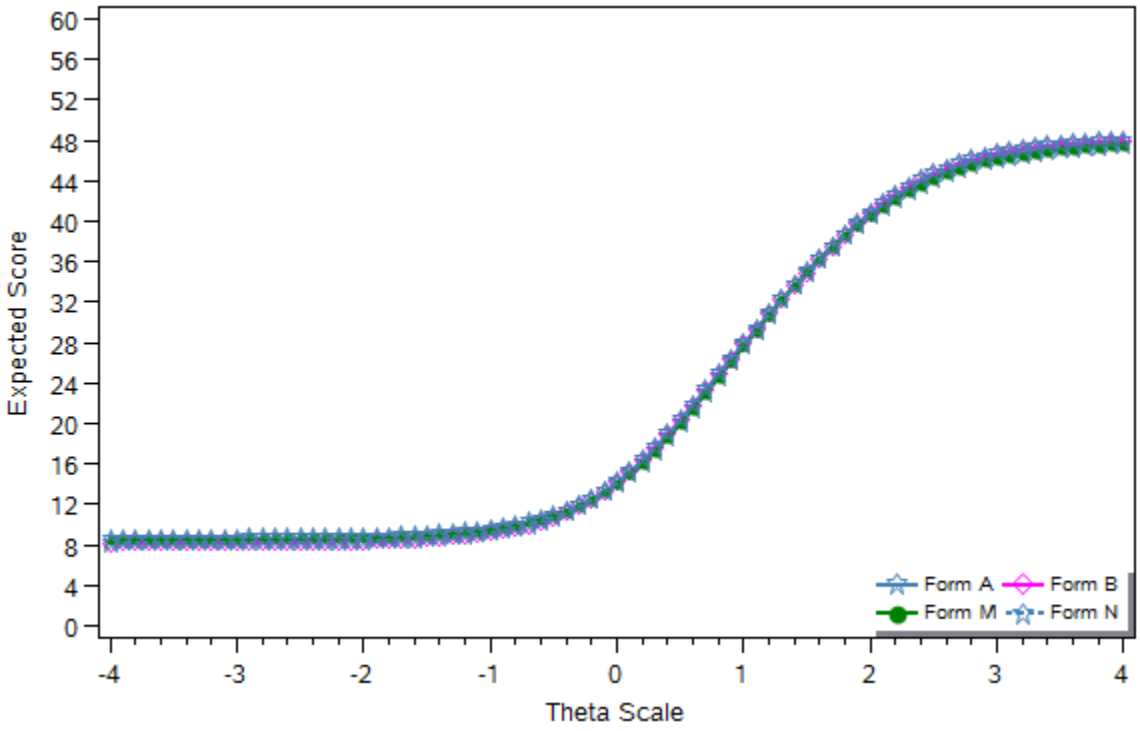


Figure 4.9 EOC Math I TCC forms A, B, M, and N



4.6 Step 16. Review of Assembled Operational Test Forms

Once forms were assembled to meet content specifications, test blueprints, target p-values, and target IRT item parameter, were sent to outside content experts (see

*Figure **3.5***) who provided an independent outside review of all assembled forms. Criteria for evaluating each test form included the following:

- The content of the test forms reflects the goals and objectives of the North Carolina *Standard Course of Study* for the subject (content validity).
- The content of 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 are balanced in relation to ethnicity, gender, socioeconomic status, and geographic district of the state (free from test/item bias); and
- 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 recorded his or her opinion as to how well the tests met the five criteria listed above in TDS. Form reviewer comments were recorded in TDS were reviewed by NCDPI and NCSU-TOPS content specialist. 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 due to cuing concerns, overemphasis on 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 Review of Computer-based Forms

After computer-based forms were exported from the Test Development System (TDS) application into the NCTest platform, a series of quality checks were perform to ensure all the

specified interactions between items and the NCTest platform were fully functional across the different end users' approved devices. NSCU-TOPS and the NCDPI technology sections have instituted a five-phase quality check system that focuses on issues ranging from technical and network comparability aspects to accessibility aspects, like verifying that high contrast, large font, read aloud files 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. Each form is assigned to a demo student for all the different presentation types (high contrast, large font, read aloud) available during operational administration. In Phase 2, NCSU-TOPS employees conduct quality checks to ensure the correctness of the forms and the items themselves. The Editing/Production groups are 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 like highlighting, audio playback, and 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 corrections, and all corrections are monitored and verified as complete by the NCDPI.

Chapter 5 Test Administration

This chapter of the technical report describes the materials and activities in which NC DPI engaged in order 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, it could undermine the comparability of the resulting test scores. This chapter presents the efforts made to standardize test administration for the NC assessments in order to reduce construct-irrelevant variance that could undermine the comparability of test scores.

5.1 Test Administration Materials

NC DPI prepared materials prescribing the means for administering the NC EOG and EOC assessments. This section describes test administration materials prepared by the NCDPI that are made available to test administrators to ensure standardized administration of EOG and EOC assessments across the state. As referenced in standard 6.1 of the *Standards,* "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).

For every assessment and grade level the NCDPI produces two comprehensive guides:

- 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 which covers: the purpose of the assessment, eligible students, testing window, 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.
- The Proctor Guide: The Proctor guide serves as the source document with detailed guidelines on selecting proctors, defining their roles, and training information. Key training topics covered in the proctor's guide includes defining proctors' responsibility, training on how to maintain test security, ensure appropriate testing conditions, maintain students' confidentiality, assist test administrator, monitor students, report test irregularities and follow appropriate procedures for accommodations.

The NCDPI also provides a guideline training manual for testing students identified as English Language Learners (ELL). This guide provides training on the following areas: ELL

testing requirements, responsibilities of test coordinators, procedures for participation, testing accommodations available, and monitoring accommodations.

Standard 4.15 states: "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 Training for Test Administrators

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 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 special needs (students with disabilities and students with limited English proficiency), 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, NCDPI has taken extensive steps to ensure the security of the assessments by establishing protocols for school employees administering tests, protocols for handling and administering paper tests, and protocols 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 *Test Administrator's Manual* and the codified North Carolina *Testing Code of Ethics* (see Appendix 2-A) prior to actual test administration. Test administrators must also follow the instructions given in the Test Administrator's Manual to ensure a standardized administration and 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,* 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 building-level 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 school system test coordinator. Upon notification, the school system test coordinator must report the discrepancies to the regional accountability coordinator and ensure all procedures in the Online Testing Irregularity Submission System 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 school system 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

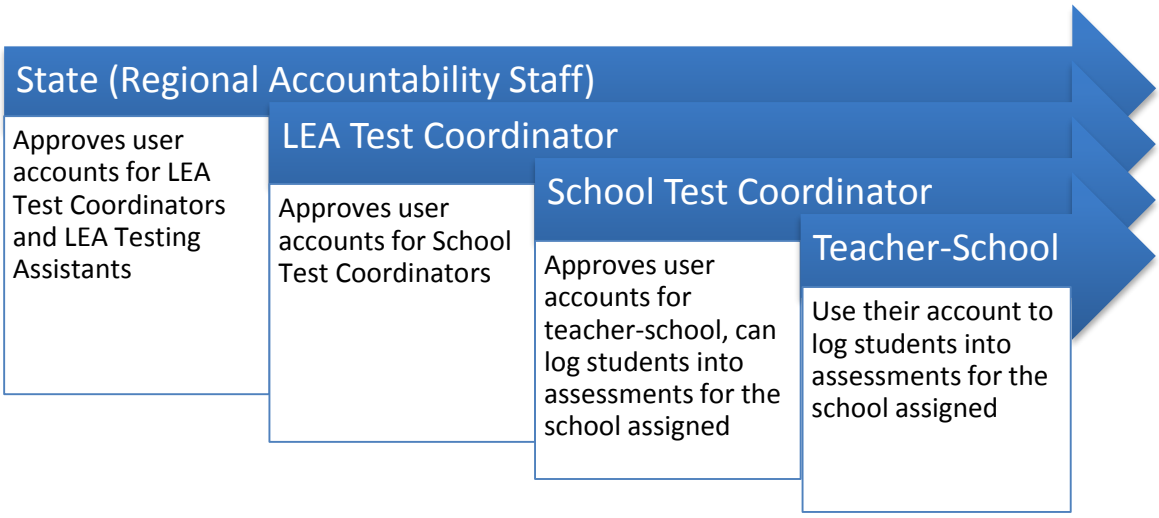
Test Material	Required Storage Time
All used answer sheets for operational tests (including scoring sheets for W-APT)	Six months after the return of students' test scores
Original responses recorded in a test book, including special print version test books (i.e., large print edition, one test item per 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	Store indefinitely
EOC: Math I, Biology, and English II	Retain unused test materials from fall for use in spring; retain unused test materials from spring for use in summer
W-APT test materials (reusable except for scoring sheets)	Store indefinitely (all forms)

5.3.3 Computer Mode Test Security Measures

The 2012–13 operational EOC Math I assessment was available in both computer and paper modes. The NCTest platform is used to administer computer-based, fixed form assessment. The NC Education system manages student enrollments, monitors assessment start and stoppage times, and manages accommodation needs.

NCDPI limits all LEA access to the computer-based assessment to specific testing days. An LEA's 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*



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 login, 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 are sent at that time. Graphics and audio files (for computer read aloud accommodation) are sent as students move from item to item within the assessment.

Student responses are securely sent after each item is answered 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 cache and cookies from local machines.

After online student assessments are finalized, they are transferred nightly to the NCDPI and/or to 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 the *NCDPI's Maintaining the Confidentiality and Security of Testing and*

Accountability Data Guidance. The NCDPI systems 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 3–8 were required to participate in the EOG assessments administered within the last 15 days of the school year. Based on the traditional school calendar, EOG assessments are administered in late spring of the school academic calendar.

The EOC has two administration windows: one in fall and another in spring. Students enrolled in a semester schedule are required to take EOC assessment with the last 15 days of the semester. Students enrolled in a yearlong course schedule are administered the EOC assessment within the last 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 3–8 are administered the EOG assessment during the last ten days of the school year. The testing window for the EOC assessment was also modified. Beginning with the 2013–14 school year, the EOC administration window was changed to the last five days of the instructional period for the semester courses or the last 10 days of the instructional period for the yearlong courses. Districts can request a waiver to increase the testing window by five days.

5.4.2 Timing Guidelines

The Math 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* (2014) 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" (p.51). In keeping with the *Standards* (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 test and analyzed in section 4.4, the NCDPI recommended time allotted for both the EOG Math and EOC Math I is 180 minutes, with a maximum of 240 minutes. Students with approved accommodations may take even longer as specified by their particular Individualized Education Plan (IEP).

5.4.3 Testing Accommodations

State and federal law requires that all students, including students with disabilities (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 are provided with "test accommodations, when appropriate and feasible, to remove construct-irrelevant barrier 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 IEP, Section 504 Plan, ELL documentation, or transitory impairment documentation, and (2) the documentation must reflect routine use 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 guides:

- Computer Reads Test Aloud—Student Controlled (computer-based assessments only; not approved for ELA EOG grades 3–8 and EOC English II)
- 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
- Interpreter/Translator Signs/Cues Test (not approved for ELA EOG grades 3 – 8 and EOC English II)
- 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 prior to the test administration.

According to the *Standards*, an appropriate accommodation addresses a 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, the modified assessment should be treated like a newly developed assessment. The NCDPI assessment guide recommends that students should only be allowed the same accommodations for assessments as those routinely used during classroom instruction and other classroom assessments that measure the same construct.

5.4.4 English Language Learners

Per State Board policy GCS-C-021, students identified as English Language Learners (ELL)[i] 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 WIDA-ACCESS Placement Test and are in their first year in United States schools are exempt from taking the ELA EOG assessment or the English II EOC assessment.

For both EOG and EOC, ELL students are provided with an ELL reading accommodation 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 Math include:

- Multiple testing session
- Scheduled extended time
- Testing in a separate room
- Student read aloud to self
- English/Native Language word-to-word Bilingual Dictionary/Electronic Translator
- Test administrator reads test aloud in English

[i] 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®.

- 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

<table>
<tr><th>Subtest</th><th>1
Entering</th><th>2
Emerging</th><th>3
Developing</th><th>4
Expanding</th><th>5
Bridging</th><th>6
Reaching</th></tr>
<tr><td>Reading</td><td colspan="4">Eligible to Receive State-Approved ELL Testing Accommodations for All State Tests</td><td colspan="2">Must Participate in General State Test Administration without ELL Testing Accommodations</td></tr>
</table>

5.4.5 Mode of Test Administration

The EOG assessments may be administered in either as paper or computer-based fixed forms. The state's goal is to gradually transition test administration for EOG and EOC to the computer mode as districts are able to build their resources and technology capacity. For the 2012–13 administration, all EOGs were administered in paper mode. Beginning with the 2014–2015 administration, the grade 7 EOG ELA/reading and math was available in both paper and computer mode.

The EOC Math I assessment was developed as a computer-based fixed form. Districts could opt to use paper-based forms in place of the computer-based form. *Table 5.2* shows the total number of students who took the Math EOG and EOC tests by mode during the 2013, 2014, and 2015 test administration windows. As shown in the table, the percentage of students who are administered the computer-based EOC forms continues to increase gradually from 2013 to 2015. In 2015, 55% of students took Math I computer-based forms compared to 52% in 2013. EOG

computer-based forms were administered for the first time in 2015 at grade 7, and approximately 21% of students took the computer-based form.

Table 5.2 Math EOG and EOC Test Administered by Mode

Type and Year		Test Administration Mode			
		Paper Mode		Computer Mode	
		Total Test	Percent	Total Test	Percent
EOG Grade 3	*2013*	106,518	100%		
	2014	116,083	100%		
	2015	118,510	100%		
EOG Grade 4	*2013*	114,669	100%		
	2014	107,388	100%		
	2015	115,798	100%		
EOG Grade 5	*2013*	114,435	100%		
	2014	115,544	100%		
	2015	108,385	100%		
EOG Grade 6	*2013*	116,314	100%		
	2014	115,280	100%		
	2015	116,500	100%		
EOG Grade 7	*2013*	115,381	100%		
	2014	117,606	100%		
	2015	92,935	79%	24,143	21%
EOG Grade 8	*2013*	112,944	100%		
	2014	116,256	100%		
	2015	118,869	100%		
EOC Math I	*2013*	61,247	48%	65,893	52%
	2014	56,684	46%	65,337	54%
	2015	55,763	45%	69,521	55%

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 (16 NCAC 6G.0305[g]). For the EOG, all students in grades 3 through 8 are required to participate in the end-of-grade assessments or the corresponding alternate assessment, as indicated by the student's Individualized Education Program (IEP) or

appropriate ELL documentation. For the EOC, all students enrolled in Math I must be administered the EOC test. 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 district-wide 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 district-wide and/or state-mandated tests to students and parents or guardians within 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 due to 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 http://www.ncpublicschools.org/docs/accountability/1516medexcept.pdf.

Chapter 6 Scoring and Scaling

This chapter describes the processes used for scoring items and procedure adopted to create final reportable scale scores. The first section of this chapter summarizes the automated scoring procedures to transform students' responses into a number correct score for fixed response items. Section two and four 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 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 Fixed Response Items

The NCDPI WinScan software program is used for scoring all EOG responses. WinScan is a specialized scoring and reporting software program created and managed by the NCDPI accountability division. At the beginning of each testing window, a new release of WinScan is updated and distributed to all LEAs and charter schools. 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 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. 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-based forms are scored electronically via a centrally-hosted server at NCDPI using WinScan software. Once WinScan assigns scores for each item, data are then merged with student-level records then electronically made available to test coordinators.

Once the data are available, school system test coordinators can generate school rosters, class rosters, and individual reports. Initial district school-level reporting occurs at the LEA level. North Carolina Administrative Code (i.e., 16 NCAC 06D .0302) requires districts to report scores resulting from the administration of district-wide and State-mandated tests to students and parents or guardians along with available score interpretation information within 30 days from generation of the score at the district level or from the receipt of the score and interpretive documentation from the department.

Student's response choices for gridded response items are re-evaluated again before the scores are certified, any recorded response format not previously accounted for in the WinScan scoring key list for these items are verified and updated to ensure all valid response choice are properly scored.

6.2 Scale Scores

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

- They provide a standard metric 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 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.

As stated in 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 transformed raw scores into scale scores. For in-depth review of the procedure see Thissen and Orlando (2001, p. 119). Summary of the procedure for creating summed scores as described by Thissen and Orlando is as follows:

For any IRT model with item scores indexed ($u_i = 0,1,$), the likelihood for any summed scores $x = \sum u_i$ is:

$$L_x(\theta) = \sum_{\sum u_i = x} L(u/\theta)$$

(6-1)

Where $L(u/\theta) = \prod_i T(u_i/\theta)$ 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

$$P_x = \int L_x(\theta) g(\theta)$$

(6-2)

And the expected θ associated with each summed score is

$$E(\theta/x) = \frac{\int \theta L_x(\theta) g(\theta)}{P_x}$$

(6-3)

With posterior standard deviation (PSD) given by

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

(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. The mean and standard deviation of the scale scores of the grades 3–8 math EOG tests were set to be 450 and 10, and EOC Math I was 250 and 10. 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 was administered.

6.3 Data Certification

Prior to the release of test scores for official reporting, NCDPI performs data certification to ensure all items, both automated and hand scored, were correctly scored and captured and that there were no issues reported during administration. 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 two main quality control steps: (1) independent scoring of student responses, and (2) computing CTT statistics and comparing to the field test.

During the first step, 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 2 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 due to a wrong scoring key or 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 the generation of 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 the EOG and EOC in 2012– 13. The chapter begins with a description of the random spiraling process used to administer three parallel forms across North Carolina. This chapter summarizes item analysis results from the operational administration in 2012– 13, which includes CTT (p-value, biserial correlations, Cronbach's alpha reliability estimate) 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 testing program uses a pre-equating model base on IRT to score test forms and compute raw-to-scale tables for each form prior to operational administration. This model allows the department to satisfy NCSBE policy GCS-A-001 "… School systems shall report scores resulting from the administration of district-wide 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." (Page 43 of the *Test Coordinator Manual*).

For the first administration of the North Carolina READY 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 parallel forms were constructed using data from stand-alone field tests. Field test data are usually unstable, and it is common to experience drift in item parameters between a stand-alone 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, NCDPI wanted to reanalyze all forms based on operational data to ensure item parameters and scale scores used for standard setting to set achievement levels were stable to be used as baseline.

7.2 Spiraled Form Administration

Three parallel forms in EOG grades 3–8 (A, B, C) and four parallel forms in EOC Math I (A, B, M, N) were administered operationally for the first time in the 2012–13 school year. At every grade level, all alternate forms were administered to randomly equivalent groups of examinees. Within each grade, the forms were spiraled within the classroom. Spiraling forms ensures that item parameter 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 a demographic descriptive summary for students who were administered Math EOG and EOC in 2012–13. The student counts listed in these tables is 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 3 through 8 and in high school were administered the EOG Math or EOC Math I assessments. For EOG grades 3–8 at least 35,000 were administered one of the three parallel forms. The differences across forms within a grade are negligible, which is evident of the success of the random spiral process. In EOC Math I, about 32,000 students were administered one of the two computer-based alternate forms, and about 30,000 students were administered one of the two alternate paper based forms.

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

Table 7.1 Student Demographic Summary for Math EOG and EOC Operational Test 2012–2013

Grade and Form		N	Gender (%)		Ethnicity (%)						
			Female	Male	Asian	Black	Hispanic	American Indian	Multi-racial	Native Hawaiian/ Pacific Islander	White
Grade 3	**A**	35,550	48.57	51.43	2.87	24.17	15.58	1.31	4.18	0.08	51.80
	B	35,523	48.71	51.29	2.78	24.49	15.33	1.38	4.04	0.08	51.90
	C	35,163	49.41	50.59	2.91	24.35	15.54	1.32	4.06	0.07	51.75
	All	**106,236**	**48.89**	**51.11**	**2.85**	**24.34**	**15.48**	**1.34**	**4.10**	**0.08**	**51.82**
Grade 4	**A**	38,256	49.05	50.95	2.84	24.76	15.27	1.50	3.99	0.09	51.54
	B	38,163	48.98	51.02	2.72	24.72	15.19	1.43	3.94	0.08	51.91
	C	37,900	49.10	50.90	2.80	24.67	15.16	1.35	4.05	0.08	51.89
	All	**114,319**	**49.04**	**50.96**	**2.79**	**24.72**	**15.21**	**1.43**	**3.99**	**0.08**	**51.78**
Grade 5	**A**	38,109	49.27	50.73	2.81	25.69	14.66	1.39	3.87	0.09	51.49
	B	38,043	48.73	51.27	2.71	25.17	14.85	1.32	3.88	0.12	51.94
	C	38,000	49.11	50.89	2.78	25.31	15.04	1.39	3.64	0.08	51.76
	All	**114,152**	**49.04**	**50.96**	**2.77**	**25.39**	**14.85**	**1.37**	**3.80**	**0.10**	**51.73**
Grade 6	**A**	38,796	49.16	50.84	2.62	26.05	14.35	1.38	3.58	0.10	51.93
	B	38,652	48.97	51.03	2.54	26.03	14.02	1.38	3.76	0.09	52.18
	C	38,326	49.00	51.00	2.68	26.07	13.83	1.41	3.57	0.08	52.37
	All	**115,774**	**49.05**	**50.95**	**2.61**	**26.05**	**14.07**	**1.39**	**3.64**	**0.09**	**52.16**
Grade 7	**A**	38,428	49.37	50.63	2.51	26.33	13.29	1.52	3.58	0.09	52.68
	B	38,394	48.65	51.35	2.70	26.22	13.23	1.50	3.52	0.09	52.75
	C	38,003	49.41	50.59	2.63	26.25	13.10	1.49	3.52	0.10	52.91
	All	**114,825**	**49.14**	**50.86**	**2.61**	**26.27**	**13.21**	**1.50**	**3.54**	**0.09**	**52.78**
Grade 8	**A**	37,778	49.34	50.66	2.57	26.91	12.34	1.48	3.44	0.11	53.16
	B	37,452	49.33	50.67	2.59	26.51	12.49	1.44	3.51	0.12	53.35
	C	37,326	49.48	50.52	2.44	26.29	12.44	1.40	3.46	0.08	53.89
	All	**112,556**	**49.38**	**50.62**	**2.53**	**26.57**	**12.42**	**1.44**	**3.47**	**0.10**	**53.46**
Math I	**A**	30,685	48.19	51.81	3.31	30.64	12.09	2.86	3.42	0.10	47.57
	B	29,748	49.13	50.87	3.29	30.47	12.27	2.95	3.45	0.10	47.47
	M	32,349	48.99	51.01	1.75	23.88	12.01	1.08	3.65	0.10	57.54
	N	31,978	48.84	51.16	1.79	23.71	12.13	1.05	3.54	0.12	57.67
	All	**124,760**	**48.79**	**51.21**	**2.51**	**27.07**	**12.12**	**1.95**	**3.52**	**0.11**	**52.72**

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 tables and provide item statistics and student level data for standard setting study. 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 IRT Calibration for Parallel Forms

To evaluate the overall impact of item parameter drift, the parallel forms' test characteristic curves created from field test statistics were re-evaluated using operational administration data. Using the psychometric criteria presented in section 4.5.1, all items were re-evaluated based on their operational item parameters, and problematic items were effectively removed from the form before final item calibration. No items from EOG forms were dropped from the operational set. Single-group 3PL IRT model for multiple-choice items and 2PL IRT model if there were gridded response items were used in each calibration to establish the final IRT parameters for scaling. In IRT, the need for equating is a non-issue if parameters from alternate forms are put on the same IRT scale either through the data collection design—as is the case with random spiraling of forms—or through the concurrent calibration method. Once all items are calibrated onto the same IRT scale, then raw-to-scale tables are created for each alternate form, and scores from parallel forms can be used interchangeably. The data collection design together with the IRT calibration method applied provide evidence referenced in standard 5.12 of the *Standards* which states "A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably" (p. 105).

7.3.2 EOC IRT Calibration Across Modes

For Math I, all operational items in the two pairs of parallel forms (A and M, B and N) created from field test data were reviewed using the psychometric criteria presented in section 4.5.1. Following these analyses, 1 item was removed from each pair of EOC Math I parallel forms. These forms are marked with an asterisk in *Table 7.3*.

Concurrent calibration with differential item functioning (DIF) sweep in IRTPRO was used for each pair of parallel forms across modes to establish final parameters. The DIF sweep option in IRTPRO (Cai, Thissen, & du Toit, 2011) 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 a 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 effect were classified as candidate DIF items. During the second step, items that did not show any mode effect 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 were 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.

7.3.3 Parallel Forms Test Characteristic Curves (TCC)

Figure 7.1 through Figure 7.7 show TCCs computed from post administration parameters for parallel forms. The TCC plot shows 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 TCC for parallel forms across the entire ability range. TCC for parallel forms across grades show small variations at different sections along the ability scale. Small variations in TCC of parallel forms are tolerated and accounted for in the raw-to-scale tables. Also, students' experiences are not noticeably different across forms, as there are no artificial restrictions of range imposed by taking a form that is 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 4.9*. Major differences between the TCCs from operational and stand-alone field test administration are that the gradient of the operational TCCs is slightly lower, and the steepest sections of the TCCs from the operational analysis are slightly shifted to the left of the ability scale, indicating the forms had gotten easier.

Figure 7.1 Grade 3 TCC Math Operational Forms A, B, and C

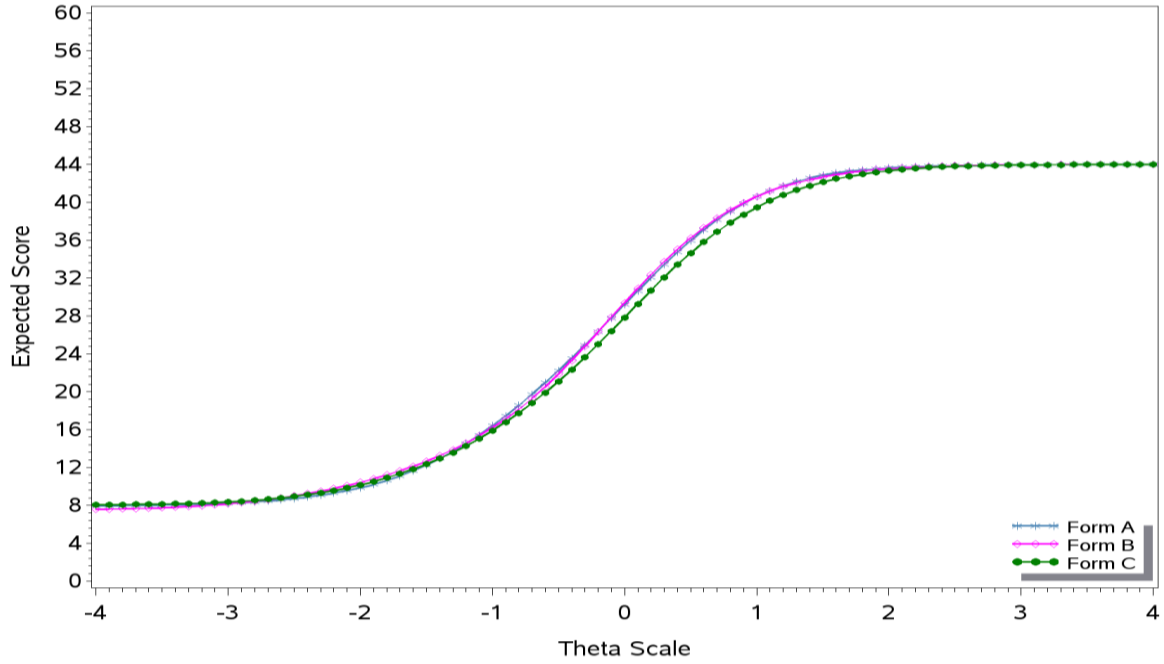


Figure 7.2 Grade 4 TCC Math Operational Forms A, B, and C

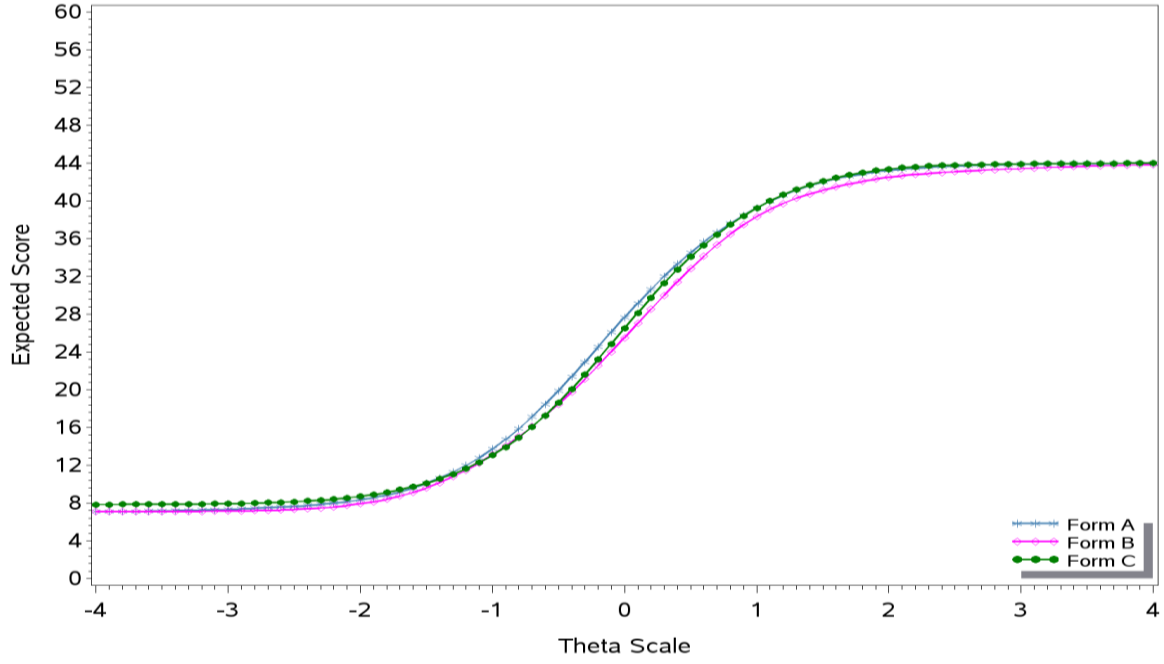


Figure 7.3 Grade 5 TCC Math Operational Forms A, B, and C

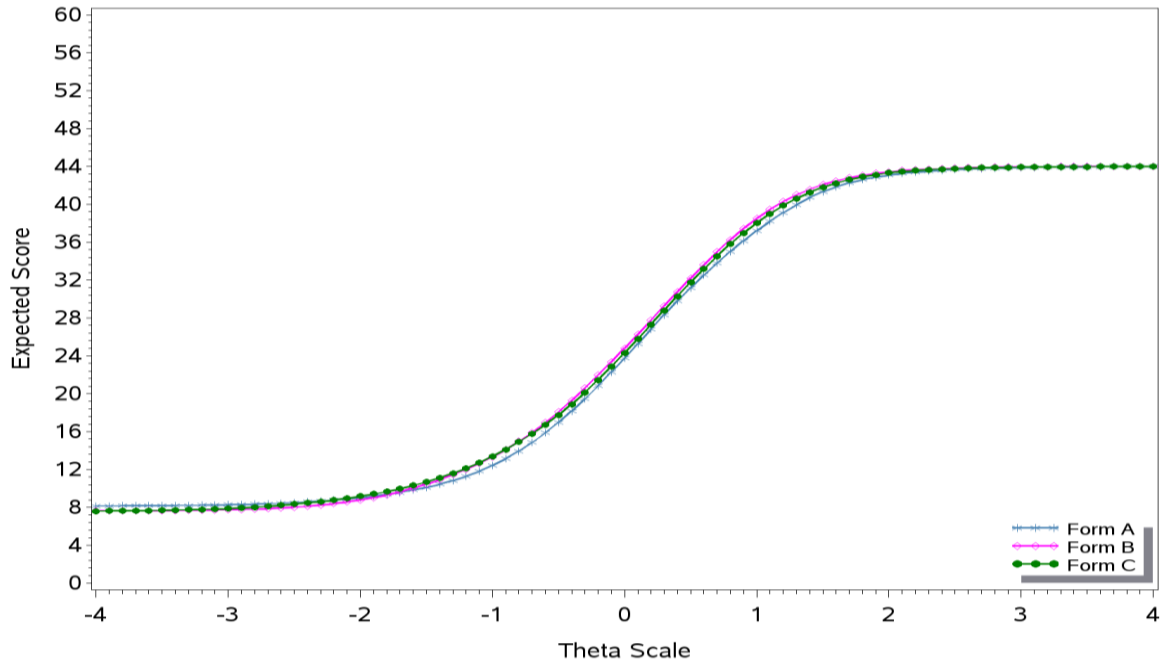


Figure 7.4 Grade 6 TCC Math Operational Forms A, B, and C

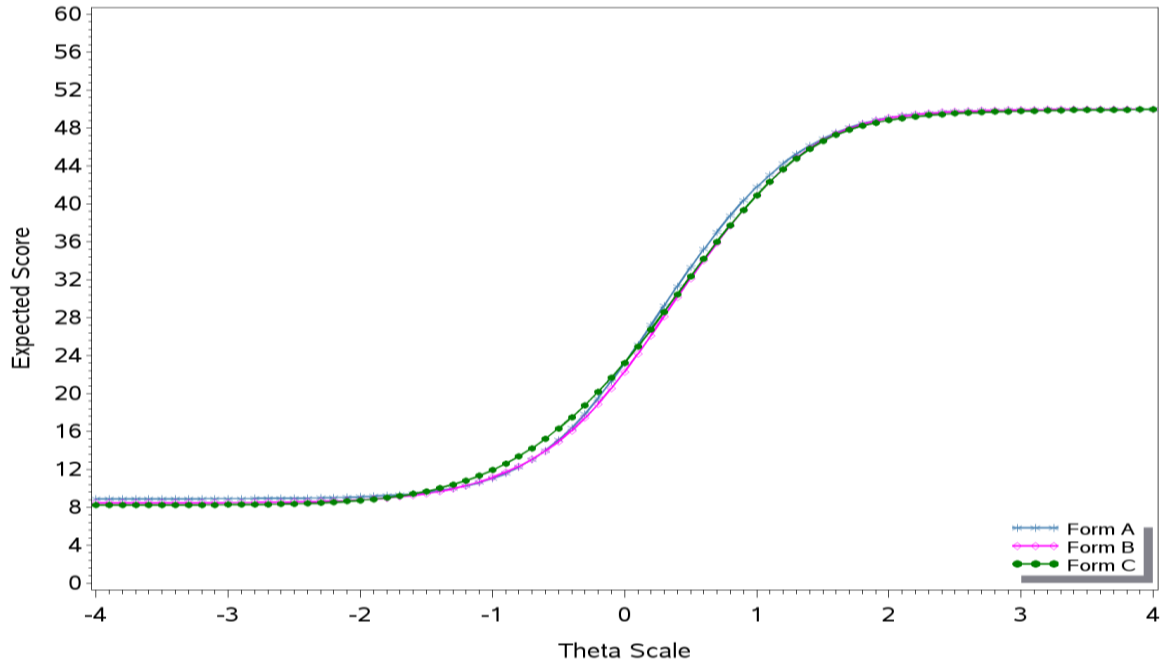


Figure 7.5 Grade 7 TCC Math Operational Forms A, B, and C

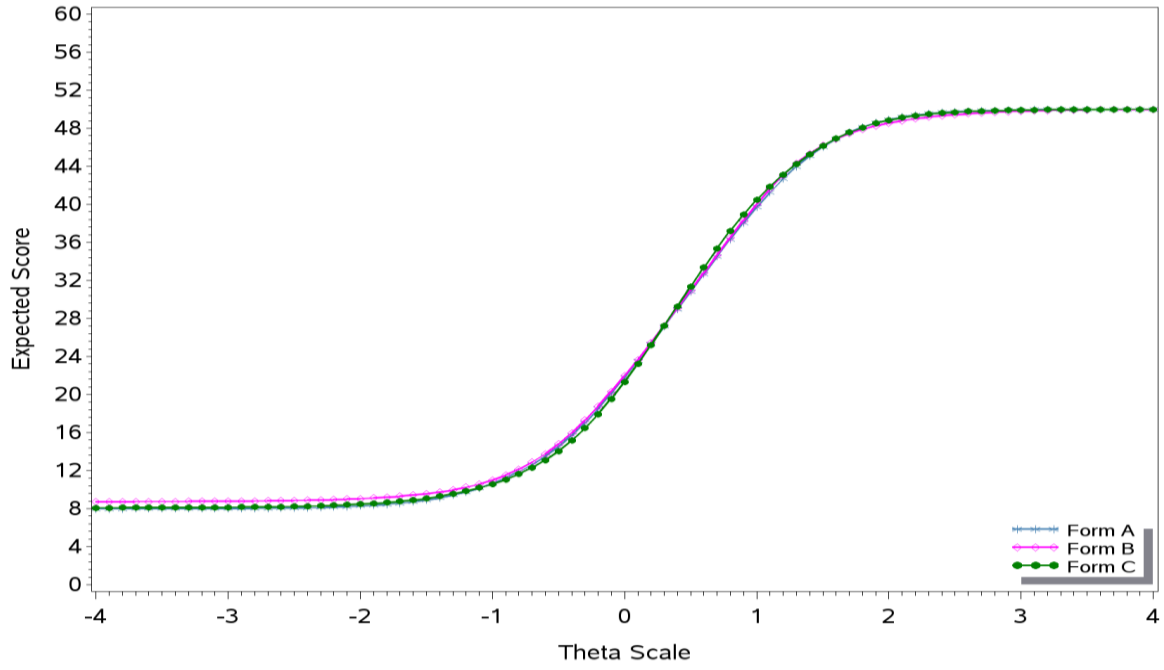


Figure 7.6 Grade 8 TCC Math Operational Forms A, B, and C

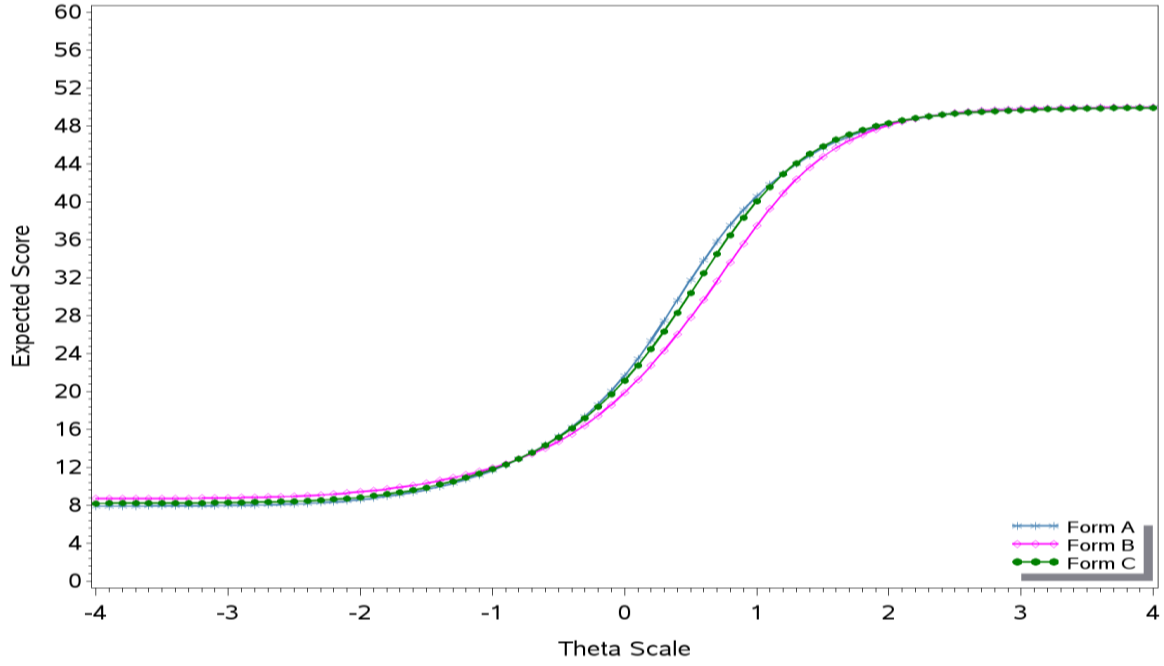
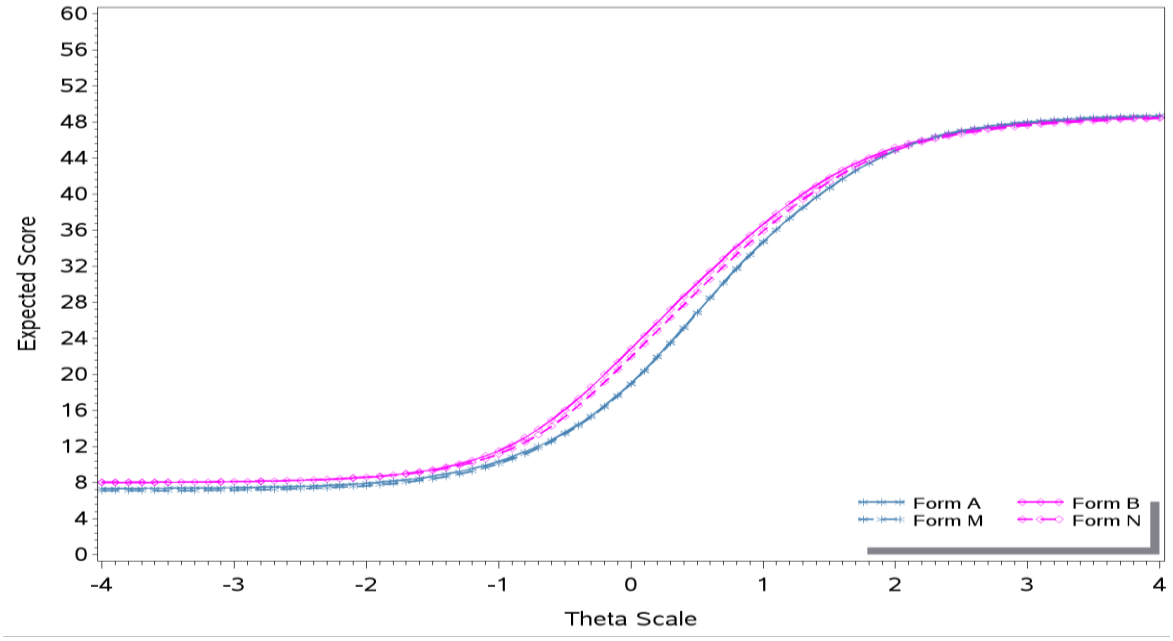


Figure 7.7 Math I TCC Operational Forms A and M, B and N



7.3.4 Measurement Precision-Test Information Function and Conditional Standard Error

In CTT, the concept of reliability is at the center of evaluating the 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 7.8* to *Figure 7.14* show TIF by forms with their associated standard error of measurement. Because NCDPI used TCCs as targets for building parallel forms, the goal was to

select items that minimize the differences between TCCs. As a result, the displayed TIFs for parallel forms are not as closely uniform as the TCCs. The implication is that relative efficiency across forms varies slightly along the ability scale. But overall, the forms are most efficient between theta ranges of -1 to 1.

In terms of standard errors, the figures show they are inversely related to TIF across all forms and are lowest between the theta ranges of -2 and 2. Between the range of -2 and 2 standard errors for alternate forms are uniform and max at about 0.5 around the tails.

Figure 7.8 Math Grade 3 Test Information and Standard Errors for Operational Forms

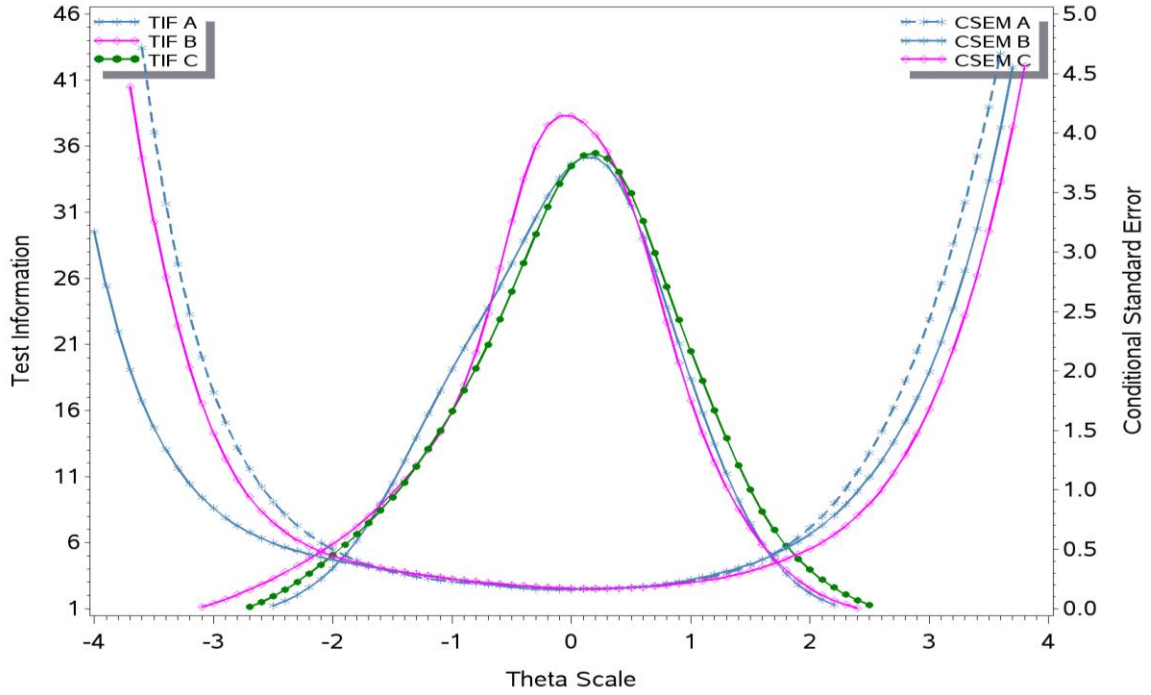


Figure 7.9 Math Grade 4 Test Information and Standard Errors for Operational Forms

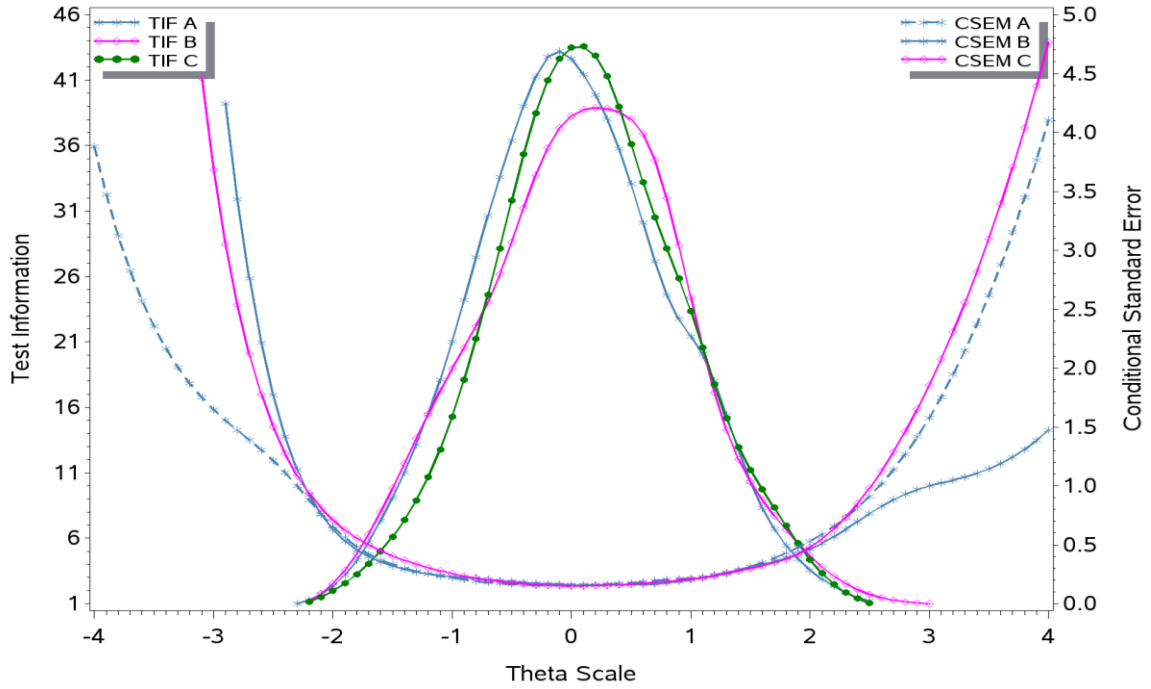


Figure 7.10 Math Grade 5 Test Information and Standard Errors for Operational Forms

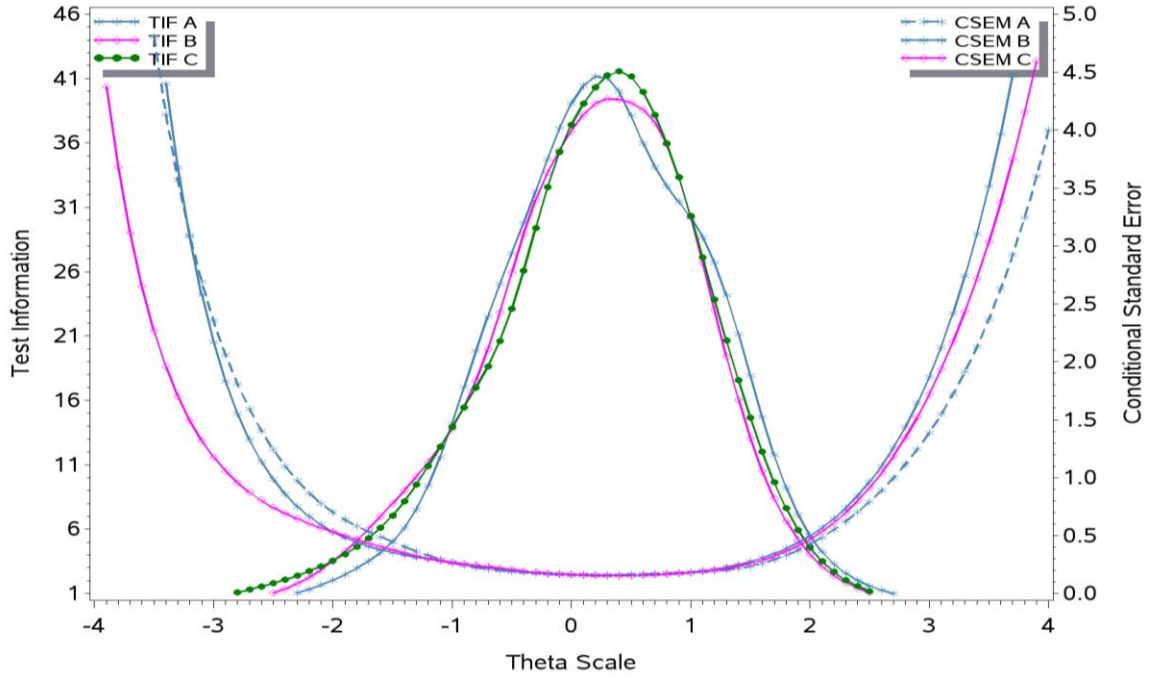


Figure 7.11 Math Grade 6 Test Information and Standard Errors for Operational Forms

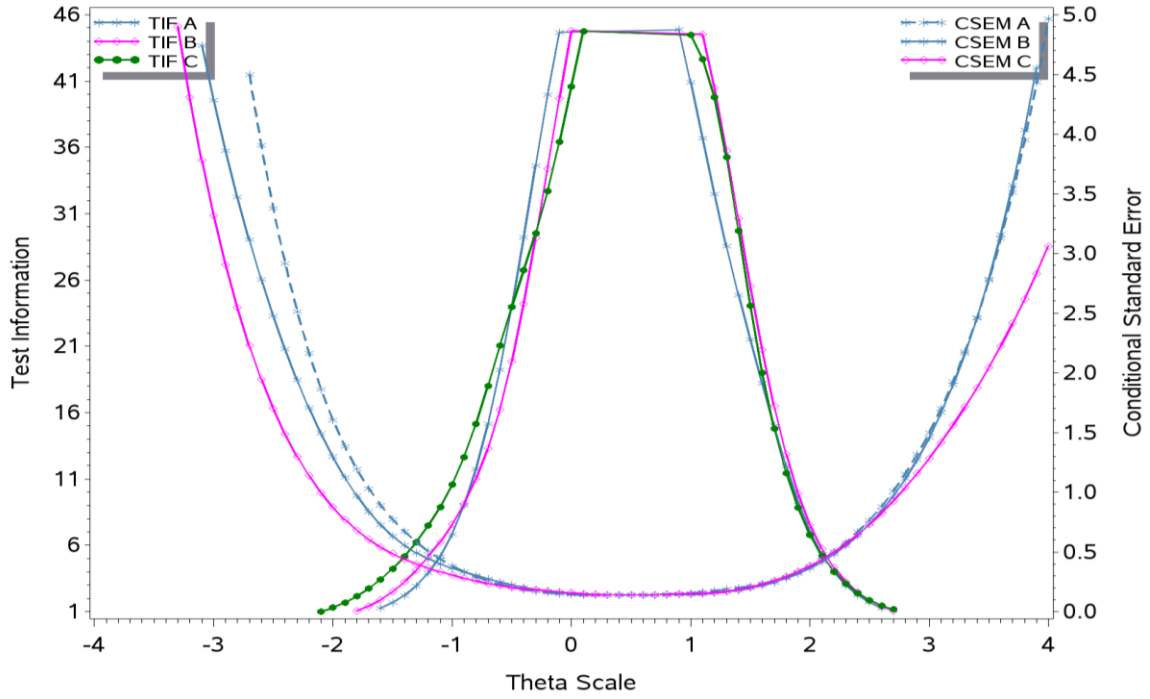


Figure 7.12 Math Grade 7 Test Information and Standard Errors for Operational Forms

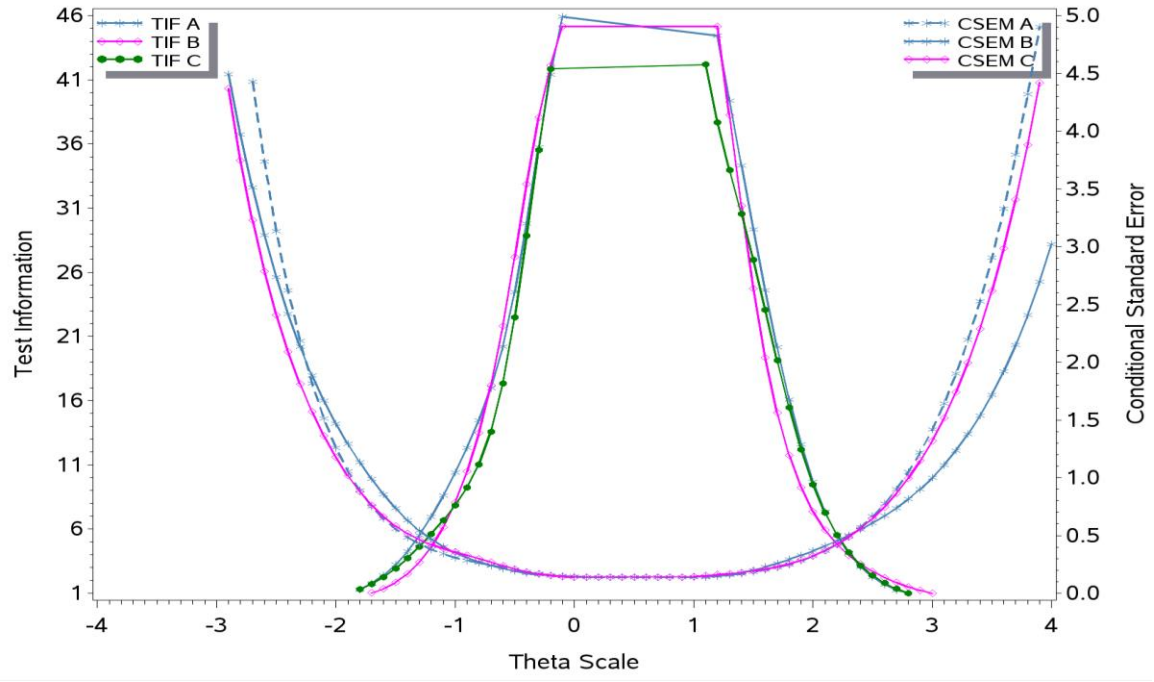


Figure 7.13 Math Grade 8 Test Information and Standard Errors for Operational Forms

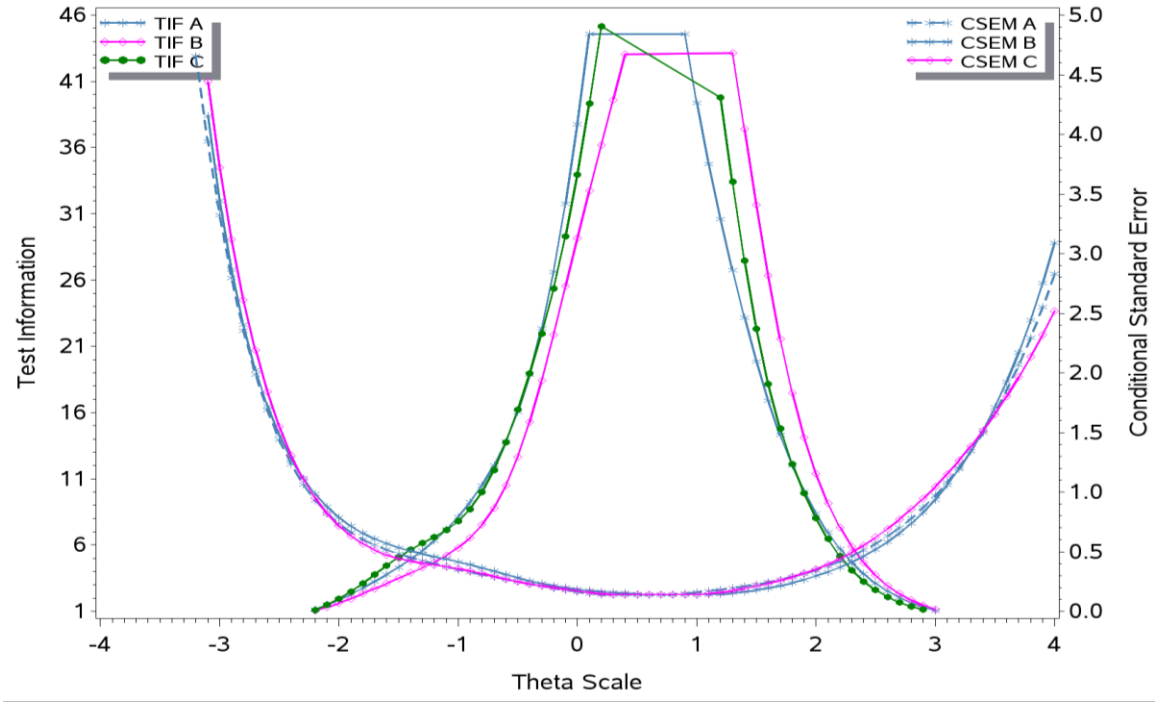
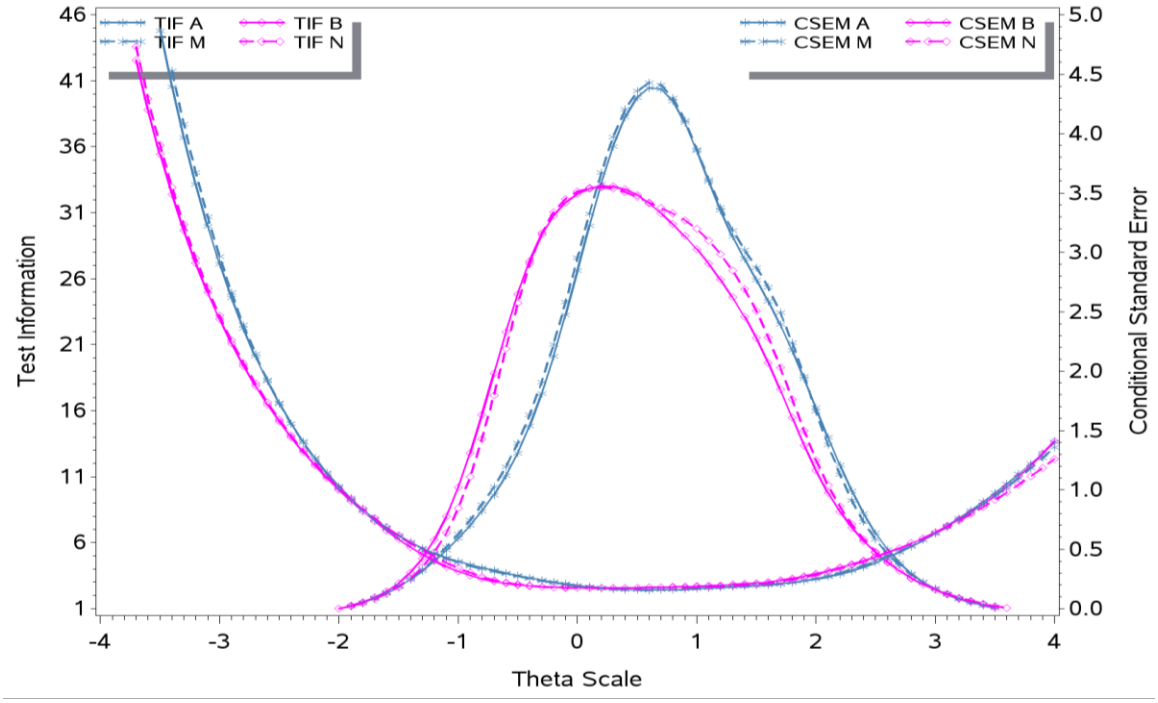


Figure 7.14 Math I Test Information and Standard Errors for Operational Forms



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 will drift from stand-alone 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 baseline. Results from these studies provided evidence in support of the hypothesis of parameter drift and NCDPI's decision to use operational data in conducting standard setting study.

Table 7.2 presents comparison form-level average CTT summary statistics (p-values and point biserials) from the field test and operational administration. The general trend was that the average p-value increased from field test to operational administration ranging from 0.07 to 0.12 across all EOG and EOC forms. This indicated that students' performance on test items on average was higher than estimated from field test data, sometimes significantly. The reliability of

the operational forms ranged from 0.90 to 0.94, which reflects a good error to true score variance ratio for large scale standardized assessment. .

IRT parameters calibrated using field test data and again after the operational administration are presented in *Table 7.3*. A similar trend as noted in the p-values was confirmed by the IRT b-parameter, with the average absolute difference between 0.39 and 0.68 across forms. The ICCs from the post administration calibration on average shifted to the left, indicating that the items were less difficult for students during the operational administration. Complete distributional summaries of the difference in IRT difficulty parameter (b-parameters) between operational and field test administration are shown using boxplots in *Figure 7.15* through *Figure 7.21*. The middle 50% (25^{th} to 75^{th} percentile) of the differences across all forms by grades are 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 better during operational administration.

Table 7.2 CTT Average Descriptive Statistics for Math EOG and EOC 2012–2013

Grade and Form		Number of Items	Field Test CTT Summary		Operational Test CTT Summary		
			P-value	Biserial Correlation	P-value	Biserial Correlation	Reliability (Cronbach's Alpha)
Grade 3	A	44	0.54	0.48	0.64	0.46	0.91
	B	44	0.54	0.47	0.64	0.47	0.92
	C	44	0.54	0.48	0.62	0.45	0.91
Grade 4	A	44	0.50	0.51	0.61	0.48	0.92
	B	44	0.51	0.52	0.58	0.47	0.92
	C	44	0.50	0.51	0.59	0.48	0.92
Grade 5	A	44	0.46	0.49	0.56	0.46	0.92
	B	44	0.46	0.50	0.58	0.47	0.92
	C	44	0.47	0.50	0.58	0.46	0.91
Grade 6	A	50	0.44	0.49	0.52	0.47	0.93
	B	50	0.44	0.50	0.51	0.48	0.93
	C	50	0.44	0.49	0.52	0.47	0.93
Grade 7	A	50	0.43	0.52	0.50	0.49	0.93
	B	50	0.43	0.50	0.50	0.48	0.93
	C	50	0.43	0.51	0.50	0.49	0.94
Grade 8	A	50	0.41	0.47	0.50	0.46	0.92
	B	50	0.41	0.46	0.48	0.45	0.92
	C	50	0.41	0.47	0.50	0.46	0.92
Math I	A*	49	0.39	0.39	0.47	0.43	0.91
	B*	49	0.39	0.39	0.47	0.43	0.91
	M*	49	0.39	0.39	0.46	0.42	0.90
	N*	49	0.39	0.39	0.46	0.41	0.90

*Note: * one item was dropped from the form.*

Table 7.3 IRT Average Descriptive Statistics for Math EOG and EOC 2012–2013

Grade and Form		Number of Items	Average IRT Summary Field Test Administration			Average IRT Summary Operational Administration		
			Slope (a)	Threshold (b)	Asymptote (g)	Slope (a)	Threshold (b)	Asymptote (g)
Grade 3	A	44	1.68	0.29	0.20	1.62	-0.30	0.18
	B	44	1.73	0.29	0.21	1.71	-0.31	0.17
	C	44	1.66	0.30	0.20	1.68	-0.19	0.18
Grade 4	A	44	1.90	0.45	0.19	1.82	-0.13	0.16
	B	44	1.92	0.44	0.19	1.79	0.03	0.16
	C	44	1.90	0.45	0.20	1.81	-0.04	0.18
Grade 5	A	44	1.88	0.66	0.19	1.89	0.14	0.19
	B	44	1.85	0.66	0.19	1.79	0.02	0.17
	C	44	1.94	0.60	0.20	1.86	0.04	0.17
Grade 6	A	50	1.89	0.76	0.18	1.87	0.32	0.18
	B	50	1.86	0.71	0.17	1.89	0.31	0.17
	C	50	1.93	0.74	0.18	1.86	0.28	0.16
Grade 7	A	50	2.04	0.77	0.18	2.06	0.37	0.16
	B	50	1.97	0.79	0.18	2.08	0.40	0.17
	C	50	2.02	0.81	0.18	2.10	0.38	0.16
Grade 8	A	50	1.76	0.89	0.17	1.78	0.33	0.16
	B	50	1.78	0.92	0.18	1.83	0.48	0.17
	C	50	1.83	0.93	0.19	1.85	0.38	0.16
Math I	A*	49	1.58	1.18	0.18	1.54	0.59	0.15
	B*	49	1.62	1.16	0.17	1.56	0.48	0.16
	M*	49	1.58	1.18	0.18	1.56	0.60	0.14
	N*	49	1.62	1.16	0.17	1.57	0.57	0.16

Note: * one item was dropped from the form.

Figure 7.15 Grade 3 Math b-parameter Difference Operational and Field Test

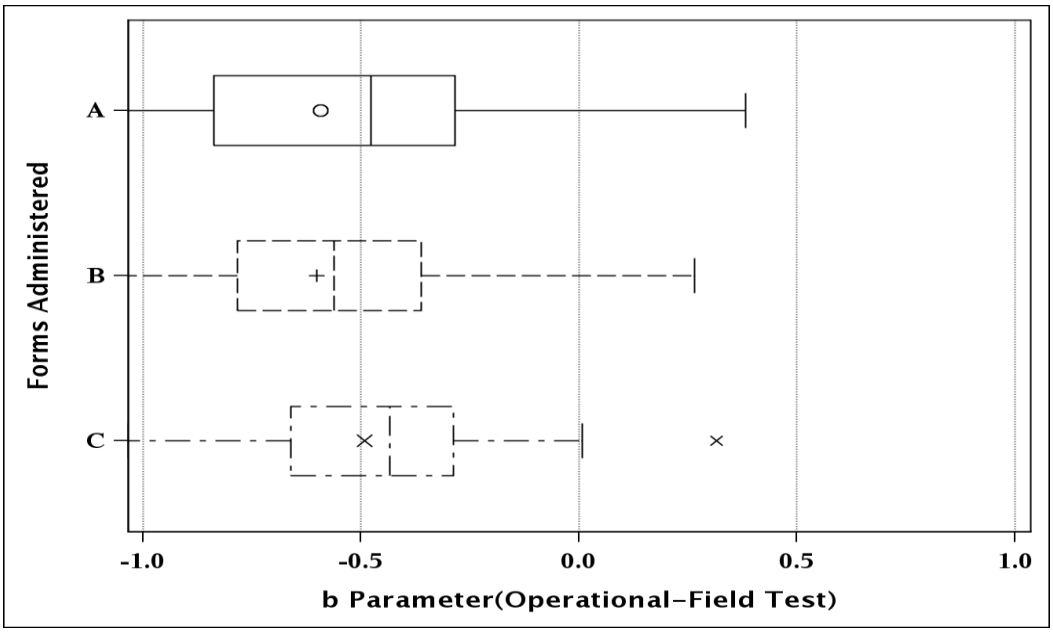


Figure 7.16 Grade 4 Math b-parameter Difference Operational and Field Test

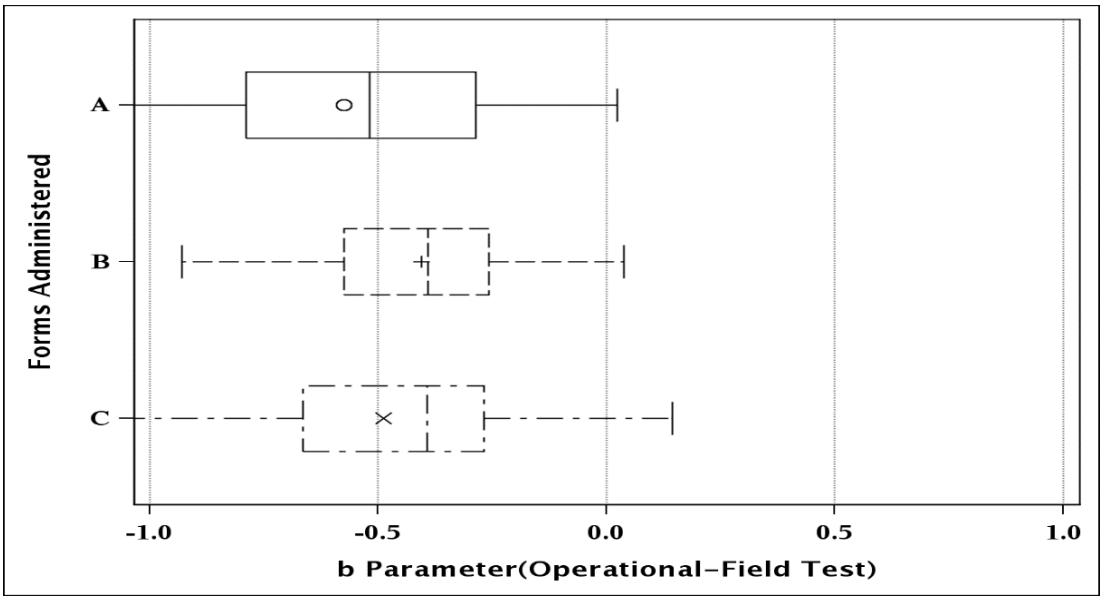


Figure 7.17 Grade 5 Math b-parameter Difference Operational and Field Test

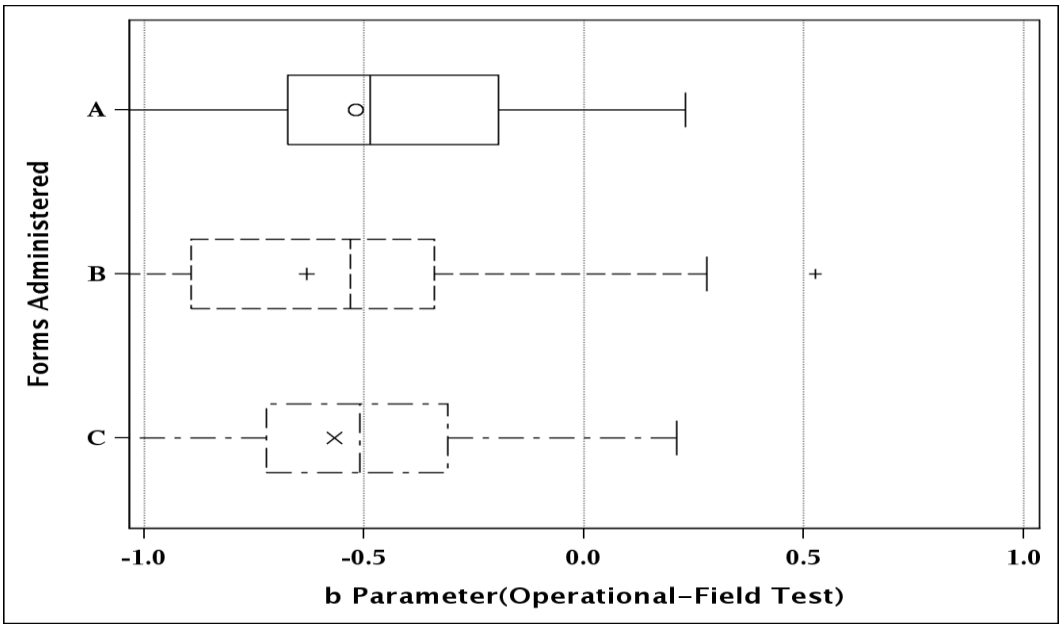


Figure 7.18 Grade 6 Math b-parameter Difference Operational and Field Test

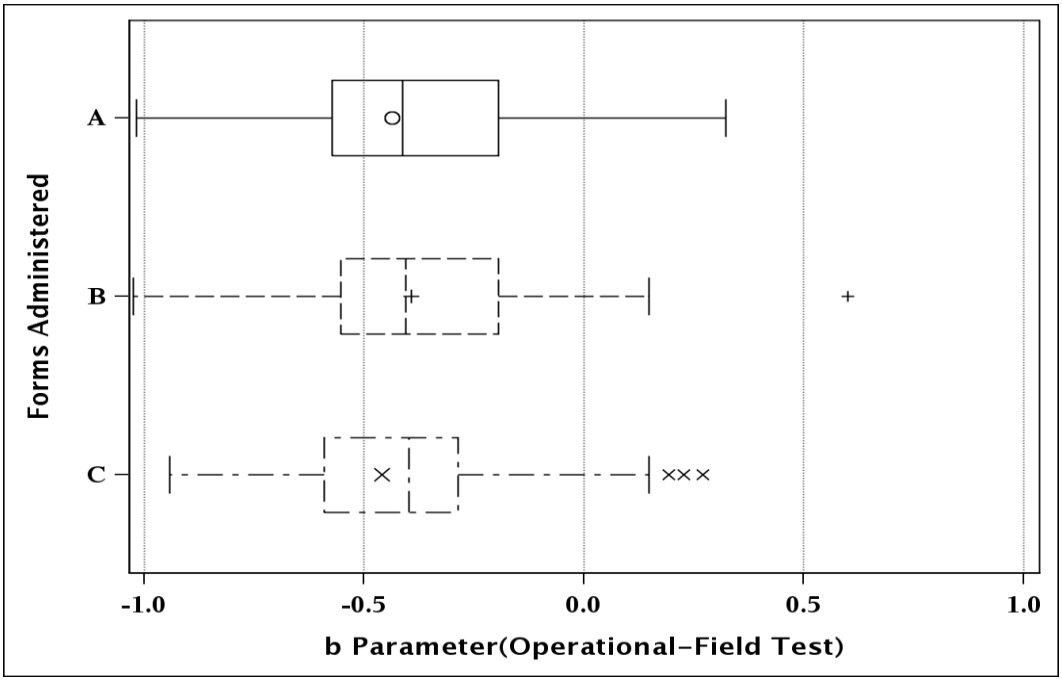


Figure 7.19 Grade 7 Math b-parameter Difference Operational and Field Test

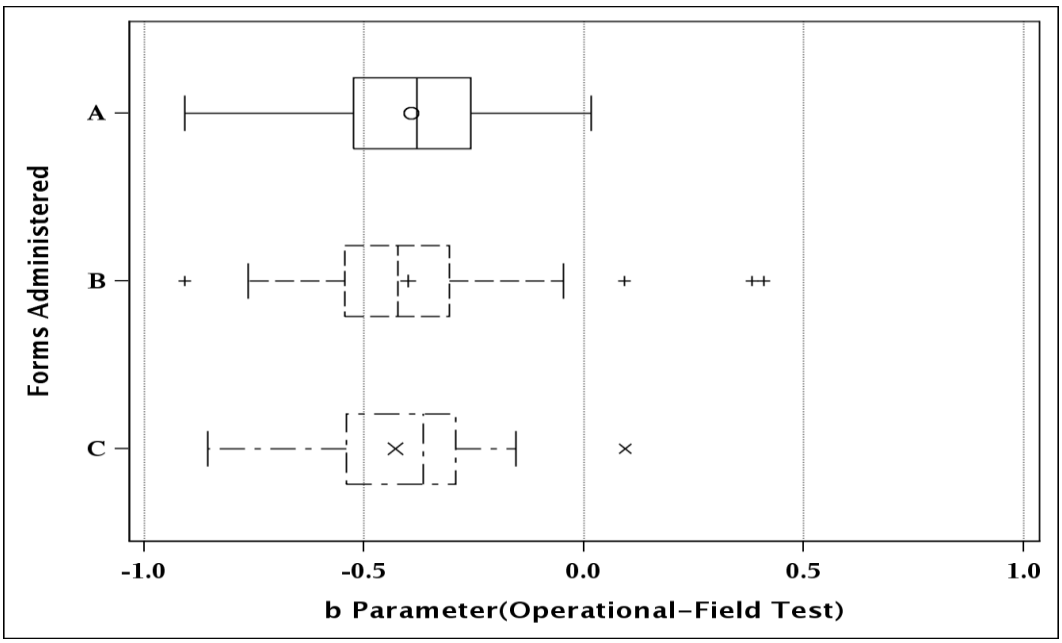


Figure 7.20 Grade 8 Math b-parameter Difference Operational and Field Test

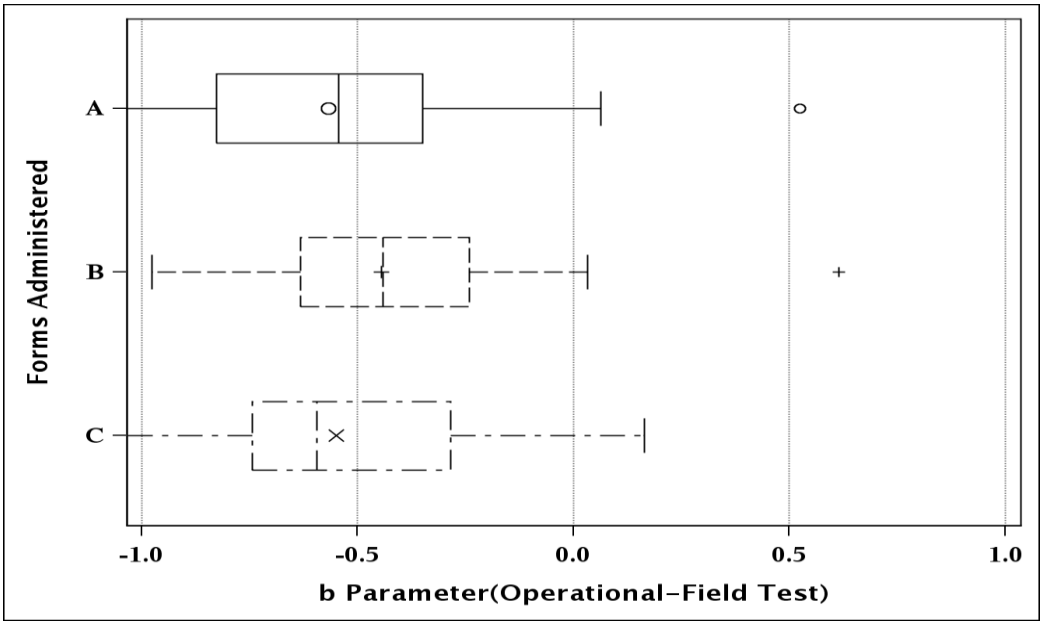
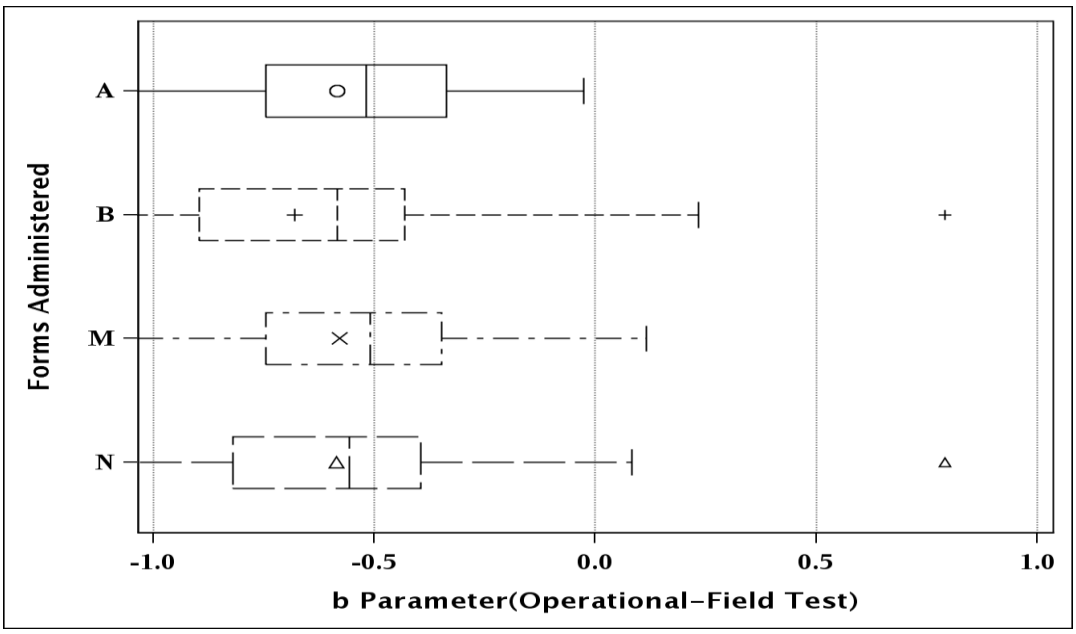


Figure 7.21 Math I b-parameter Difference Operational and Field Test



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.

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

- 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 (1988) classification most of the effect sizes for p-value ranged from 0.40 to 0.74 and b-parameter range from -0.48 to as large as -0.85 indicating on average a medium to large effect from field test to operational parameters.

Table 7.4 Math Effect Size Summary of Operational and Field Test Statistics

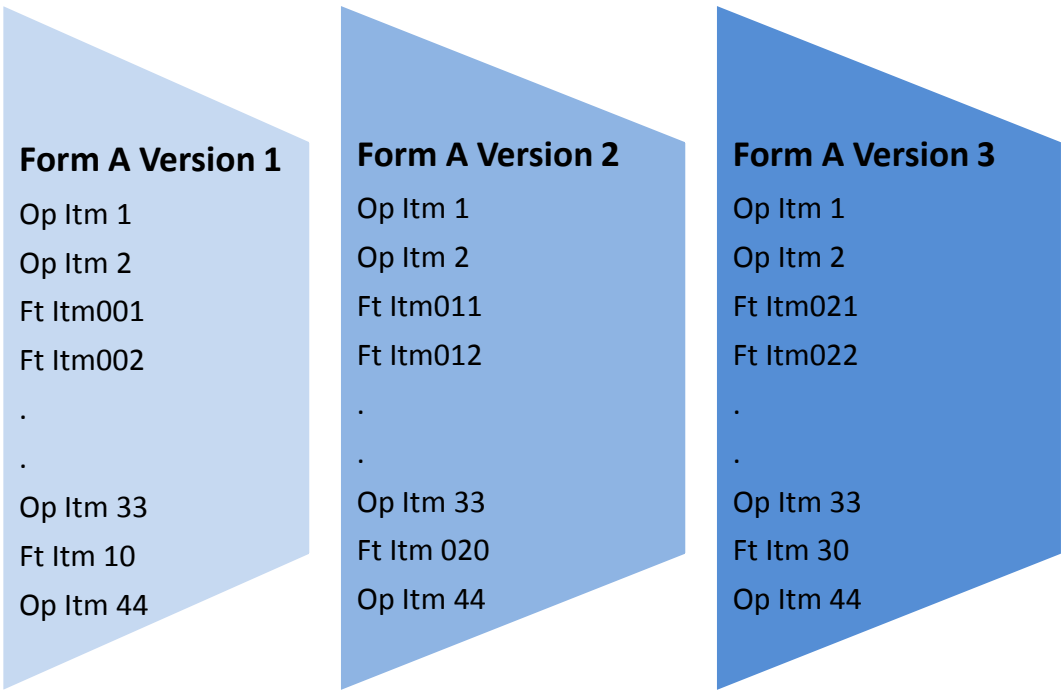
Grade and Form		Operational Items	P-value Standardized Mean Difference	Threshold Standardized Mean Difference
Grade 3	**A**	44	0.60	-0.72
	B	44	0.59	-0.64
	C	44	0.52	-0.57
Grade 4	**A**	44	0.64	-0.70
	B	44	0.45	-0.48
	C	44	0.55	-0.67
Grade 5	**A**	44	0.65	-0.69
	B	44	0.74	-0.85
	C	44	0.59	-0.70
Grade 6	**A**	50	0.60	-0.73
	B	50	0.45	-0.64
	C	50	0.48	-0.60
Grade 7	**A**	50	0.42	-0.56
	B	50	0.46	-0.55
	C	50	0.45	-0.61
Grade 8	**A**	50	0.59	-0.79
	B	50	0.40	-0.58
	C	50	0.51	-0.75
Math I	**A***	49	0.53	-0.66
	B*	49	0.58	-0.73
	M*	49	0.42	-0.65
	N*	49	0.46	-0.60

Note: * one item was dropped from the form.

7.5 Ongoing Form maintenance and Item Development.

As indicated in chapter 1 and 7 of this report NCDPI relies on a continuous item field testing embedding plan for ongoing item development. During operational administration field test items are embedded within operational items and administered to students. For EOG Math a total of 10 field test items are embedded within each operational version of the EOG assessment. For each operational test form, distinct versions are created following a predefined embedding plan See *Figure 7.22* for a schematic example.

Figure 7.22 Item Field Test Embedding Plan



The figure shows field test items (Ft Itm…) embedded within operational items (Op Itm). 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 depends on future form building needs and overall number of students expected to be administered the EOG or EOC. 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 statistics are generalizable to the expected item parameter for the state at the given grade level.

7.6 Development of Forms C and O for EOC Math I

As part of ongoing form rotation NCDPI created two new base forms for Math I using field test items embedded within operational forms A and M during the 2013–14 administration. During operational administration in 2013 – 14 NCDPI had issues with the spiraling procedure in Math I which resulted in only 1 version of forms B and N being administered. *Table 7.5* shows the distribution of number of students by version and form for Math I. As indicated in the table only version 8 in forms B and N were effectively rotated. Whereas, all 15 versions in forms A and M, were effectively rotated as evident by the equal distribution of students who took each version. The implication was that only item statistics from field test items embedded in forms A and M were generalizable to state item parameters and used for subsequent form building.

Table 7.5 EOC Math I Forms by Versions Administered in 2013–14.

Version Number	Base Form							
	A		B		M		N	
	N	Col %	N	Col %	N	Col %	N	Col %
8	1,941	7.01	27,926	99.93	4,283	6.93	3,511	100.00
9	1,890	6.82	11	0.04	3,902	6.31	.	.
10	1,941	7.01	1	0.00	4,718	7.63	.	.
11	1,900	6.86	.	.	3,410	5.52	.	.
12	1,833	6.62	.	.	4,361	7.05	.	.
13	1,862	6.72	1	0.00	3,624	5.86	.	.
14	1,881	6.79	1	0.00	3,430	5.55	.	.
15	1,868	6.74	.	.	4,397	7.11	.	.
16	1,857	6.70	1	0.00	3,523	5.70	.	.
17	1,834	6.62	.	.	4,460	7.21	.	.
18	1,884	6.80	1	0.00	3,609	5.84	.	.
19	1,778	6.42	1	0.00	4,593	7.43	.	.
20	1,824	6.58	.	.	4,875	7.89	.	.
21	1,722	6.22	2	0.01	4,336	7.01	.	.
22	1,688	6.09	1	0.00	4,305	6.96	.	.
All	27,703	100.00	27,946	100.00	61,826	100.00	3,511	100.00

Note: Col = Column

Table 7.6 Field Test Item Pool for EOC Math I in 2013–14

EOC Test Mode	Number of Versions	Items Per Version	Total Items Field Tested
Paper	15	11	165
Computer	15	11	165

The classical statistics (p-values and biserial correlations) and the descriptive statistics for IRT parameters (a, b, and g) are presented in *Table 7.7* and Table 7.8 for the field tested items

used to build Forms C and O. NCDPI item quality criteria (see Section 4.5.1) were used to determine if items met the technical standards to be considered for operational use. Any exception to the criteria is done only under exceptional cases and with thorough vetting from the content experts and psychometricians.

Table 7.7 CTT Field Test Item Pool Descriptive Statistics for EOC Math I 2013 - 14

EOC Math I Test Mode	Number of Items		P-Value				Biserial Correlation			
	MC	GR	Mean	SD	Min	Max	Mean	SD	Min	Max
Paper	135	30	0.35	0.17	0.00	0.82	0.42	0.16	0.01	0.78
Computer	135	30	0.33	0.17	0.00	0.82	0.39	0.17	0.01	0.75

Note: MC = Multiple-choice; GR = Gridded-response

Table 7.8 *IRT Field Test Item Pool Descriptive Statistics for EOC Math I 2013 - 14*

EOC Math I Test Mode	Number of Items		Slope(a)				Threshold(b)				Asymptote(g)			
	MC	GR	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Paper	135	30	1.97	0.65	0.45	3.85	0.90	0.77	-1.49	3.16	0.17	0.11	0.00	0.47
Computer	135	30	2.01	0.66	0.49	3.85	0.88	0.75	-1.49	2.29	0.17	0.11	0.00	0.47

Note: MC = Multiple-choice; GR = Gridded-response

The number of items classified into the “Delete,” “Reserve,” and “Keep” categories from EOC Math I item pool are shown in *Table 7.9* A total of 184 (56%) of items field tested in forms A and M met the psychometric “Keep” criteria. These item pool provided sufficient items to build the new forms C and O based on the same content domain blueprint adopted for Math I and used to build parallel forms A, B, M and N.

Table 7.9 Psychometric Status for Item Pool 2013 -14

EOC Math I Test Mode	Psychometric Evaluation Summary					
	Keep		Reserve		DELETE	
	N	Row %	N	Row %	N	Row %
Paper	95	58	27	16	43	26
Computer	89	54	32	19	44	27
Total	184	56	59	18	87	26

The TCCs of forms C and O plotted together with forms A, B, M, and N are shown in *Figure 7.23*, and TIFs with their corresponding CSEs are shown in *Figure 7.24*. The TCCs for A/M and C/O are closely overlapped, indicating that the new forms are psychometrically similar with operational forms A/M in terms of form difficulty across the ability range. The TIFs and CSEs also indicate that the new forms (C and O) are most informative between the ability ranges of 0 and 2. Between ability ranges of -2.5 and 2.5 CSEs are similar for all parallel forms. These IRT results confirm the new forms C and O are parallel with forms A, B, M, and N. All forms share the same blueprint.

Figure 7.23 TCCs for Math I Operational Forms A, B, C, M, N and O

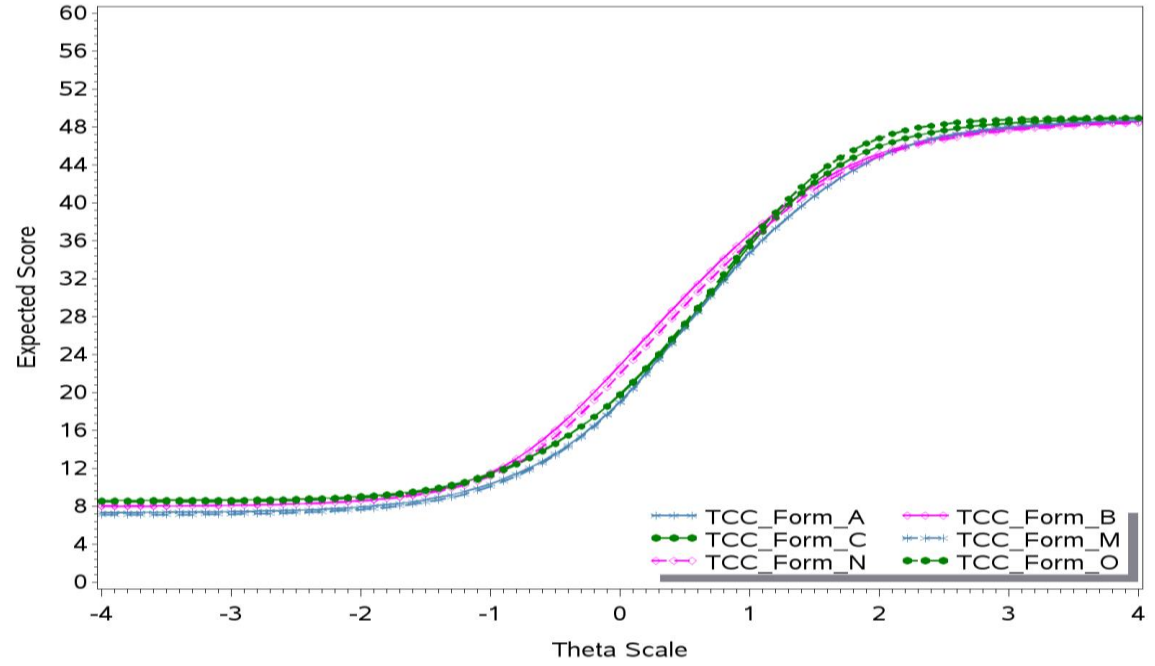
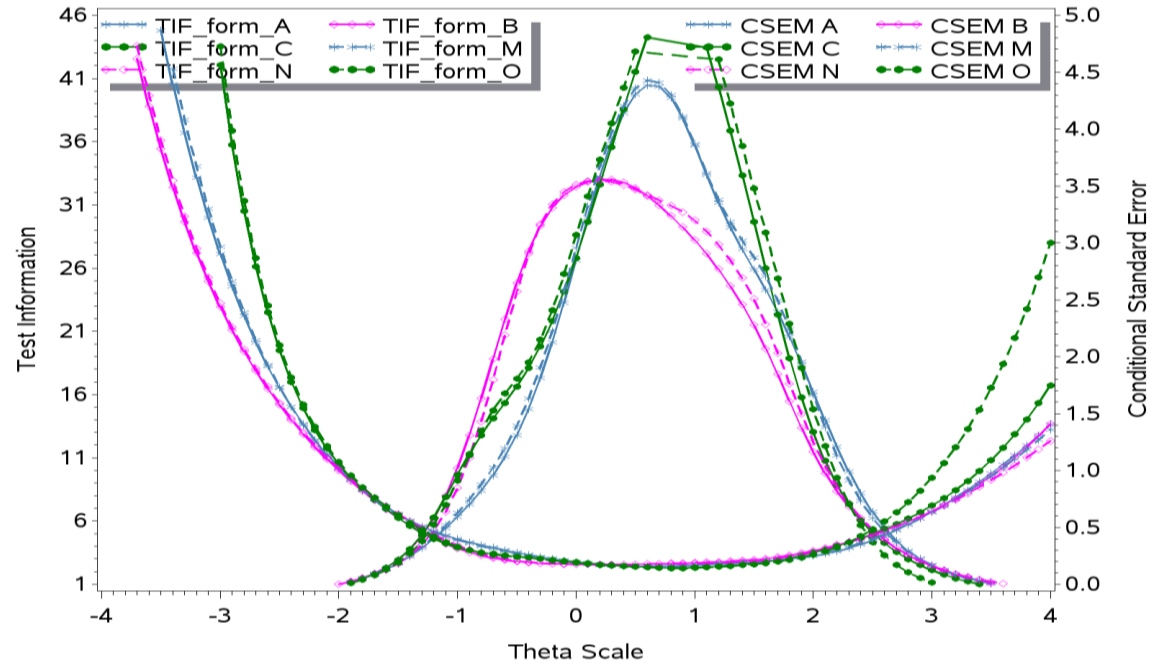


Figure 7.24 TIFs and CSEs for Math I Operational Forms A, B, C, M, N, and O



Chapter 8 **Standard Setting**

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 NC*SCS* 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, NCDPI contracted with Pearson Inc. to conduct a standard setting workshop to recommend cut scores and achievement levels for the newly developed Math EOG and EOC assessments.

8.1 Standard Setting Overview

Standard 5.21 (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" (p. 107). 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.

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 NCSCS and the development of The READY accountability assessment system to measure students' college- and career-readiness. In July of 2013 after the first operational administration of EOG and EOC, NCDPI contracted with Pearson Inc. to conduct a full standard setting workshop with the main goal of recommending cut scores and achievement levels for the newly developed Math EOG and EOC assessments.

Three panels (grades 3–5, grades 6–8, and Math I) with a total of 57 (20 for grade 3–5, 16 for grades 6–8, and 21 for Math I) North Carolina Math educators convened in Chapel Hill, North Carolina, between July 22 and July 26, to make cut score recommendations for the Math EOG and EOC assessments. The item mapping

procedure (Lewis, Green, Mitzel, Baum, & Patz, 1998; Mitzel, Lewis, Patz, & Green, 2001) based on ordered item booklets prepared by 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 staff. The full report of the standard setting can be found in 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 present to the NCSBE for adoption. An abbreviated version of the final standard setting study prepared by Pearson[j] for the North Carolina Department of Public Instruction 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 8.6***. 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 57 panelist ranged from less than 5 years to more than 21 years of experience. The table shows that a very diverse group of educators participated in 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
Math 3–5	20	1	4	8	2	4	1
Math 6–8	16	2	3	4	5	2	0
Math I	21	4	3	5	2	7	0

Note: NR = no response.

The panelists' professional backgrounds are summarized in *Table 8.2* and *Table 8.3*. Panelists in the Math 3–5 and 6–8 groups made cut score recommendations for three grade levels of EOG Math, and the 21 panelist in the Math I group made cut score

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

recommendations for EOC Math I. From these tables, teachers reported as teaching in lower, middle, or upper grades are reported in the context of their committees. For example, a lower-grade panelist in the Math 3–5 panel teaches Grade 3 Math, while a lower-grade panelist in the Math 6–8 panel teaches Grade 6 Math. 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 Math 3–5 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 these tables, all grade levels were represented by panels, plus a variety of professional backgrounds was also represented on these panels.

Table 8.2 Panelist Professional Background: Three-Grade Panels

Panel	LOW	MID	UP	MUL	OFF	SED	SPE	COA	GNS	OTH
Math 3–5	3	6	5	2	1	0	2	1	0	0
Math 6–8	7	3	3	1	0	1	1	0	0	0

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.

Table 8.3 Panelist Professional Background: Single-Grade Panels

Panel	ON	OFF	SED	SPE	COA	HED	OTH	RET	NR
Math 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.

In addition to reporting their own demographic characteristics (*Table 8.4*), panelists were asked to report their district geographic location within the state (*Table 8.5*) as well as district size and community setting (*Table 8.6*). As demonstrated by the information provided in these tables, panelists making up the standard setting committees showed representative diversity among geographic regions, district sizes, and community settings across North Carolina.

Table 8.4 Panelist Gender and Ethnicity

Panel	Gender			Ethnicity						
	F	M	NR	AA	AS	HI	NA	WH	MU	NR
Math 3–5	18	2	0	7	0	0	0	12	0	1
Math 6–8	11	5	0	3	0	1	0	12	0	0
Math 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, NR = no response.

Table 8.5 Panelist Geographic Region

Panel	C	NC	NE	NW	SC	SE	SW	W	MU	NR
Math 3–5	4	1	0	1	4	4	5	1	0	0
Math 6–8	1	2	1	1	2	3	4	2	0	0
Math 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.

Table 8.6 Panelist District Characteristics

	District Size				Community Setting			
Panel	NR	SM	MD	LG	NR	RU	SU	UR
Math 3–5	0	4	6	10	1	10	4	5
Math 6–8	0	4	5	7	0	9	4	3
Math I	1	7	6	7	0	6	8	7

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 21 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 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.

8.1.3 Method and Procedure

A total of nine panels set standards for 17 grades and subjects. Panelists on the three-grade committees recommended standards for three adjacent grade levels within Math (i.e., grades 3–5 or 6–8). For the single-grade committees, panelists recommended standards for a single grade/subject. Although all nine panels used a similar methodology for panelists to render their judgments, the scope of activities varied across the two panel types. The three-grade panels convened between July 22 through 26, 2013, while the single-grade panels convened between July 24 and 25, 2013.

8.1.4 Table Leader Training

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.

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 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. Breakout session facilitators provided panelist 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 were turned over to NCDPI for review and future revisions, as deemed necessary.

8.1.7 Standard Setting

"Just Barely" Level Descriptors

Following ALD writing activities, panelists performed tasks to set standards for their assigned subject area and grade(s). 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 descriptions 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 test with reading selections only), and linked content standard. After completing the OIB review, panelists were given an opportunity to share their thoughts 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 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 the three-grade panels, standard setting activities began at the lower grade level (i.e., grade 3 for the panels assigned to grade 3–5, grade 6 for panels assigned to grades 6–8). 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. 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. The breakout session facilitator collected all of the committee's 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 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 those 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's, 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 their thoughts and perspectives with the other panelists at the table. Panelists compared bookmarks and discussed the differences between them. 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 8.7.

Table 8.7 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 8.8

Table 8.8 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**
	Minimu	**14**	**42**	**60**
	Maximu	**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 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.

As shown in *Table 8.9*, cut scores shared with panelists were translated into page numbers in the ordered item book to help facilitate comparisons between the external data and their own cut score judgments. 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 8.9 Linked Page Cuts from the Teacher Survey and ACT Explore®

Assessment	Level 2	Level 3	Level 4	Explore®
Math 3	6	22	66	48
Math 4	1	14	60	44
Math 5	1	8	56	38
Math 6	1	3	48	29
Math 7	1	3	46	30
Math 8	1	3	34	28
Math I	1	1	38	*

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

Following discussion of Round 1 cut scores and the provided 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 both 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 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 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 by 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. For the three-grade committees, the breakout session facilitator guided panelists through the same process for the middle and upper grades, starting with the ordered item book review and then proceeding directly to Round 1. 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 3–8 Math) 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; but the lead facilitator did provide 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:

$$SE_{Median} = \frac{\sigma}{\sqrt{N}} \quad (8\text{-}1)$$

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 full report Appendix M) 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 full report Appendix N).

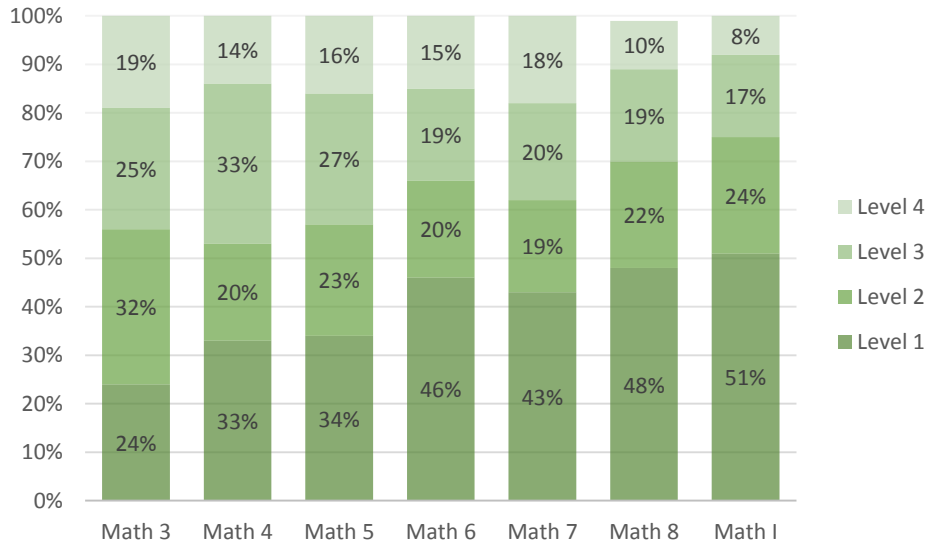
8.3 Results

The standard setting panels' final recommended cut scores, obtained prior to the vertical articulation session, are presented in Table 8.10. 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 figure 8.1 and figure 8.2 display impact data for the Mathematics EOG and EOC assessments respectively based upon these cut score recommendations. Tables and figures showing individual panelists' page cuts across rounds are provided in the full report Appendix I.

Table 8.10 Pre-Vertical Articulation Page Cuts

Assessment	Level 2	Level 3	Level 4
Math 3	16	41	69
Math 4	15	34	70
Math 5	9	33	65
Math 6	10	32	67
Math 7	9	28	59
Math 8	10	30	70
Math I	9	29	60

Figure 8.1 Pre-Vertical Articulation Impact Data

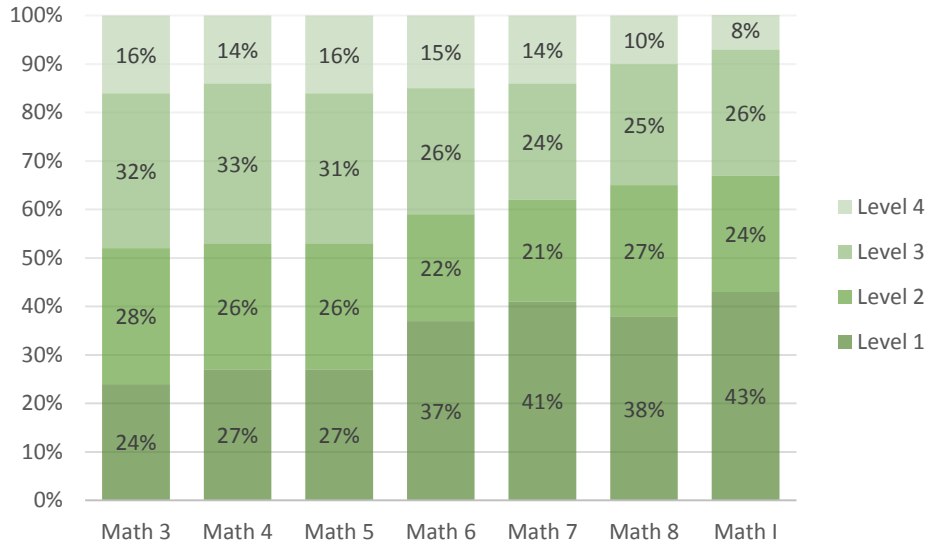


Cut scores obtained following the vertical articulation session are shown in Table 8.11 and impact data associated with these recommended cut scores are displayed in the subsequent figures.

Table 8.11 Post-Vertical Articulation Page Cuts

Assessment	Level 2	Level 3	Level 4
Math 3	16	38	73
Math 4	10	34	70
Math 5	7	30	65
Math 6	4	24	67
Math 7	6	28	65
Math 8	5	25	70
Math I	16	38	73

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 8.12.

The scale scores cut represent the lower cuts for the adjacent achievement level. For example, the Math 3 "Level 2" cut of 443 is interpreted as students with a scale score of 442 or lower are placed in "Achievement Level 1" and student who score between 443 and 450 are considered to be performing at "Achievement Level 2".

Table 8.12 Scale Scores Cuts Based on Four Achievement Levels 2012–2013.

Assessment	Level 2	Level 3	Level 4
Math 3	443	451	460
Math 4	444	451	460
Math 5	444	451	460
Math 6	447	453	461
Math 7	447	453	461
Math 8	447	454	463
Math I	247	253	264

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, calling 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 (NCSBE) adopted College- and Career-Readiness Academic Achievement Standards and Academic Achievement descriptors for the End-of-Grade (EOG) and End-of-Course (EOC) assessments. After considering much input on the importance of having more definitive discrimination for student achievement in the reported levels, the NCSBE 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 will be reported based on five achievement levels as described in table 8.13

.

Table 8.13 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

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

The new Achievement Level 3 "Sufficient Command" identifies students who met on-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. For EOG and EOC Math the new Level 3 minimum scale score was created subtracting one standard error of measurement (SEM) from the original Level 3 scale score. The one standard error adjustment was also done to the original Level 1 "Limited Command" and Level 2 "Partial Command" cuts because the gap in terms of scale scores between level 3 and 2 after the adjustment became very small. Thus new Levels 1/2 and 2/3 cuts were defined whereas old Levels 2/3 and 3/4 cuts became Levels 3/4 and 4/5 cuts respectively (see Table 10.2).

Chapter 9 Test Results and Reports

This chapter is divided into two main sections and presents test-level summary statistics for Math EOG and EOC based on reported scale scores and achievement levels from 2012–13 through 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 NCDPI, which are available to LEA to share assessments results with stakeholders.

9.1 Scale Score Summary

9.1.1 Scale score population

The scale scores distribution from the first operational administration of the EOG and EOC in 2012–13 are displayed in the bar charts in *Figure 9.1* through *Figure 9.7*. Scale scores across all EOG grade levels consistently have means around 450 and standard deviations around 9.5. For EOC Math I, score distribution is skewed slightly to the right, with a mean at 249.4 and a standard deviation at 9.6.

Figure 9.1 Math Grade 3 Scale Score Distribution 2012–2013

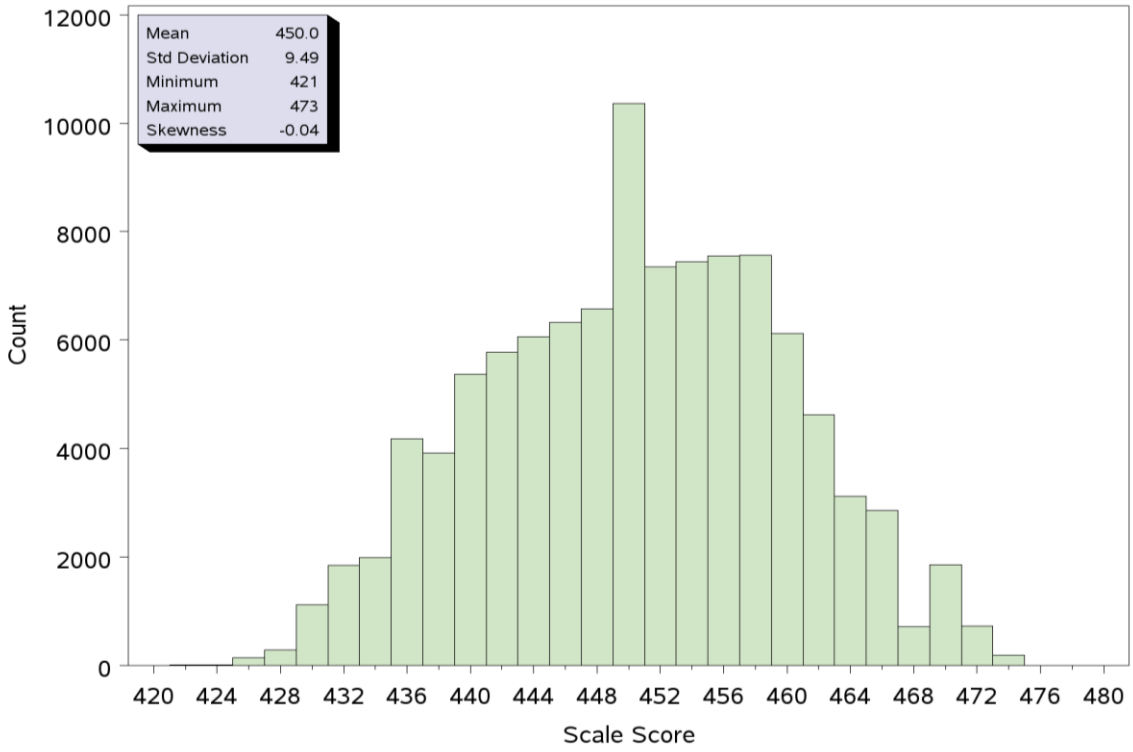


Figure 9.2 Math Grade 4 Scale Score Distribution 2012–2013

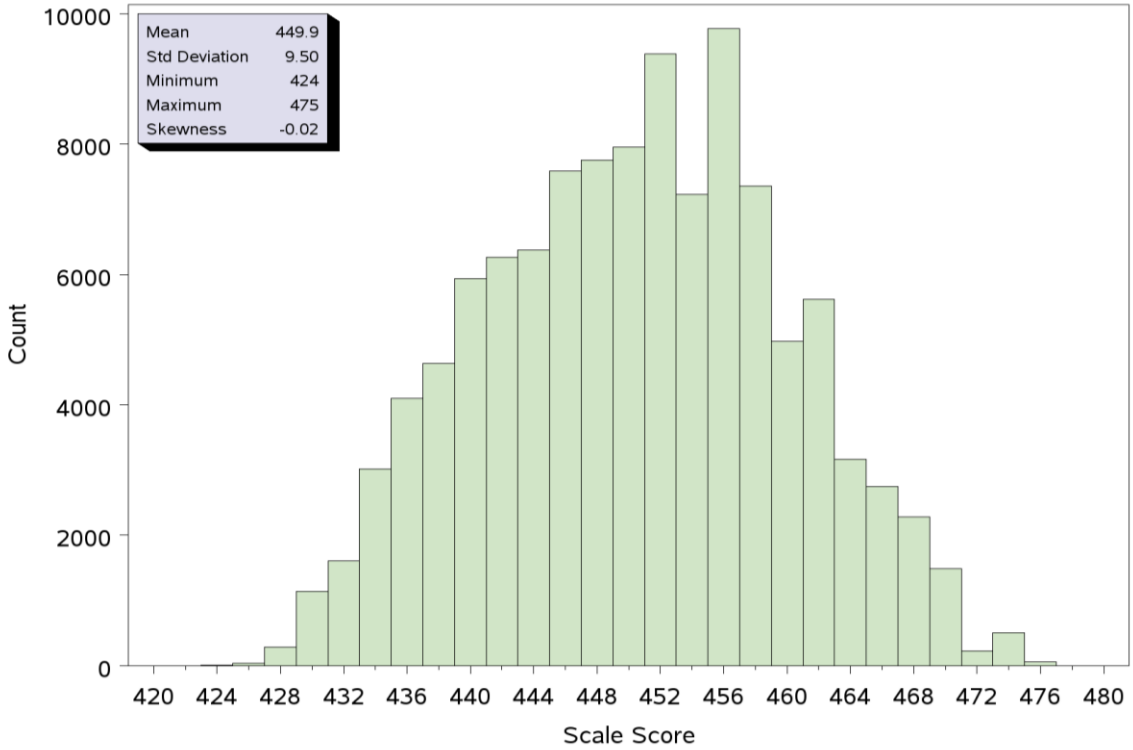


Figure 9.3 Math Grade 5 Scale Score Distribution 2012–2013

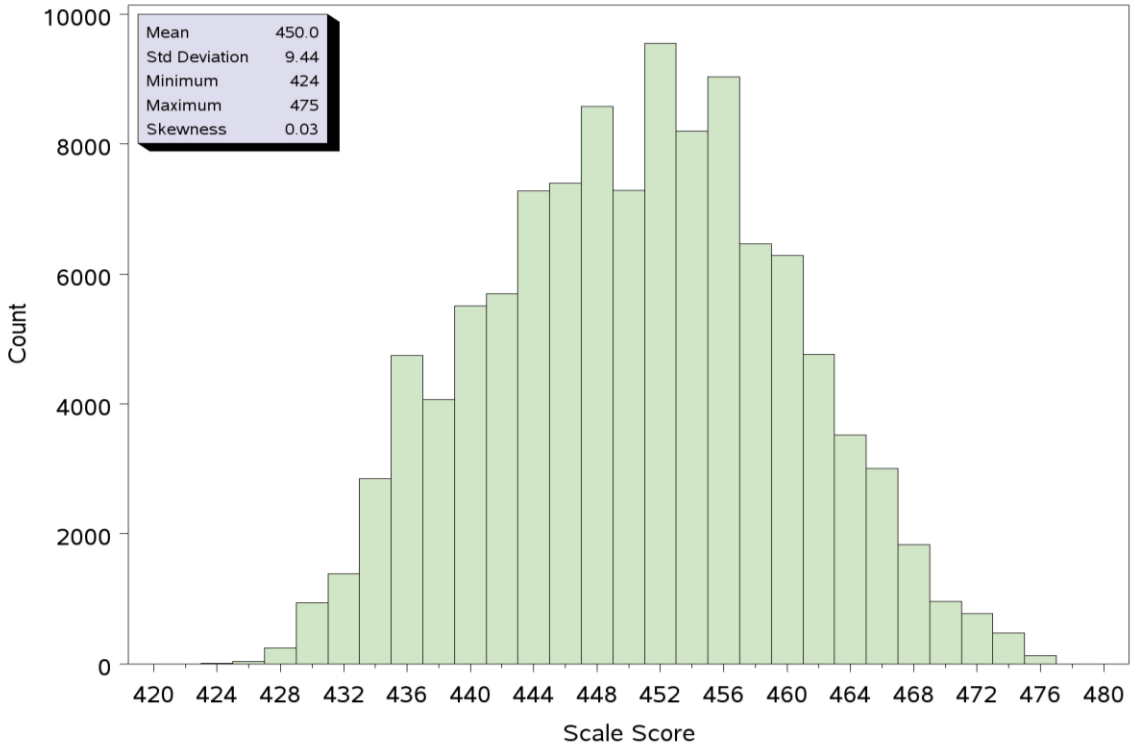


Figure 9.4 Math Grade 6 Scale Score Distribution 2012–2013

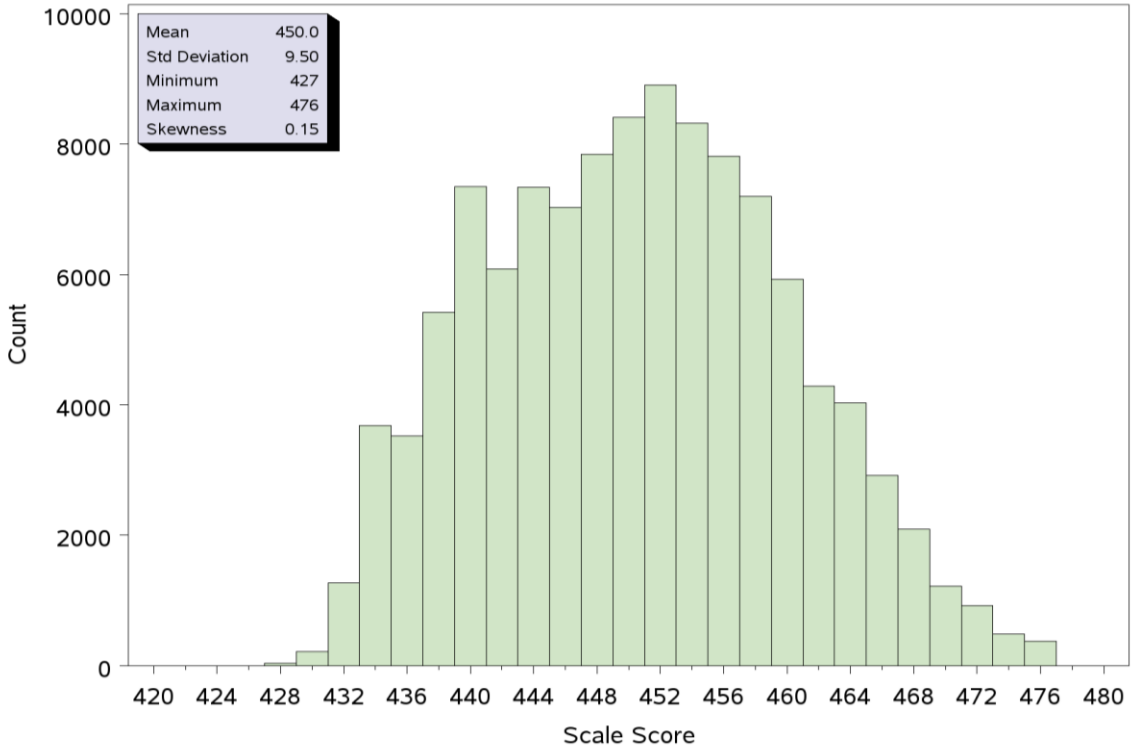


Figure 9.5 Math Grade 7 Scale Score Distribution 2012–2013

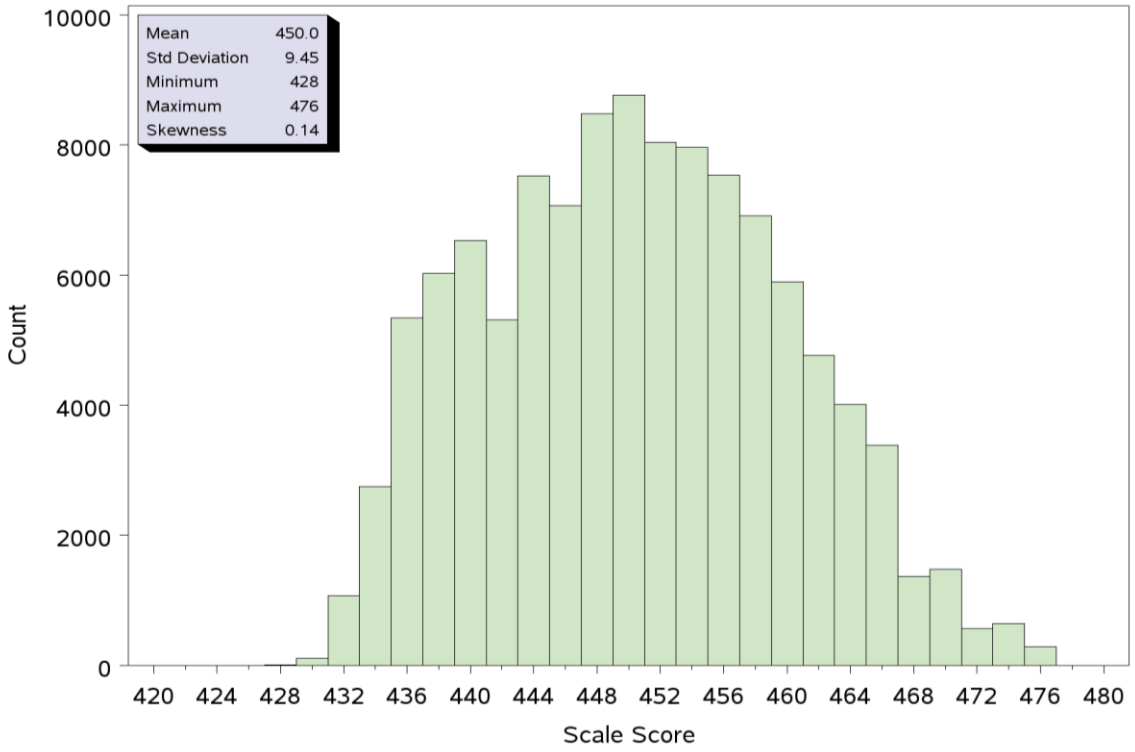


Figure 9.6 Math Grade 8 Scale Score Distribution 2012–2013

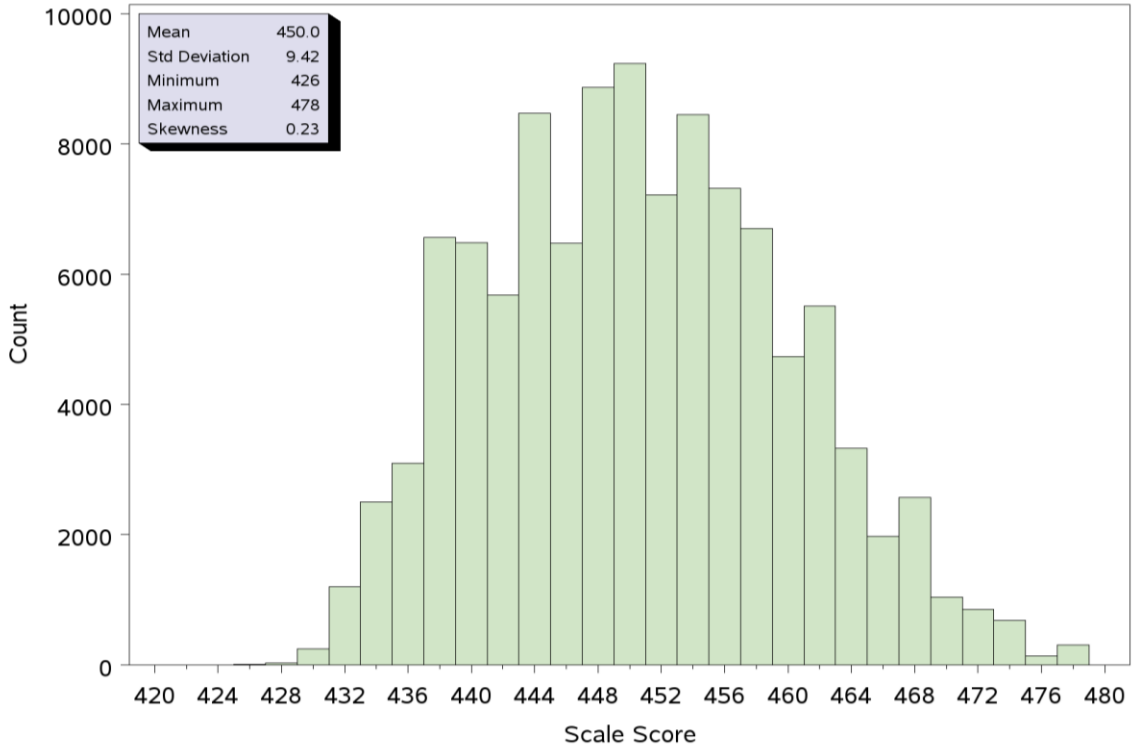
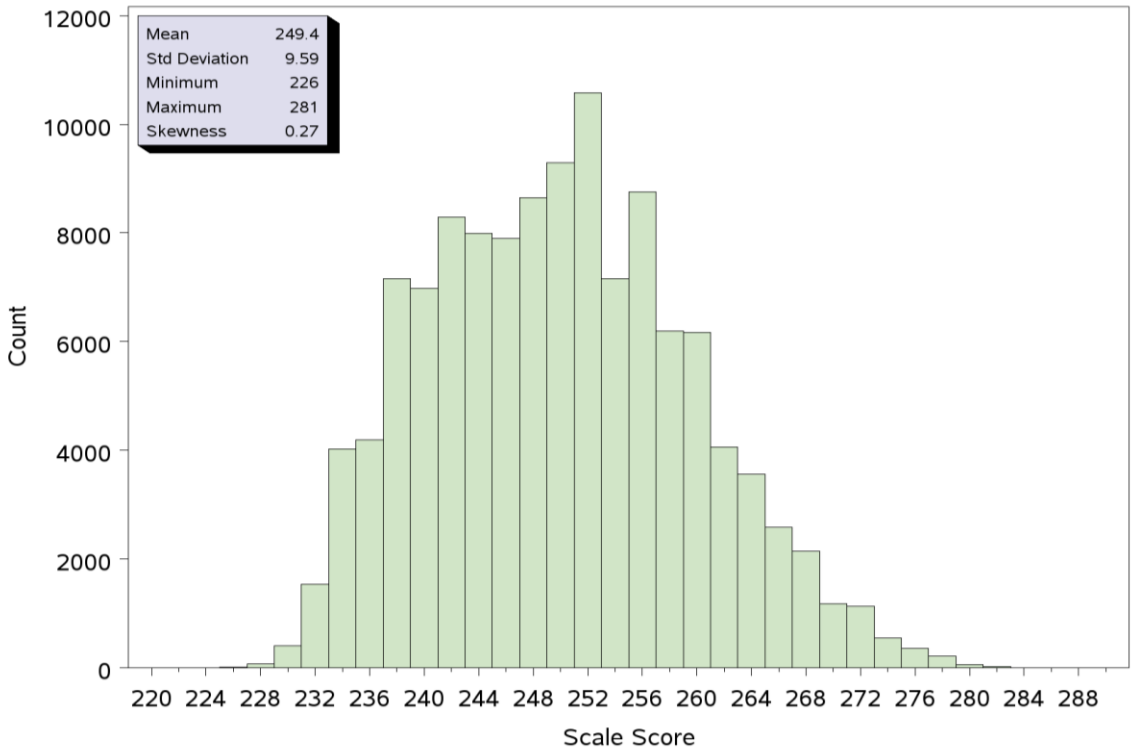


Figure 9.7 Math I Scale Score Distribution 2012–2013



A longitudinal summary of EOG and EOC scale scores for the past three administrations (2012–13, 2013–14 and 2014–15) is presented in *Table 9.1*. The number of students taking EOG and EOC assessments across the state has been on a small but steady increase across the years in general, with some exceptions. Descriptive summary evidence from *Table 9.1* indicates average scale scores have been consistent across the past three years (around 450). In general, average scales scores across all assessment for the past three years have either stayed flat or show slight fluctuation from the base year. The effect of the difference across years is very small and can be explained mainly by sampling variability across years. In the 2014–15 administration cycle, NCDPI also administered EOG grade 7 on computers. Overall variability summarized using the standard deviation (SD) also indicates a flat to slightly upward trend in overall variability from 2012–13 to 2014–15 but only of a small magnitude.

Table 9.1 Descriptive Statistics of Scale Scores by Grade across Administrations, Population

Type	2012-13			2013-14			2014-15		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
EOG 3	103,594	450.0	9.5	112,017	450.2	9.6	116,404	450.0	9.7
EOG 4	110,987	449.9	9.5	103,977	449.6	9.9	113,968	449.7	10.0
EOG 5	110,599	450.0	9.4	111,718	450.2	9.7	106,611	450.3	10.1
EOG 6	112,257	450.0	9.5	111,470	449.9	9.7	114,473	449.8	10.0
EOG 7	111,333	450.0	9.4	113,416	449.9	9.7	114,662	449.7	10.1
EOG 8	109,199	450.1	9.4	112,243	450.1	9.6	116,739	449.7	10.2
EOC Math I	116,988	249.7	9.5	116,462	250.7	9.5	118,802	250.2	10.0

9.1.2 Scale Score by Gender

Scale score summaries by gender for EOG and EOC across three administration cycles show similar trend observed in the population distribution. Across all grades, the distribution between males and females is almost even with male students having a slight majority. In terms of performance, females on average score 0.1 to 1.1 scale points higher than males, except in Grade 4 where males on average slightly out performed females across all three years. Scale score variances were very similar in both gender groups, with variability among scores for males slightly larger than for females; and the trend shows slightly increasing score variability recorded across years.

Table 9.2 Scale Scores by Grade and Gender, Population

	Gender	2012-13			2013-14			2014-15		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
EOG 3	Female	51,003	450.0	9.3	55,329	450.2	9.4	56,938	450.2	9.4
	Male	52,591	450.0	9.6	56,688	450.1	9.8	59,466	449.8	10.0
EOG 4	Female	54,829	449.7	9.3	50,995	449.5	9.7	55,849	449.6	9.9
	Male	56,158	450.1	9.6	52,982	449.8	10.0	58,119	449.7	10.2
EOG 5	Female	54,693	450.2	9.1	55,065	450.4	9.4	51,936	450.6	9.7
	Male	55,906	449.9	9.7	56,653	450.1	10.0	54,675	450.0	10.4
EOG 6	Female	55,440	450.3	9.4	54,754	450.2	9.6	55,841	450.2	9.9
	Male	56,817	449.8	9.6	56,716	449.6	9.9	58,632	449.4	10.2
EOG 7	Female	55,105	450.3	9.3	55,884	450.3	9.5	55,933	450.2	9.9
	Male	56,228	449.8	9.5	57,532	449.6	9.8	58,729	449.2	10.2
EOG 8	Female	54,349	450.1	9.2	55,443	450.2	9.4	57,161	450.1	9.9
	Male	54,850	450.0	9.6	56,800	449.9	9.8	59,578	449.4	10.3
EOC Math I	Female	57,423	249.9	9.2	57,020	251.0	9.2	57,519	250.8	9.8
	Male	59,565	249.6	9.8	59,442	250.4	9.8	61,283	249.7	10.3

9.1.3 Achievement Levels

The achievement level classifications for the population across grades and administrations are displayed in *Table 9.3* through

*Table **9.5***. Note that the cut scores for the base administration (2012–13) were different from 2013–14 administration and beyond. As a result, in 2012–13, NCDPI classified students using 4 achievement levels. From 2013–14 onwards students are classified based on a 5-achievement level scale. Therefore, achievement levels proportions 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 in *Table 9.3* there is no data for Level 3. Levels 3 and 4 proportion for 2012 – 13 has been displayed as Levels 4 and 5 respectively. The largest movement for students classified as college- and career-readiness (Levels 4 and 5) occurred in Math I with a 3.6% increase from 2012–13 to 2013–14. Grades 5 and 3 also had 2.2% and 1.1% more students classified at Levels 4 and 5 between 2012–13 and 2013–14. Between 2013–14 and 2014–15 short-term trends

within grades indicated very small fluctuations, on average about 0.3% for students at Achievement Levels 4 and 5.

The achievement level classifications by gender across grades and administrations are presented in *Table 9.4* and

Table 9.5. These tables follow the same degree of caution as the previous table with regards to interpretation of achievement levels for 2012 – 13. A similar trend as the total population can be observed for each gender. The results across administrations and grades further indicated that in general there are higher proportions of female students over male students who performed at Level 4 or above (college- and career-readiness), with some exceptions. In Grade 4, there are 0.9% to 2.4% more males classified at college- and career-readiness than females over all three administrations. Across all other grade levels anywhere from 0.1% to 3.8% more females were classified as college- and career-readiness.

Table 9.3 Achievement Level Classifications by Grade and Year

	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***
EOG 3	2012–13*	103,594	23.5	28.2		32.3	16.0
	2013–14	112,017	15.5	22.3	12.8	32.6	16.8
	2014–15	116,404	16.3	21.8	12.8	32.5	16.5
EOG 4	2012–13*	110,987	27.0	23.8		32.6	16.7
	2013–14	103,977	20.6	23.9	7.3	30.8	17.5
	2014–15	113,968	21.1	22.9	7.4	30.7	17.9
EOG 5	2012–13*	110,599	26.1	24.3		32.9	16.8
	2013–14	111,718	18.0	24.1	6.0	34.1	17.8
	2014–15	106,611	19.6	22.8	6.1	31.9	19.6
EOG 6	2012–13*	112,257	37.1	22.4		26.0	14.5
	2013–14	111,470	29.1	22.8	7.3	26.1	14.8
	2014–15	114,473	30.4	21.1	7.3	25.7	15.5
EOG 7	2012–13*	111,333	37.2	22.7		25.4	14.8
	2013–14	113,416	29.5	23.1	7.1	25.7	14.7
	2014–15	114,662	31.8	21.1	6.8	24.8	15.5
EOG 8	2012–13*	109,199	37.0	27.5		25.6	10.0
	2013–14	112,243	27.5	29.1	7.7	25.6	10.2
	2014–15	116,739	30.6	26.1	7.3	25.0	11.0
Math I	2012–13*	116,988	38.6	24.0		29.0	8.5
	2013–14	116,462	26.8	18.6	13.4	31.4	9.7
	2014–15	118,802	30.2	17.6	11.5	30.4	10.3

**Cut scores and achievement levels were different in 2012-13 hence the results are not comparable with 2013–14 and 2014–15*

Table 9.4 EOG Achievement Level Classifications by Gender

				% Achievement Level				
	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***
EOG 3	2012–13*	Female	51,003	23.0	28.7		32.9	15.5
		Male	52,591	24.0	27.8		31.7	16.5
	2013–14	Female	55,329	14.5	22.6	13.2	33.4	16.3
		Male	56,688	16.4	22.0	12.4	31.9	17.3
	2014–15	Female	56,938	14.9	22.4	13.2	33.5	16.0
		Male	59,466	17.7	21.3	12.4	31.7	17.0
EOG 4	2012–13*	Female	54,829	27.3	24.7		32.5	15.6
		Male	56,158	26.7	22.9		32.8	17.7
	2013–14	Female	50,995	20.5	24.5	7.5	31.0	16.5
		Male	52,982	20.7	23.2	7.1	30.6	18.5
	2014–15	Female	55,849	20.5	23.6	7.6	31.2	17.0
		Male	58,119	21.6	22.1	7.2	30.3	18.8
EOG 5	2012–13*	Female	54,693	24.6	25.6		33.8	16.0
		Male	55,906	27.5	23.0		32.0	17.5
	2013–14	Female	55,065	16.5	24.6	6.5	35.3	17.1
		Male	56,653	19.5	23.5	5.5	33.0	18.5
	2014–15	Female	51,936	17.2	23.6	6.4	33.7	19.1
		Male	54,675	21.8	22.1	5.7	30.3	20.2
EOG 6	2012–13*	Female	55,440	35.8	22.8		26.8	14.7
		Male	56,817	38.4	22.0		25.3	14.4
	2013–14	Female	54,754	27.3	23.4	7.5	26.7	15.1
		Male	56,716	30.8	22.2	7.0	25.5	14.4
	2014–15	Female	55,841	28.3	21.8	7.5	26.7	15.8
		Male	58,632	32.4	20.5	7.1	24.8	15.2
EOG 7	2012–13*	Female	55,105	35.7	23.3		26.0	15.0
		Male	56,228	38.7	22.0		24.8	14.6
	2013–14	Female	55,884	27.5	23.6	7.5	26.4	15.0
		Male	57,532	31.4	22.5	6.8	25.0	14.4
	2014–15	Female	55,933	29.1	21.8	7.1	25.9	16.2
		Male	58,729	34.5	20.5	6.5	23.7	14.8
EOG 8	2012–13*	Female	54,349	36.2	28.4		25.9	9.5
		Male	54,850	37.8	26.7		25.2	10.4
	2013–14	Female	55,443	26.0	30.2	7.9	26.1	9.8
		Male	56,800	29.0	28.0	7.4	25.1	10.5
	2014–15	Female	57,161	28.4	27.1	7.6	26.0	10.9
		Male	59,578	32.7	25.2	6.9	24.1	11.1

Table 9.5 EOC Math I Achievement Level Classifications by Gender

				% Achievement Level				
		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
Math I	2012–13*	Female	57,423	37.2	25.4		29.5	7.9
		Male	59,565	40.0	22.5		28.5	9.0
	2013–14	Female	57,020	24.5	19.5	14.2	32.4	9.5
		Male	59,442	29.1	17.9	12.5	30.5	10.0
	2014–15	Female	57,519	27.0	18.1	12.3	32.3	10.4
		Male	61,283	33.2	17.1	10.8	28.7	10.2

**Cut scores for Proficiency levels were different in 2012-13 hence the results are not comparable with 2013–14 and 2014–15*

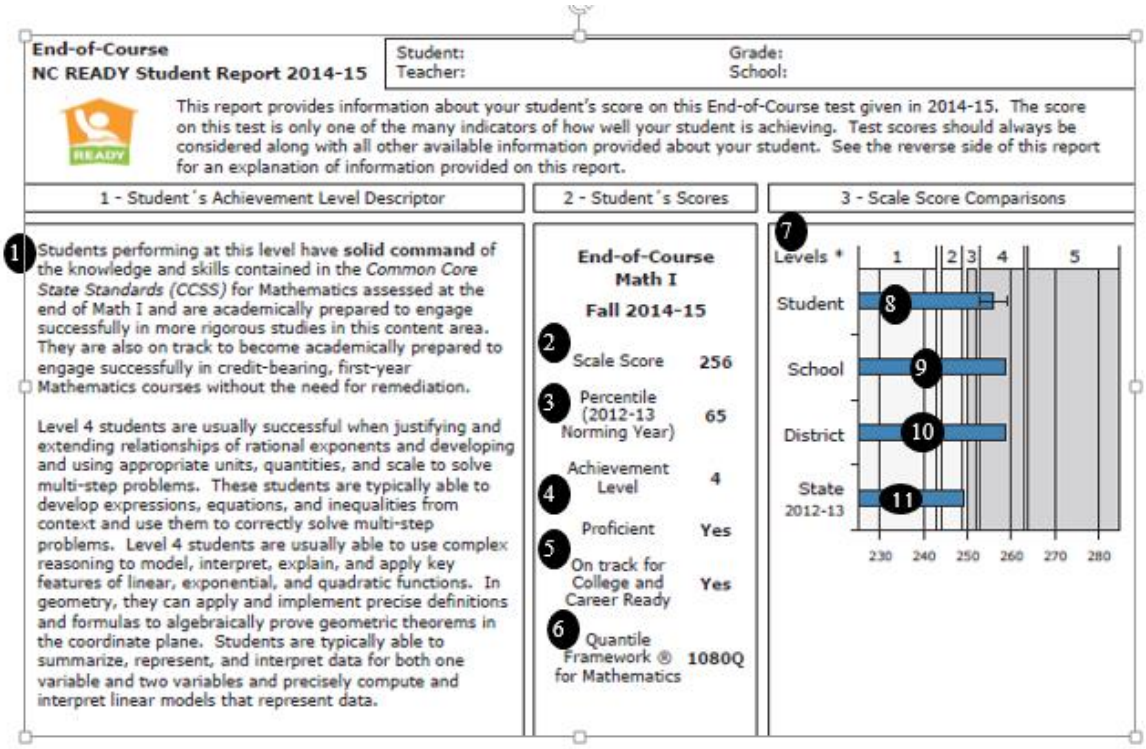
9.2 Sample Reports

To address fairness in reporting and valid interpretation and use of individual test scores, NCDPI produces a series of custom reports along with interpretive guides. This ensures students, teachers and stakeholders are able to make valid interpretations about test scores. The sample reports, along with the complete interpretive guide, is published on the NCDPI public webpage. This next section presents examples of the score reports with brief explanations about their use and interpretation.

9.2.1 Individual Student Report (ISRs)

For students at grades 3–8, the ISR for the EOG provides information concerning performance on the EOG for ELA/reading and mathematics. For students at grades 5 and 8, ISRs provide information about the EOG science assessments. A sample ISR report is shown in figure 9.8. Key features are labeled and explained in the *Index of Terms by Label Number* section in the ISR.

Figure 9.8 Sample Individual Student Report for Math I EOC Assessment



The Student's Achievement Level Descriptor section (label 1) describes the expected performance of the student given his or her score on the assessments as agreed upon during standard setting. The achievement level descriptors can be viewed at http://www.ncpublicschools.org/accountability/testing//shared/achievelevel.

The Scale Score (label 2) shows the student's transformed score obtained from the test administration. The Percentile (2013 Norming Year) (label 3) compares a student's performance on the assessment relative to all North Carolina students at that grade level who took the assessment in the norming year (2013). The norming year for an assessment is generally the first year the assessment was administered, and data from that year was used to set achievement levels. The percentile shows a student performed at a level better than the stated percentage displayed on the report. For example, the student with a scale score of 256 on Math I, and a percentile of 65, is said to have performed better than 64% of students who took the assessment during the norming year.

The Achievement Level (label 4) 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 (label 5). Achievement levels of 4 and 5 indicate college- and career-readiness.

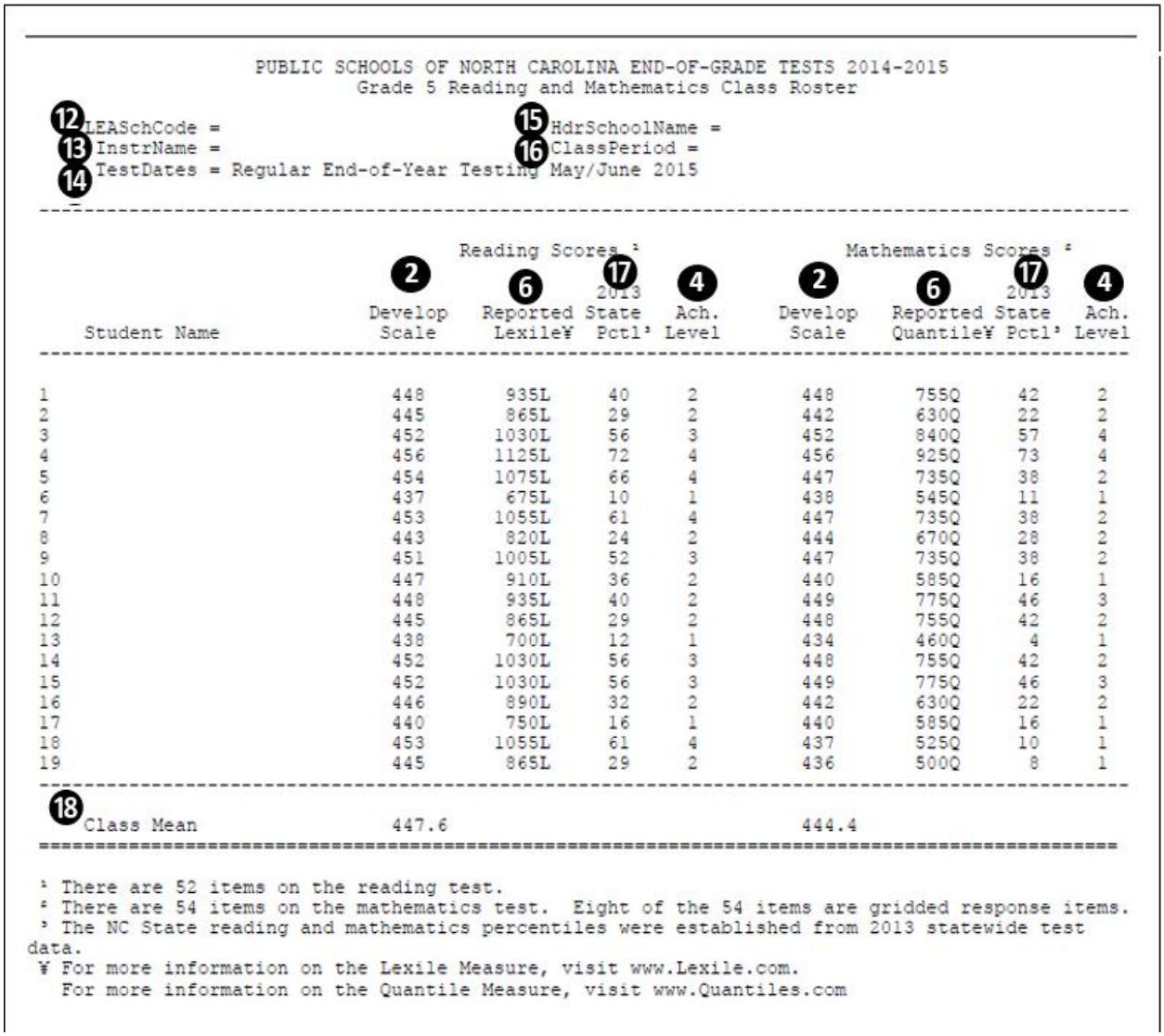
The Quantile Framework for Mathematics (label 6) shows the Quantile Framework level that is associated with the EOC or EOG scale score. Additional information on Quantiles can be found at https://www.quantiles.com/.

The Levels (label 7) refers to Achievement Levels, which 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. The Student (label 8) scale score is represented by a blue bar. Surrounding the student's scale score is a confidence interval, indicated by a black line. The confidence interval indicates the range of scores that would likely result if the same student completed similar tests many times. For example, if this student were to take a similar tests a second time, the scale score would very likely fall to around level 3 or 4. The average school score (label 9) is represented by this blue bar. The average scale score for the school is based on the fall or spring test administration for the given school year of the report. The average district score (label 10) is represented by the third blue bar. The average scale score for the district is 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 Reports take on many different combinations. A Class Roster Report can contain grade-specific student scores for each content area independently, or a class roster report can contain grade-specific student scores for combinations of content areas. The most typical combination for the EOG is a Class Roster Report that displays reading and mathematics scores together on one report for a specific grade. *Figure 9.9* displays a sample EOG Class Roster Report and a brief explanation of the labels listed below the report. This report is often produced at the class level and the school level. The report's features and layout do not differ across levels.

Figure 9.9 Sample Class Roster Report for EOG Grade 5



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2014-2015
Grade 5 Reading and Mathematics Class Roster

(12) LEASchCode = (15) HdrSchoolName =
(13) InstrName = (16) ClassPeriod =
(14) TestDates = Regular End-of-Year Testing May/June 2015

	Reading Scores [1]				Mathematics Scores [2]			
Student Name	(2) Develop Scale	(6) Reported Lexile¥	(17) 2013 State Pctl[3]	(4) Ach. Level	(2) Develop Scale	(6) Reported Quantile¥	(17) 2013 State Pctl[3]	(4) Ach. Level
1	448	935L	40	2	448	755Q	42	2
2	445	865L	29	2	442	630Q	22	2
3	452	1030L	56	3	452	840Q	57	4
4	456	1125L	72	4	456	925Q	73	4
5	454	1075L	66	4	447	735Q	38	2
6	437	675L	10	1	438	545Q	11	1
7	453	1055L	61	4	447	735Q	38	2
8	443	820L	24	2	444	670Q	28	2
9	451	1005L	52	3	447	735Q	38	2
10	447	910L	36	2	440	585Q	16	1
11	448	935L	40	2	449	775Q	46	3
12	445	865L	29	2	448	755Q	42	2
13	438	700L	12	1	434	460Q	4	1
14	452	1030L	56	3	448	755Q	42	2
15	452	1030L	56	3	449	775Q	46	3
16	446	890L	32	2	442	630Q	22	2
17	440	750L	16	1	440	585Q	16	1
18	453	1055L	61	4	437	525Q	10	1
19	445	865L	29	2	436	500Q	8	1
(18) Class Mean	447.6				444.4			

[1] There are 52 items on the reading test.
[2] There are 54 items on the mathematics test. Eight of the 54 items are gridded response items.
[3] The NC State reading and mathematics percentiles were established from 2013 statewide test data.
¥ For more information on the Lexile Measure, visit www.Lexile.com.
For more information on the Quantile Measure, visit www.Quantiles.com

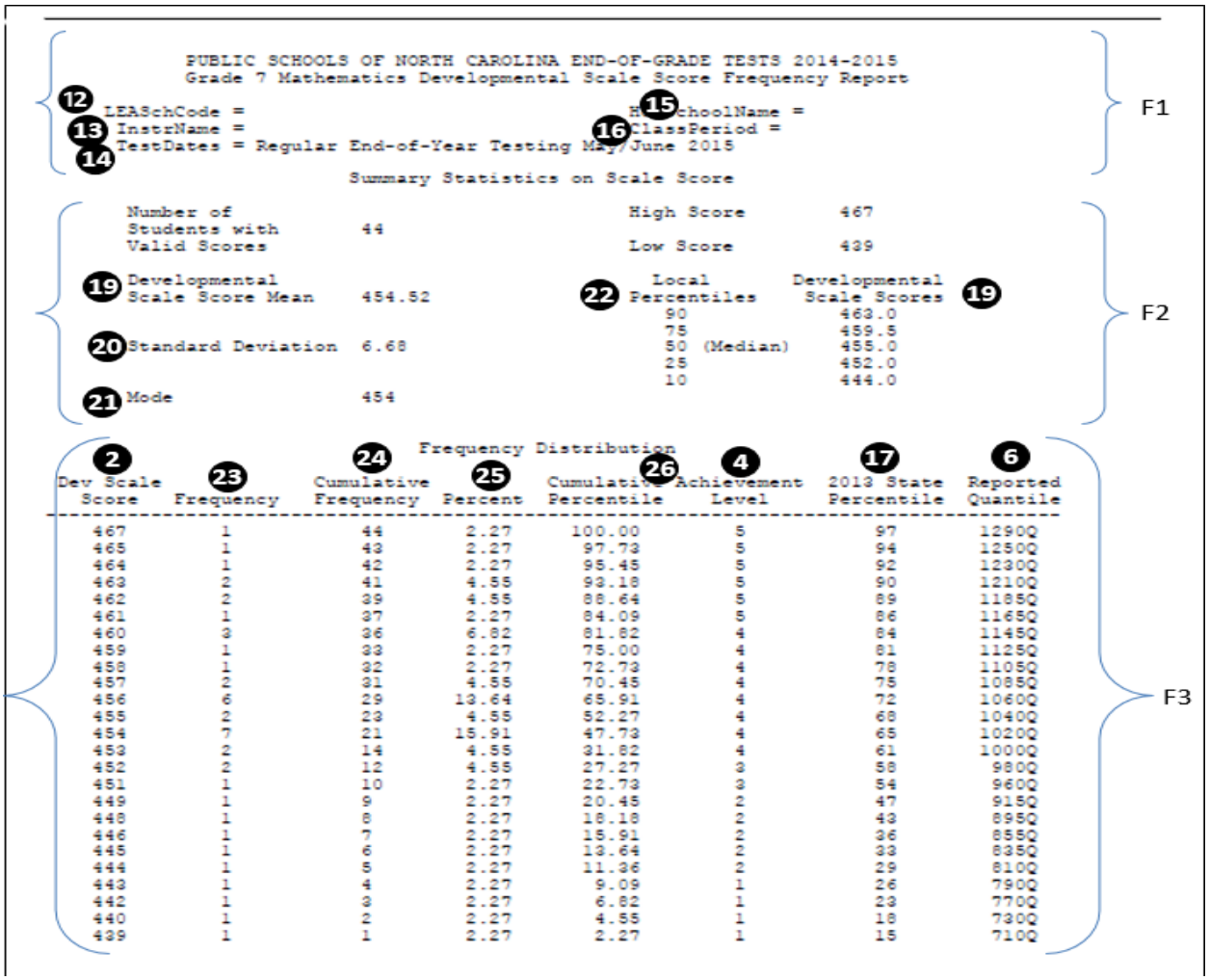
General information is reported from label 12 to label 16. LEASchCode (label 12) refers to the Local Education Agency (LEA) school code. InstrName (label 13) refers to the instructor's name. TestDates (label 14) refers to the time of year in which the exam was administered. HdrSchoolName (label 15) refers to the school name. ClassPeriod (label 16) refers to the class period. This report presents the same information as the ISR but the main difference is that it displays the score summary for the entire class. For mathematics, Reported Quantile (label 6) shows the Quantile Framework® level that is associated with the EOG math scale score. Note that this Quantile® score for math is

similar to the Quantile score for ELA. Additional information on Quantile measures can be found at http://www.Quantiles.com. The Class Mean (label 18) is the average of the class scores. The mean is the sum of all scores in the roster divided by the number of scores in the roster. For example, the class in the report got an averaged scale score at 447.6 in reading and 444.4 in math.

9.2.3 Scale Score Frequency Reports

Frequency tables are used to summarize large quantities of scores. The Scale Score Frequency Reports available in WinScan are used to summarize scale score information at the class, school, district, and state levels. The WinScan Scale Score Frequency Report presents 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 9.10 presents a sample Score Frequency Report for the EOG Mathematics Assessment.

Figure 9.10 Sample Score Frequency Report for EOG Grade 7 Math.



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2014-2015
Grade 7 Mathematics Developmental Scale Score Frequency Report

(12) LEASchCode = (15) HdrSchoolName =
(13) InstrName = (16) ClassPeriod =
(14) TestDates = Regular End-of-Year Testing May/June 2015

F1

Summary Statistics on Scale Score

Number of Students with Valid Scores	44	High Score	467
		Low Score	439
(19) Developmental Scale Score Mean	454.52		
(20) Standard Deviation	6.68		
(21) Mode	454		

(22) Local Percentiles	Developmental Scale Scores (19)
90	463.0
75	459.5
50 (Median)	455.0
25	452.0
10	444.0

F2

Frequency Distribution

(2) Dev Scale Score	(23) Frequency	(24) Cumulative Frequency	(25) Percent	(26) Cumulative Percentile	(4) Achievement Level	(17) 2013 State Percentile	(6) Reported Quantile
467	1	44	2.27	100.00	5	97	1290Q
465	1	43	2.27	97.73	5	94	1250Q
464	1	42	2.27	95.45	5	92	1230Q
463	2	41	4.55	93.18	5	90	1210Q
462	2	39	4.55	88.64	5	89	1185Q
461	1	37	2.27	84.09	5	86	1165Q
460	3	36	6.82	81.82	4	84	1145Q
459	1	33	2.27	75.00	4	81	1125Q
458	1	32	2.27	72.73	4	78	1105Q
457	2	31	4.55	70.45	4	75	1085Q
456	6	29	13.64	65.91	4	72	1060Q
455	2	23	4.55	52.27	4	68	1040Q
454	7	21	15.91	47.73	4	65	1020Q
453	2	14	4.55	31.82	4	61	1000Q
452	2	12	4.55	27.27	3	58	980Q
451	1	10	2.27	22.73	3	54	960Q
449	1	9	2.27	20.45	2	47	915Q
448	1	8	2.27	18.18	2	43	895Q
446	1	7	2.27	15.91	2	36	855Q
445	1	6	2.27	13.64	2	33	835Q
444	1	5	2.27	11.36	2	29	810Q
443	1	4	2.27	9.09	1	26	790Q
442	1	3	2.27	6.82	1	23	770Q
440	1	2	2.27	4.55	1	18	730Q
439	1	1	2.27	2.27	1	15	710Q

F3

The Score Frequency Report consists of three sections: the header (F1), a summary table of statistics (F2), and a score frequency table (F3).

The first line of the sample Score Frequency Report header describes the type of assessment (EOG) and the school year (2014–15). The second line of the header displays the specific type of assessment, the grade, the subject area, and the type of report. The LEASchCode (label 12) indicates the Local Educational Agency school code, the InstrName (label 13) indicates the instructor's name; TestDates (label 14) indicates the time of year in which the exam was administered, the HdrSchoolName (label 15) indicates the school name, and the ClassPeriod (label 16) indicates the class period.

The arithmetic mean of the scale score was 454.52 (label 19), the standard deviation was 6.68 (label 20), and the mode was 454 (label 21). The percentile scores are

listed at the far right of the table (label 22). The scale scores are listed for the 10th, 25th, 50th, 75th, and 90th percentiles (label 19). In this sample, a scale score of 459.5 corresponds to a percentile of 75. This means that 75% of the 44 students earned a score of 459.5 or less.

In the score frequency table (F3), the Dev Scale Score column (label 2) displays every score earned by the 44 students. The Frequency column (label 23) on the report displays the number of students with their respective scale score earned. For example, 6 students earned a scale score of 456. A "Missing" label would indicate that one student did not receive a score.

The Cumulative Frequency column (label 24) displays the total number of students who earned up to and including a given scale score. This column shows 29 students earned up to and including a scale score of 456.

The Percent column (label 25) presents the percentage of students that earned a given scale score (number of students that earned the score divided by total number of observations). This column shows that 13.64% of the students earned a score of 456.

The Cumulative Percentile column (label 26) displays the percentage of students that earned up to and including a given scale score. This column shows 65.91% of the students earned up to and including a scale score of 456.

The Achievement Level column (label 4) displays the achievement level associated with each scale score. In this example, a scale score of 456 corresponds to an achievement level of 4.

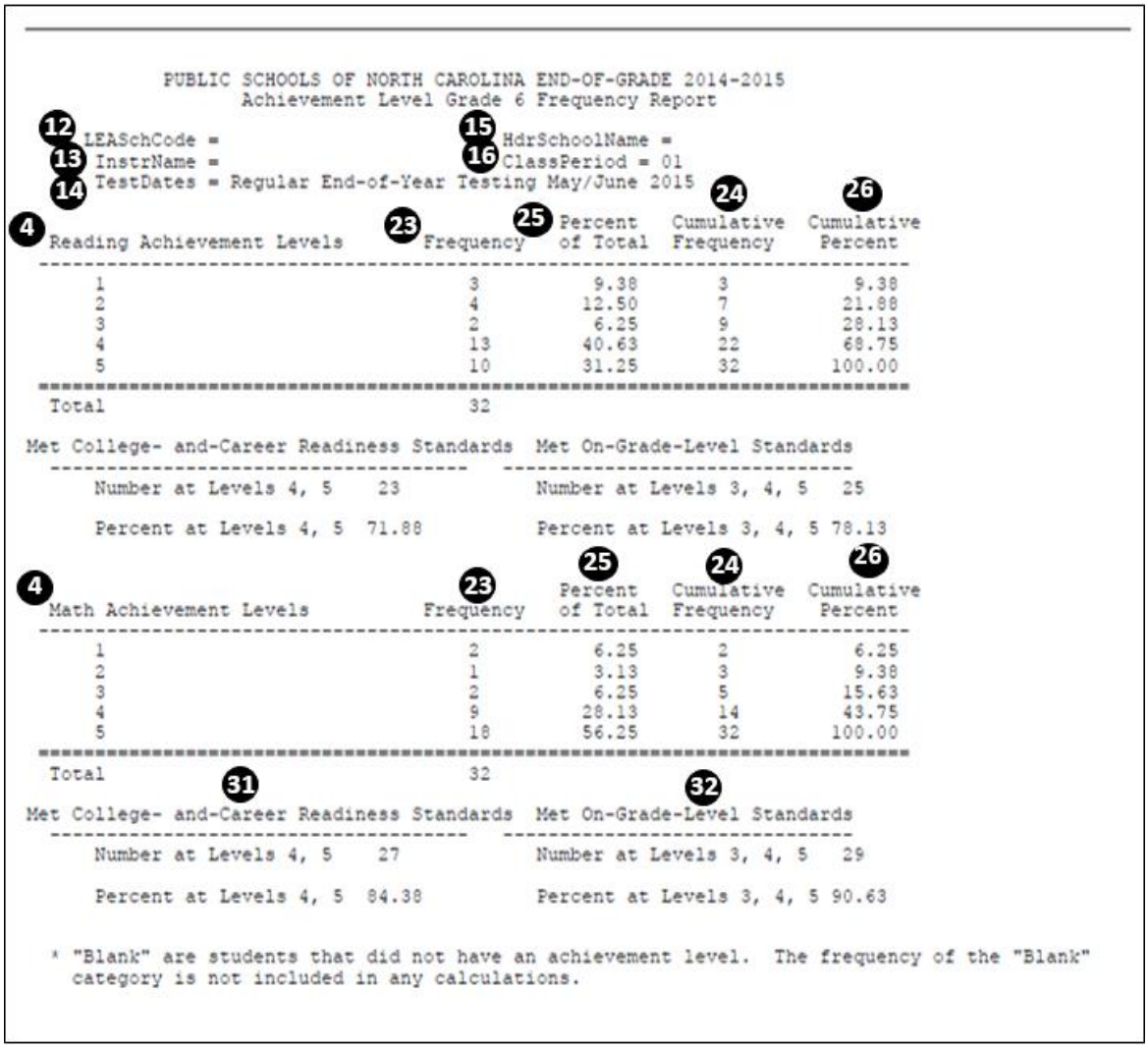
The 2013 State Percentile column (label 17) displays to the ELA/reading and mathematics percentiles that were established from 2013 statewide assessment data. This column shows that a scale score of 456 was in the 72nd percentile in 2013. The Reported Quantile column (label 6) displays the Quantile Score. This example shows that a scale score of 456 is linked to a Quantile of 1060Q.

9.2.4 Achievement Level Frequency Reports

A sample Achievement Level Frequency Report for EOG ELA and Mathematics assessment is displayed in Figure 9.11. This report presents similar information as the

Scale Score Frequency Report described above but uses achievement level as the main reporting variable.

Figure 9.11 Sample Achievement Level Frequency Report for EOG Grade 6 ELA and Math.



PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE 2014-2015
Achievement Level Grade 6 Frequency Report

12 LEASchCode = 15 HdrSchoolName =
13 InstrName = 16 ClassPeriod = 01
14 TestDates = Regular End-of-Year Testing May/June 2015

4 Reading Achievement Levels	23 Frequency	25 Percent of Total	24 Cumulative Frequency	26 Cumulative Percent
1	3	9.38	3	9.38
2	4	12.50	7	21.88
3	2	6.25	9	28.13
4	13	40.63	22	68.75
5	10	31.25	32	100.00
Total	32			

Met College- and-Career Readiness Standards		Met On-Grade-Level Standards	
Number at Levels 4, 5	23	Number at Levels 3, 4, 5	25
Percent at Levels 4, 5	71.88	Percent at Levels 3, 4, 5	78.13

4 Math Achievement Levels	23 Frequency	25 Percent of Total	24 Cumulative Frequency	26 Cumulative Percent
1	2	6.25	2	6.25
2	1	3.13	3	9.38
3	2	6.25	5	15.63
4	9	28.13	14	43.75
5	18	56.25	32	100.00
Total	32			

Met College- and-Career Readiness Standards		Met On-Grade-Level Standards	
Number at Levels 4, 5	27	Number at Levels 3, 4, 5	29
Percent at Levels 4, 5	84.38	Percent at Levels 3, 4, 5	90.63

31 32

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

LEASchCode (label 12) indicates the Local Educational Agency school code, the InstrName (label 13) indicates the instructor's name, TestDates (label 14) indicates the exam was administered as a regular End-of-Year assessment in May/June 2015, the HdrSchoolName (label 15) indicates the school name, and the ClassPeriod (label 16) indicates the class period.

The Reading and Mathematics Achievement Levels column (label 4) presents every achievement level earned by the students. Students who do not have an achievement level are classified as “blank”.

Columns labelled 23, 24, 25 and 26 are interpreted in a similar manner as described for Scale Score Frequency Report.

The summary statistics just below the frequency table show 23 of 32 students were classified as Level 4 or 5, and 25 of the 32 were classified as Level 3, 4, or 5 in Reading. This corresponds to 78.13% of the students at grade-level proficient (levels 3 and above) and 71.88% at college-and-career ready (levels 4 and above) in Reading. In Math, 27 of 32 students were classified as Level 4 or 5, and 29 of the 32 were classified as Level 3, 4, or 5. This indicates that 90.63% of the students were grade-level proficient (levels 3 and above) and 84.38% were college-and-career ready (levels 4 and above) in math.

9.2.5 Goal Summary Reports

The 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 level. Other reporting categories are beginning to be integrated that will provide teachers with additional information. For example, subscale scores for EOG Mathematics will be reported with regard to items designated for calculator active sections versus calculator inactive sections on the goal summary report. Additional information has already been incorporated for EOG reading in the goal summary report, which contains goal-level score reporting as well as subscale scores reflecting items related to literary reading versus items related to informational reading. A subscale reported in the goal summary is only meant to provide teachers with formative information to help instruction.

Figure 9.12 shows a sample goal summary report. Key features are labeled and explained in the *Index of Terms by Label Number* in the report. The standard protocol for reporting subscale scores requires that any goal 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.

Figure 9.12 Sample Goal Summary Report for EOG Grade 8 ELA and Math.

PUBLIC SCHOOLS OF NORTH CAROLINA END-OF-GRADE TESTS 2013-2014
Grade 8 Goal Summary Report
Regular test administration

33 SystemCode =

34 SystemName =

	19 Developmental Scale Score Mean	35 Number of Valid Scores	28 Pct of Read Items per Form [1]	29 Weighted Mean Pct Correct	30 Diff from 2013 State Mean Pct Correct [2]
Reading	455.6	1840	100.0		
State 2013 [3]	458.7	108923			
Common Core English Language Arts Concepts					
Language			20.3	64.5	-5.1
Reading: Literature			33.6	61.3	-3.1
Reading: Informational Text			46.2	56.1	-6.5

	19 Developmental Scale Score Mean	35 Number of Valid Scores	28 Pct of Math Items per Form [1]	29 Weighted Mean Pct Correct	30 Diff from 2013 State Mean Pct Correct [2]
Mathematics	447.6	1843	100.0		
State 2013 [3]	450.0	109580			
Calculator Inactive			30.0	35.0	-5.1
Gridded Response Items			18.0	22.5	-5.2
Calculator Active			70.0	48.7	-4.6
Common Core Mathematics Domains					
Functions			24.0	44.8	-5.4
The Number System			6.0	17.8	-4.8
Expressions and Equations			32.0	42.5	-5.8
Geometry			22.0	54.3	-2.2
Statistics and Probability			16.0	45.2	-5.2

[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 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 Common Core English Language Arts Standard can be found at
http://www.corestandards.org/ELA-Literacy

The Grade 8 Common Core Mathematics Overview can be found at
http://www.corestandards.org/Math/Content/8/introduction

In this sample, SystemCode (label 33) indicates the Local Education Agency (LEA) school code (label 33) and SystemName (label 34) refers to LEA or district name. The Developmental Scale Score Mean columns for Reading and Mathematics respectively (label 19) present the average of a group scale scores. Number of Valid Scores column (label 35) presents the number of valid scores. For example, EOG Grade 8 ELA/Reading administrated in 2013 has 108923 valid scores in North Carolina with a mean at 458.7.

The Pct of Read/Math Items per Form column (label 28) presents the percent of the items per form that align with each content goal. In ELA/Reading, 33.6% items in each form come from "Reading: Literature" content. The Weighted Mean Pct Correct column (label 29) provides averaged scores for each content area 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 (label 30) 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. For example, students' average score for the content "Reading: Literature" is 3.1 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 goal or objective level may be different in this year's forms and those from 2013.

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 EOG and EOC assessments. Evidence presented in these sections includes reliability, standard error estimates and classification consistency summary of reported achievement levels, and an exploratory Principal Component Analysis in support of the unidimensional analysis and interpretation of EOG and EOC scores. The final sections of the chapter document content validity evidence summarized from the alignment study, evidence based on relation to other variables summarized from the EOG/EOC Quantile Framework linking study, and the last part presents summary of procedures used to ensure EOG and EOC assessments are accessible and fair to all students.

10.1 Reliability Evidence of Math EOG and EOC Math I

Internal consistency reliability estimates provide a sample base summary statistic that describes the proportion of reported score which is the 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 is 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 0.0 to 1.0, where a coefficient of 1.0 refers to a perfectly reliable measure with no error. For high-stakes assessments, alpha estimates of 0.85 or higher are generally desirable. Cronbach's alpha (Cronbach, 1951) is calculated as

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

(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 that reliability estimates are less informative in describing the accuracy of individual students' scores, since they are sample based.

Table 10.1 EOG Math and EOC Math I Reliabilities by Form and Subgroup

EOG/EOC and Form		Gender		Ethnicity[k]			All
		Female	Male	Black	Hispanic	White	
Grade 3	*A*	0.91	0.91	0.89	0.89	0.91	0.91
	B	0.91	0.92	0.90	0.90	0.91	0.92
	C	0.91	0.91	0.88	0.89	0.90	0.91
Grade 4	*A*	0.92	0.92	0.90	0.91	0.92	0.92
	B	0.92	0.92	0.89	0.90	0.91	0.92
	C	0.92	0.92	0.89	0.91	0.92	0.92
Grade 5	*A*	0.91	0.92	0.88	0.90	0.91	0.92
	B	0.91	0.92	0.89	0.90	0.91	0.92
	C	0.91	0.92	0.88	0.89	0.91	0.91
Grade 6	*A*	0.93	0.93	0.89	0.91	0.93	0.93
	B	0.93	0.93	0.89	0.91	0.93	0.93
	C	0.93	0.93	0.89	0.91	0.93	0.93
Grade 7	*A*	0.93	0.93	0.90	0.91	0.93	0.93
	B	0.93	0.93	0.89	0.91	0.93	0.93
	C	0.93	0.94	0.90	0.92	0.93	0.94
Grade 8	*A*	0.92	0.93	0.88	0.90	0.92	0.92
	B	0.91	0.92	0.86	0.89	0.92	0.92
	C	0.92	0.93	0.88	0.90	0.92	0.92
Math I	*A*	0.90	0.91	0.85	0.89	0.91	0.91
	B	0.90	0.91	0.85	0.88	0.91	0.91
	M	0.90	0.91	0.85	0.88	0.91	0.90
	N	0.89	0.91	0.84	0.87	0.90	0.90

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

Table 10.1 shows Cronbach's alpha reliability estimates for all Math EOG and EOC forms by grade and major demographic variables. Across all forms, reliability estimates based on the 2012–2013 population range from 0.90 to 0.94. Subgroup reliabilities are also consistent across forms and subgroup, and for the most part, they are consistent and higher than the 0.85 threshold. Exceptions to this general trend are recorded in Black subgroup reliabilities for Math I form N in which the reported alpha is 0.84.

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 student scores 100 with SEM of 2, then one can conclude with a 68% certainty (1 standard error) that the student score is accurate within plus or minus 2 points. 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 error of measurement at the scale score cuts for achievement levels for the North Carolina EOG and EOC Math assessments are provided in Table 10.2 below. For students with scores within 2 standard deviations of the mean (95% of the students), standard errors are typically 2 to 3 points. For most of the EOG and EOC Math scale scores, the standard error of measurement in the middle range of scores, particularly at the cut point between Level 2 and Level 3, is generally around 3 points. Scores at the lower and higher ends of the scale (above the 97.5^{th} percentile and below the 2.5^{th} percentile) have standard errors of measurement of 5 to 6 points. This is typical for extreme scores which allow less measurement precision because of a lack of informative items at those ability ranges.

Table 10.2 Conditional Standard Errors at Achievement Level Cuts by Form and Grade Level

		LOSS		Level 2		Level 3		Level 4		Level 5		HOSS	
Math	Form	Loss	SE	Partial	SE	Sufficient	SE	Solid	SE	Superior	SE	Hoss	SE
EOG 3	A	422	5	440	3	448	3	451	3	460	3	472	5
	B	421	5	440	3	448	3	451	2	460	3	472	5
	C	422	5	440	4	448	3	451	3	460	3	473	5
EOG 4	A	424	5	441	3	449	2	451	2	460	3	473	5
	B	424	5	441	3	449	3	451	2	460	3	475	5
	C	425	5	441	3	449	2	451	2	460	3	473	5
EOG 5	A	426	5	441	4	449	3	451	2	460	3	475	5
	B	424	5	441	4	449	3	451	2	460	3	474	5
	C	424	5	441	4	449	3	451	2	460	3	474	5
EOG 6	A	428	5	444	3	451	2	453	2	461	2	476	5
	B	427	5	444	3	451	2	453	2	461	2	476	5
	C	427	5	444	3	451	2	453	2	461	2	476	5
EOG 7	A	428	5	444	3	451	2	453	2	461	2	476	5
	B	428	5	444	3	451	2	453	2	461	2	476	5
	C	429	5	444	3	451	2	453	2	461	2	476	5
EOG 8	A	426	5	444	4	452	2	454	2	463	3	477	5
	B	427	5	444	4	452	3	454	2	463	2	478	5
	C	427	5	444	4	452	2	454	2	463	2	477	5
Math I	A	227	5	244	4	250	3	253	3	264	3	281	5
	B	229	5	244	4	250	3	253	3	264	3	281	5
	C	227	6	244	4	250	3	253	3	264	3	282	5
	M	227	5	244	4	250	3	253	3	264	3	280	5
	N	229	5	244	4	250	3	253	3	264	3	281	5
	O	227	6	244	4	250	3	253	3	264	2	281	5

The SEs at Level 2 and Level 3 across forms and grades ranged from 2 to 4, and Level 4 ranged from 2 to 3. One useful application of the conditional SEs is that it can be used to estimate a band of scores around any scale score or cut score where a decision has to be precise. For example, the on-grade proficiency (Level 3) cut score for grade 3 math is 448. A student who took Form A and scored 448 with a SE of 3 has a 68% probability that his or her true score or ability ranges from 445 to 451 (448±1*3) when reported with

a 1 standard error level of precision. Similarly, if an educator wants to estimate the students' true score with less precision say 2 standard error then the 95% confidence interval of the student predicted ability will be from 442 to 454 (448±2*3).

10.3 Evidence of Classification Consistency

The *No Child Left Behind Act* of 2001 (2002) and subsequent Race to the Top Act of 2009 (2009) emphasized the measurement of adequate yearly progress (AYP) with respect to 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 for Educational and Psychological Testing (AERA, APA, & NCME, 2014) 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 know" (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[l]. The program provides results for both the Hanson and Brennan (1990) and Livingston and

[l] 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

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 the math EOG and EOC followed the HB procedures.

Table 10.3 shows the decision accuracy and consistency indexes for achievement levels at each grade. Overall, the values indicate good classification accuracy (ranging from 0.90 to 0.96) and consistency (from 0.86 to 0.95). For example, if Grade 3 Math students who were classified as Level 2 were to take a non-overlapping, equally difficult form a second time, 91% of them 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.
Grade 3	0.93	0.91	0.92	0.89	0.92	0.88	0.93	0.90
Grade 4	0.93	0.90	0.92	0.89	0.92	0.89	0.94	0.91
Grade 5	0.93	0.90	0.92	0.88	0.92	0.89	0.94	0.92
Grade 6	0.92	0.89	0.92	0.89	0.93	0.90	0.96	0.94
Grade 7	0.92	0.88	0.93	0.90	0.93	0.91	0.96	0.94
Grade 8	0.91	0.87	0.92	0.89	0.93	0.90	0.96	0.95
Math I	0.90	0.86	0.90	0.86	0.91	0.88	0.96	0.95

Note: Acc = Accuracy; Con = Consistency

10.4 EOG and EOC Dimensionality Analysis

Evidence of overall dimensionality for EOG and EOC Math assessments was 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

https://www.education.uiowa.edu/centers/casma/computer-programs#de748e48-f88c-6551-b2b8-ff00000648cd.

in the data. Two commonly used criteria to decide the number of meaningful dimensions for a set of observed variables are:

- Retain components whose eigenvalues are greater than the average of all the eigenvalues, which is usually 1.
- Use scree graph which is a plot of eigenvalues against and count the number of component 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, or from the polychoric correlation matrix for categorical scored responses, to determine the number of meaningful components. Scree graphs from the PCA analysis by grade and forms are shown in *Figure 10.1* through *Figure 10.7* for the first 16 components. The eigenvalue of the first component which 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 grade 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 1 dominant component to explain a significant amount of the total variance in the observed correlation matrices for EOG and EOC forms. Evaluation of the scree graph with the distinct break of the linear trend after the first dominant component present enough exploratory evidence in support of the assumption of unidimensionality of EOG and EOC assessments. Thus PCA results with one dominant component support treating the data as unidimensional.

Figure 10.1 Math Grade 3 Scree Plot of Operational Forms

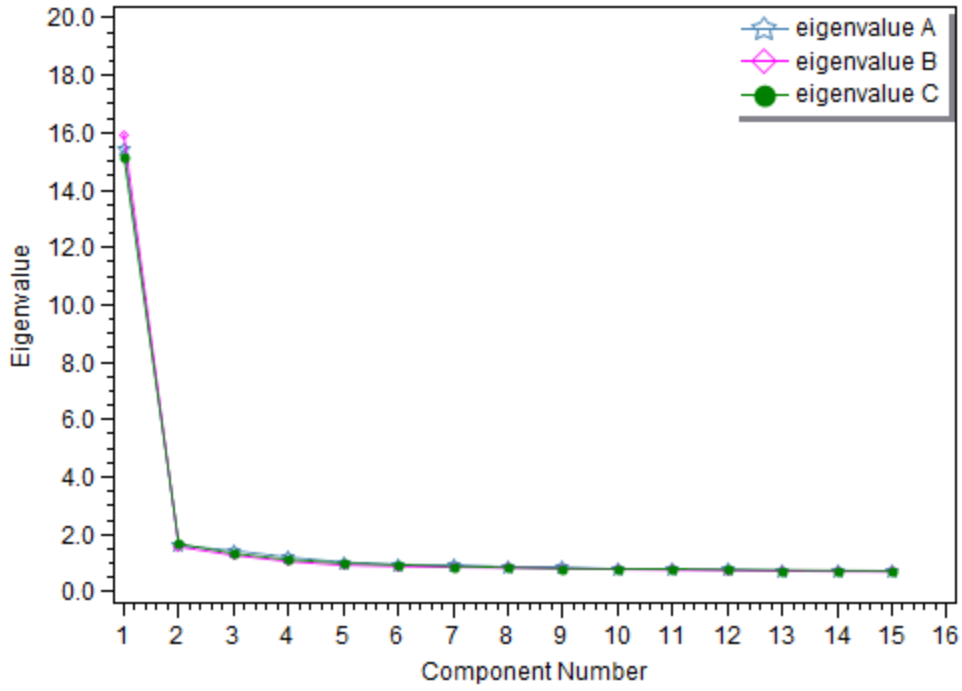


Figure 10.2 Math Grade 4 Scree Plot of Operational Forms

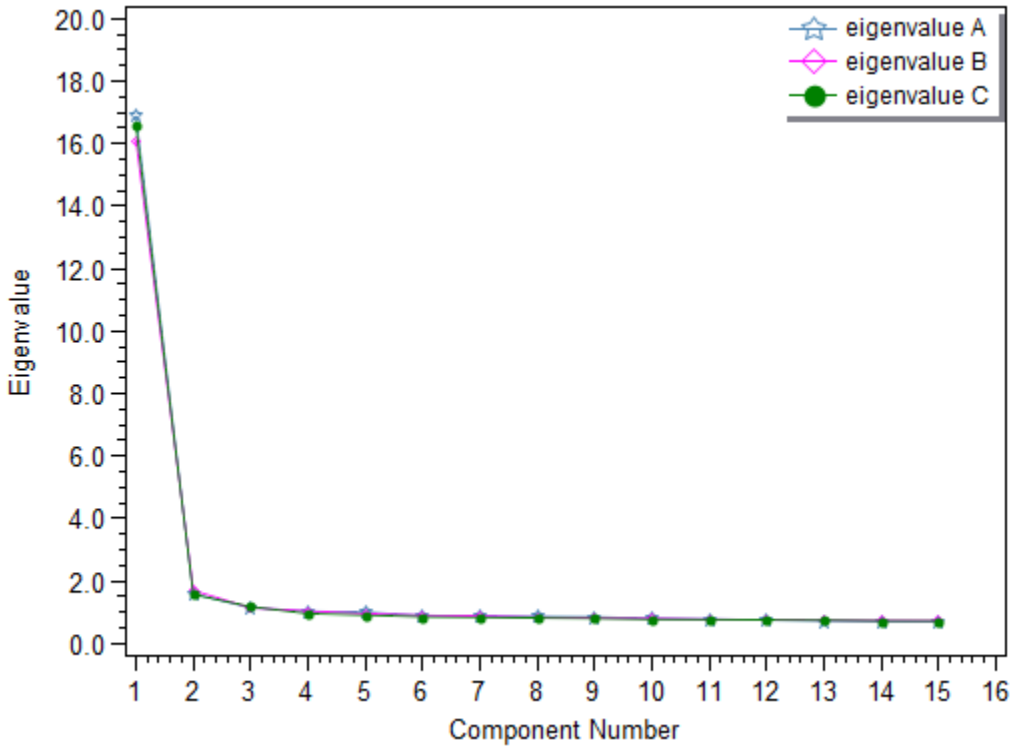


Figure 10.3 Math Grade 5 Scree Plot of Operational Forms

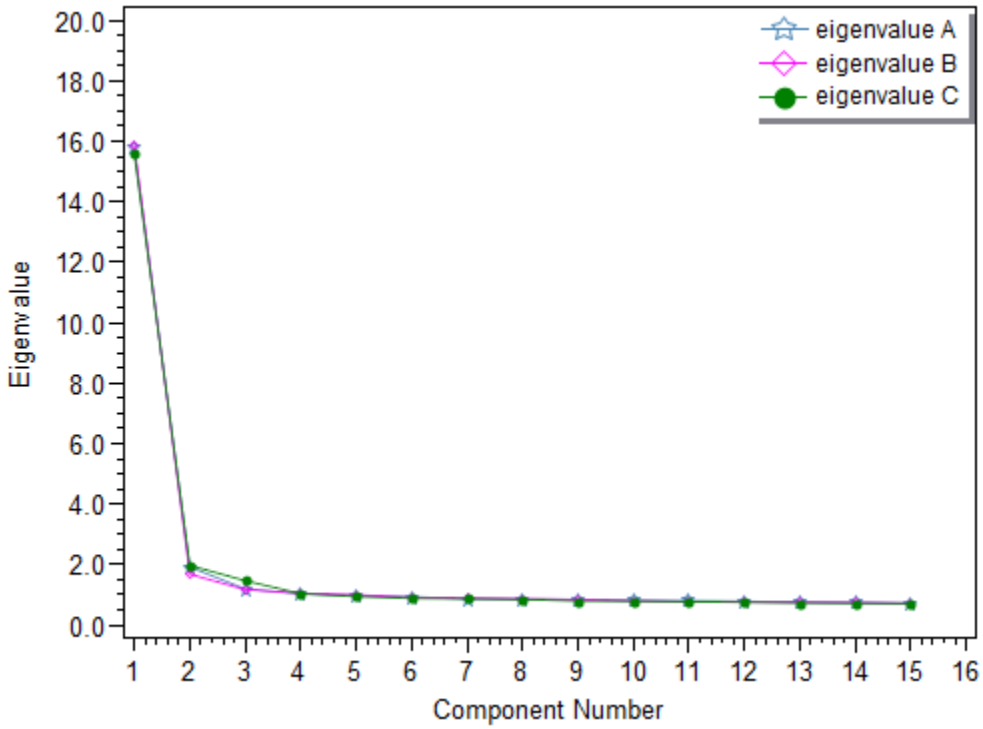


Figure 10.4 Math Grade 6 Scree Plot of Operational Forms

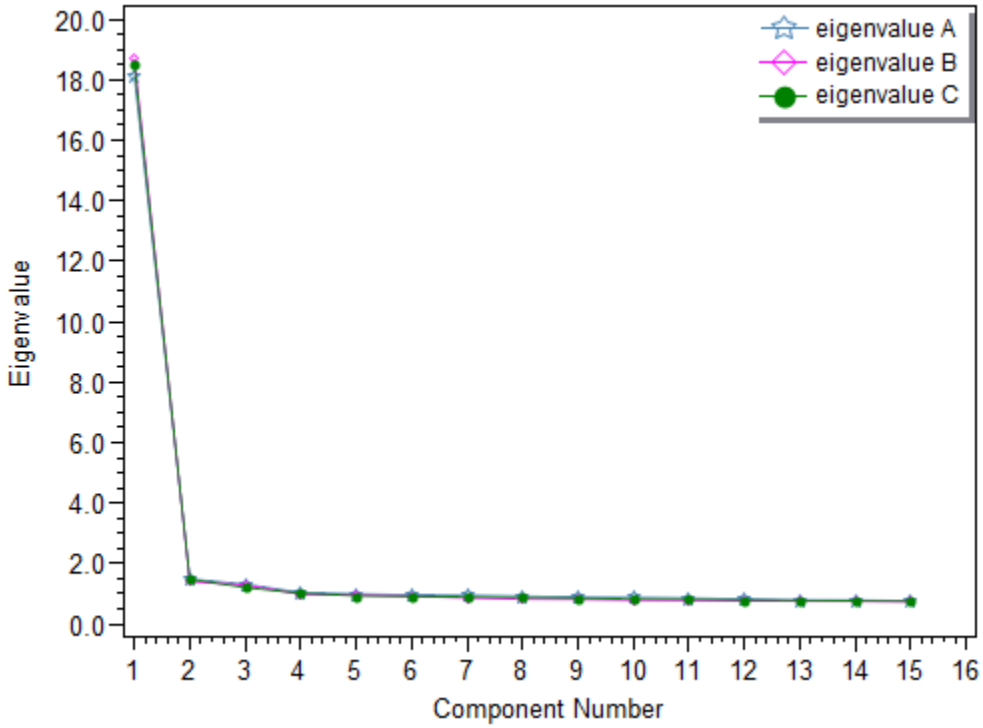


Figure 10.5 Math Grade 7 Scree Plot of Operational Forms

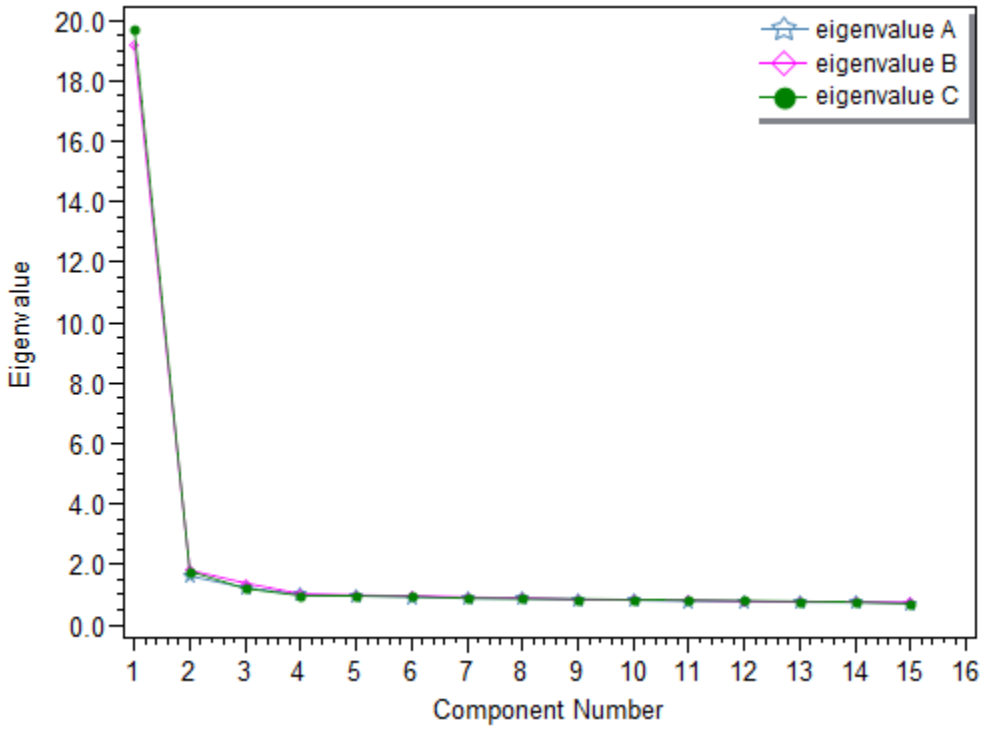


Figure 10.6 Math Grade 8 Scree Plot of Operational Forms

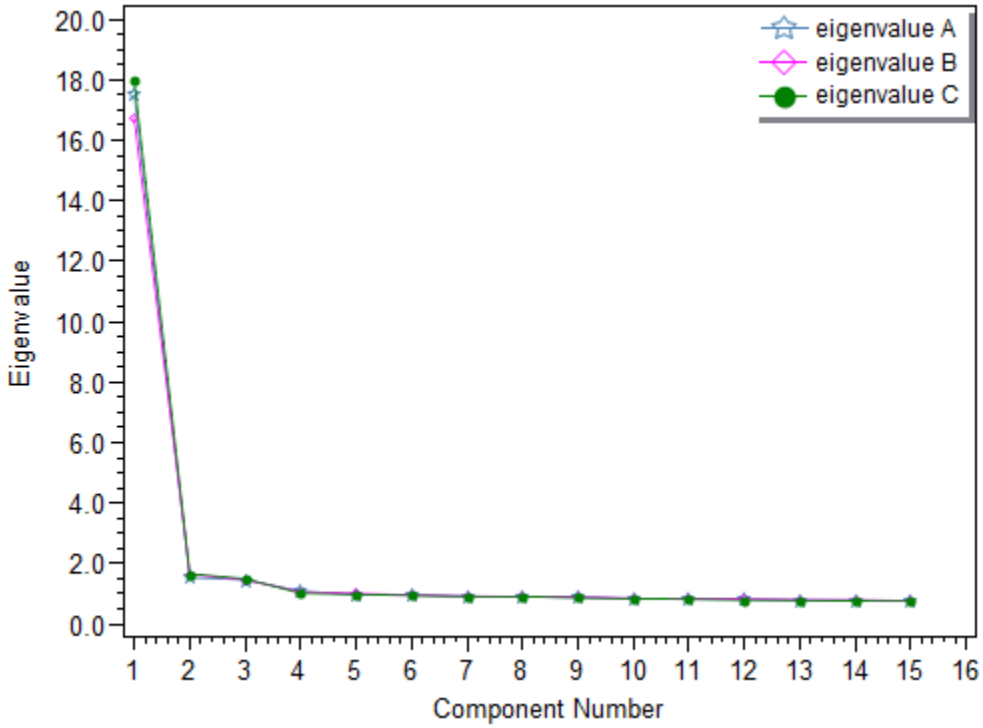
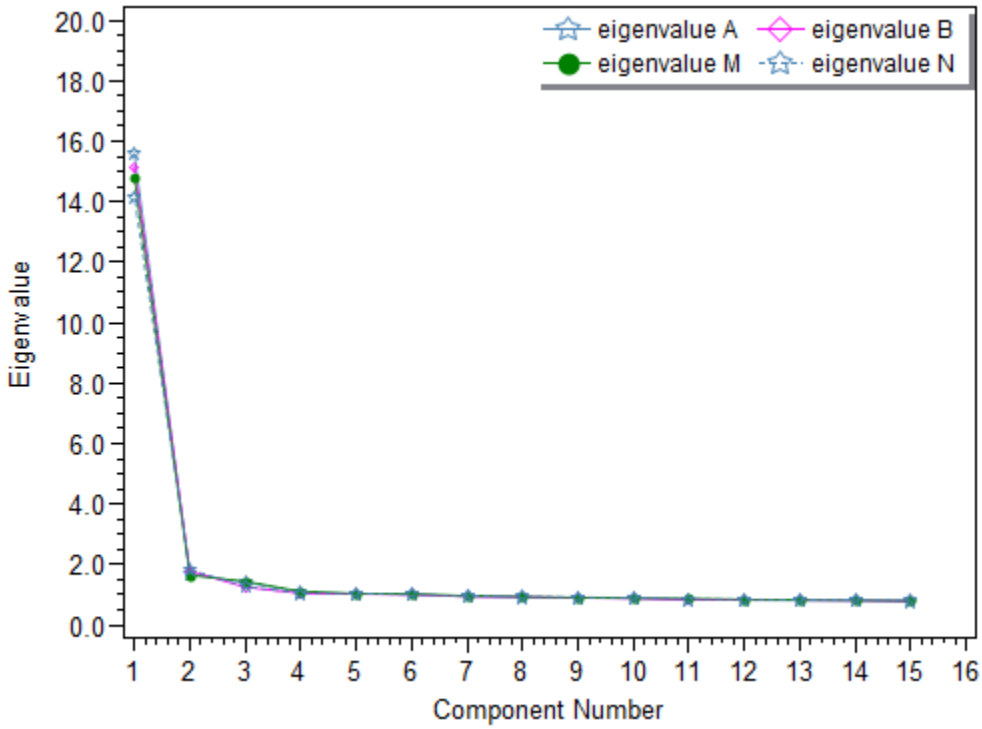


Figure 10.7 Math I Scree Plot of Operational Forms



10.5 Alignment Study

In September, 2014 the North Carolina Department of Public Instruction commissioned the Wisconsin Center for Education Research to conduct an in-depth study of the alignment of the state's newly developed assessments for mathematics, reading, and science to new standards as part of a larger effort to make a systemic examination of the state's standards-based reform efforts. The current report focuses explicitly on the relationship between new *assessments* and their respective *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 *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 Wisconsin Center for Education Research (WCER) is available on the NCDPI website. An abbreviated version of the report with highlighted summaries for reading assessments is documented as part of validity evidence in this section.

10.5.1 Rationale

Standards-based educational reform has been *the* fundamental education model employed by states, and to a growing extent federal policymakers 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 approaches alone can address whether the assumptions of the second 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 approach to examining alignment issues that is objective, systematic, low-inference, and quantifiable. 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 North Carolina's Department of Public Instruction 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 mathematics and reading in grades 3–8, as well as one high school end-of-course exam in each content area administered by the state. 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 its standards-based system, and it will have a wealth of information for 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. As with most relationships, the answers to questions about alignment aren't ever as simple 'yes' or 'no', but rather they always contain a matter of degree. Relationships also tend to be multi-dimensional; they have more than a single aspect, dimension, or quality that is important for one to fully understand the nature of the alignment relationship. All of these factors make alignment analyses a challenging activity.

Alignment measures in SEC are derived from content descriptions. That is, alignment analyses report on the relationship between two multi-dimensional 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 allowing examination of each dimension independently, the following 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 appear in more detail in Section III. 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

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. As with most relationships, the answers to questions about alignment aren't ever as simple 'yes' or 'no', but rather they always contain a matter of degree. Relationships also tend to be multi-dimensional; they have more than a single aspect, dimension, or quality that is important for one to fully understand the nature of the alignment relationship. All of these factors make alignment analyses a challenging activity.

Alignment measures in SEC are derived from content descriptions. That is, alignment analyses report on the relationship between two multi-dimensional 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 allowing examination of each dimension independently, the following 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 appear in more detail in Section III. 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.

10.5.4 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) Topical Coverage (TC); (2) performance expectations (PE); and (3) balance of representation (BR).

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 on 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. For reasons discussed further below, a threshold measure is set at 0.5 for each of the four summary indicator measures. Above the threshold alignment is considered to be at an acceptable level, and below is considered weak or questionable, indicating that a more detailed examination related to that indicator measure is warranted. Much like the results for medical tests, results that fall outside the range of "normal limits" indicate that further investigation is warranted, but does not necessarily mean that the patient is in ill-health, or that a given assessment is not appropriately aligned. It means more information is needed.

10.5.5 Content Analysis Workshop

Content descriptions used to generate visual displays like *Figure 10.8* 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 they were 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 who was familiar with the process and able to facilitate the conversations among team members. The process involves both independent analysis and group discussion, though group consensus is 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. Simple visual comparison of two content maps are 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 *Figure 10.11* suggest that, while distinctions can be identified, both have a generally similar structure which suggests reasonably good alignment of the two descriptions.

10.5.6 Balance of Representation

Of 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.

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.

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 Balance of Representation Index by Grade

Grade	**EOG 3**	**EOG 4**	**EOG 5**	**EOG 6**	**EOG 7**	**EOG 8**	**Math I**
BR	0.57	0.81	0.78	0.87	0.84	0.81	0.69

Table 10.4 displays BR index by grade for the NC End-of-Grade assessments for grades 3-8 and the End-of-Course Math I assessments. Without exception, all of the summary measures on BR for the assessed grades exceed 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. The 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.

10.5.7 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 sub-topic 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 Topic Coverage Index by Grade

Grade	**EOG 3**	**EOG 4**	**EOG 5**	**EOG 6**	**EOG 7**	**EOG 8**	**Math I**
TC	0.68	0.67	0.64	0.73	0.72	0.74	0.61

Table 10.5 provides the summary alignment results for TC for each of the assessed grades in mathematics analyzed for this study. Once again the summary measures for this dimension also indicate above-threshold alignment results, suggesting that the assessments are well aligned to the standards with respect to topic coverage.

10.5.8 Performance Expectations

The SEC taxonomies enable descriptions of academic content based on two dimensions ubiquitous to the field of learning: knowledge and skills. Standards are frequently summarized with the statement "what students should know and be able to do." The "what students should know" part refers to topics, while "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 these taxonomies 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 Performance Expectations Index by Grade

Grade	**EOG 3**	**EOG 4**	**EOG 5**	**EOG 6**	**EOG 7**	**EOG 8**	**Math I**
PE	0.41	0.72	0.70	0.63	0.58	0.77	0.83

As can be seen from *Table 10.6* all but EOG grade 3 math surpass this threshold. The results for grade 3 mathematics indicate weak alignment, but based on assessment design, decisions may nonetheless represent an acceptable degree of alignment. Fine-grain analyses will provide more diagnostic results to indicate particular areas of weak alignment that explain the relatively low alignment.

10.5.9 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 combines all three dimensions into an index measure that is sensitive to the dynamic interplay of all three dimensions. This is done by comparing content descriptions at the intersection of all three dimensions. Overall alignment results are summarized in *Table 10.7*. *Figure 10.8* through *Figure 10.14* show content maps used in displaying visually informative descriptions of the academic content embedded in assessment and standards documents by grade.

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. Once again we see grade 3 Mathematics indicating weak alignment, as well as slightly below-threshold results for grade 7 Mathematics. The PE measures for both grade 3 and 7 are noticeable lower than TC and BR, again suggesting that any alignment issues related to these assessments will likely center around performance expectations.

Table 10.7 Overall Alignment Index by Grade

Grade	**EOG 3**	**EOG 4**	**EOG 5**	**EOG 6**	**EOG 7**	**EOG 8**	**Math I**
OAI	0.40	0.59	0.54	0.55	0.46	0.64	0.57

Table 10.8 Overall Alignment Index for Grades 3 and 7

Grade	**OAI**	**BR**	**TC**	**PE**
Grade 3 Math	0.40	0.94	0.67	0.41
Grade 7 Math	0.46	0.76	0.72	0.58

Table 10.8 reports all four indicators for grades 3 and 7 math. Based on those results, it appears that in each case alignment issues mostly concern performance expectations. Grade 7 math appears more borderline insofar as each of the sub-measures are above 0.5, but the PE measure for both is noticeable lower than TC and BR, again suggesting that any alignment issues related to these assessments will likely center around performance expectations.

The content description maps displayed in *Figure 10.8* through *Figure 10.14* are projected along three axes or dimensions: the Y-axis, represented by the list of 16 mathematics topic areas 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.

The map to the right in *Figure 10.8* indicates that the topics with the strongest emphasis in North Carolina's grade 3 math standards ("Target Content Areas") are Measurement, Operations and, Number sense, and the performance expectation for these topics are Procedures and Demonstrate (equivalent to DOK levels 2 and 3). A careful visual comparison with the content map for grade 3 forms (left map) in terms of the three alignment dimensions indicates the following:

- Balance of Representation (BR): The two figures are shaped similarly which indicates a very good balance of representation for EOG grade 3 assessments. This is also confirm by a BR index of 0.94 see *Table 10.8*

- Topic Coverage (TC): Topics with the strongest emphasis on both maps are Measurement and Operations. This indicates the assessment blueprint is aligned to the content standards with respect to TC. The TC index for EOG grade 3 is 0.67 above the threshold of 0.50.

- Performance Expectation (PE): PE focuses on what students should "*be able to do*" more generally summarized by DOK levels. From the grade 3 assessment map (left) the two strongest topics of emphasis are mostly assessed with recall and explain type items (DOK levels 1 and 2). Whereas, the expectation of the standards focus on Procedures and Demonstrate (DOK 2 and 3). Analysis from the content map suggest that the weak alignment in grade 3 and 7 EOG is likely centered on performance expectations. The analyses results indicated that the grade 7 and especially the grade 3 assessments would benefit from a shift toward more evidence of conceptual understanding of mathematical ideas and less focus on computational proficiency.

Figure 10.8 EOG Grade 3 Assessment and Standard content map

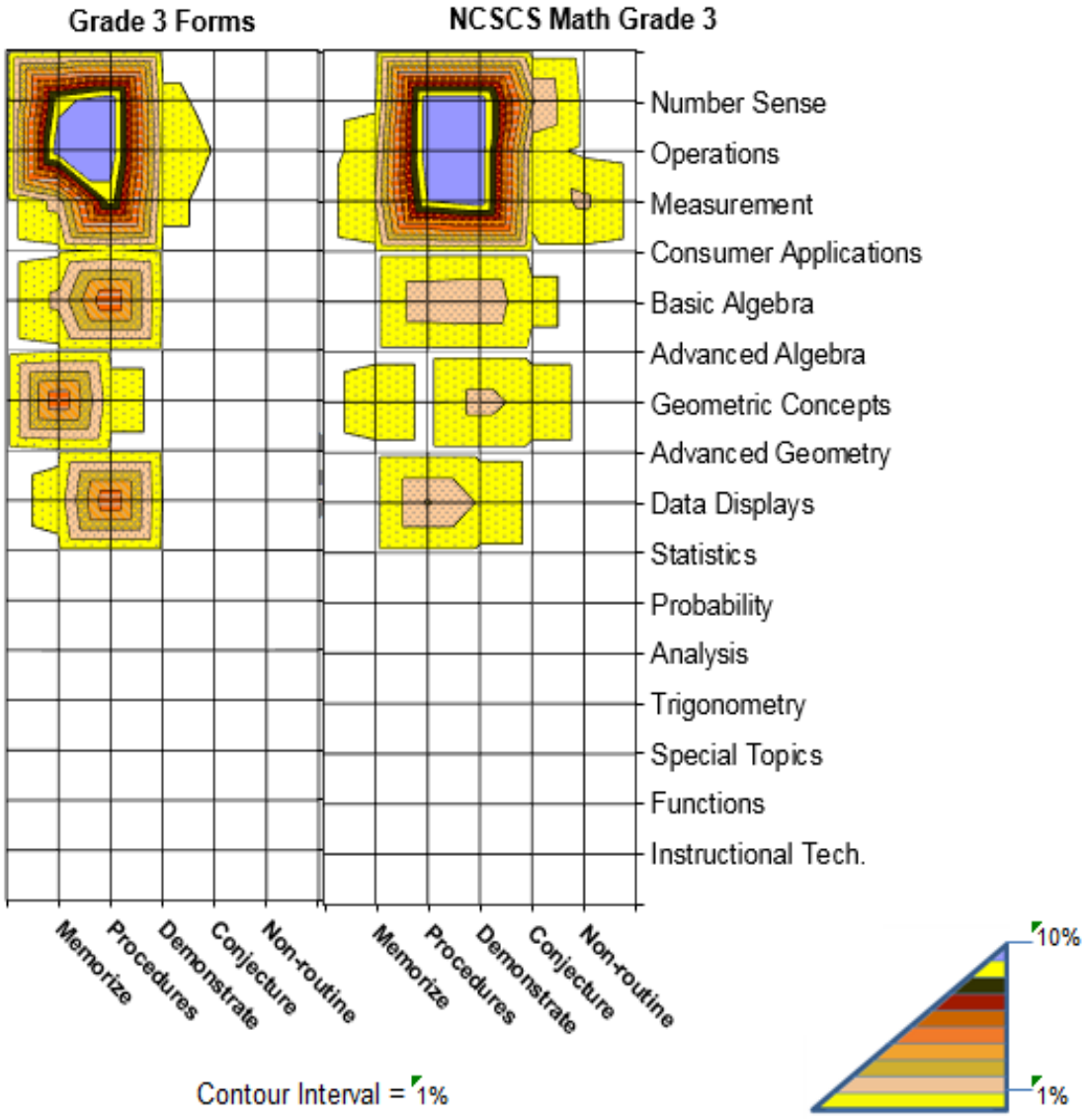


Figure 10.9 EOG Grade 4 Assessment and Standard content map

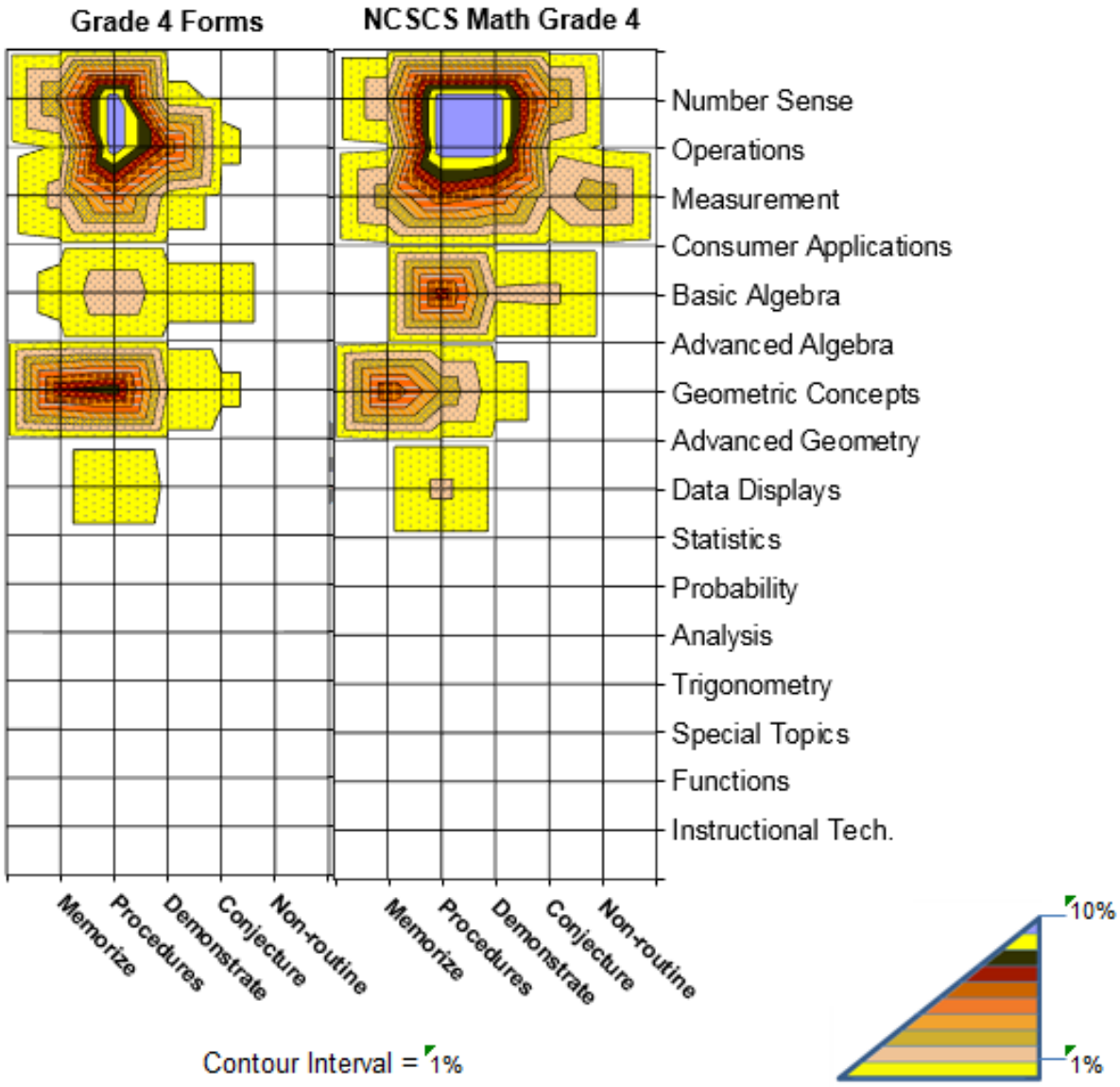


Figure 10.10 EOG Grade 5 Assessment and Standard content map

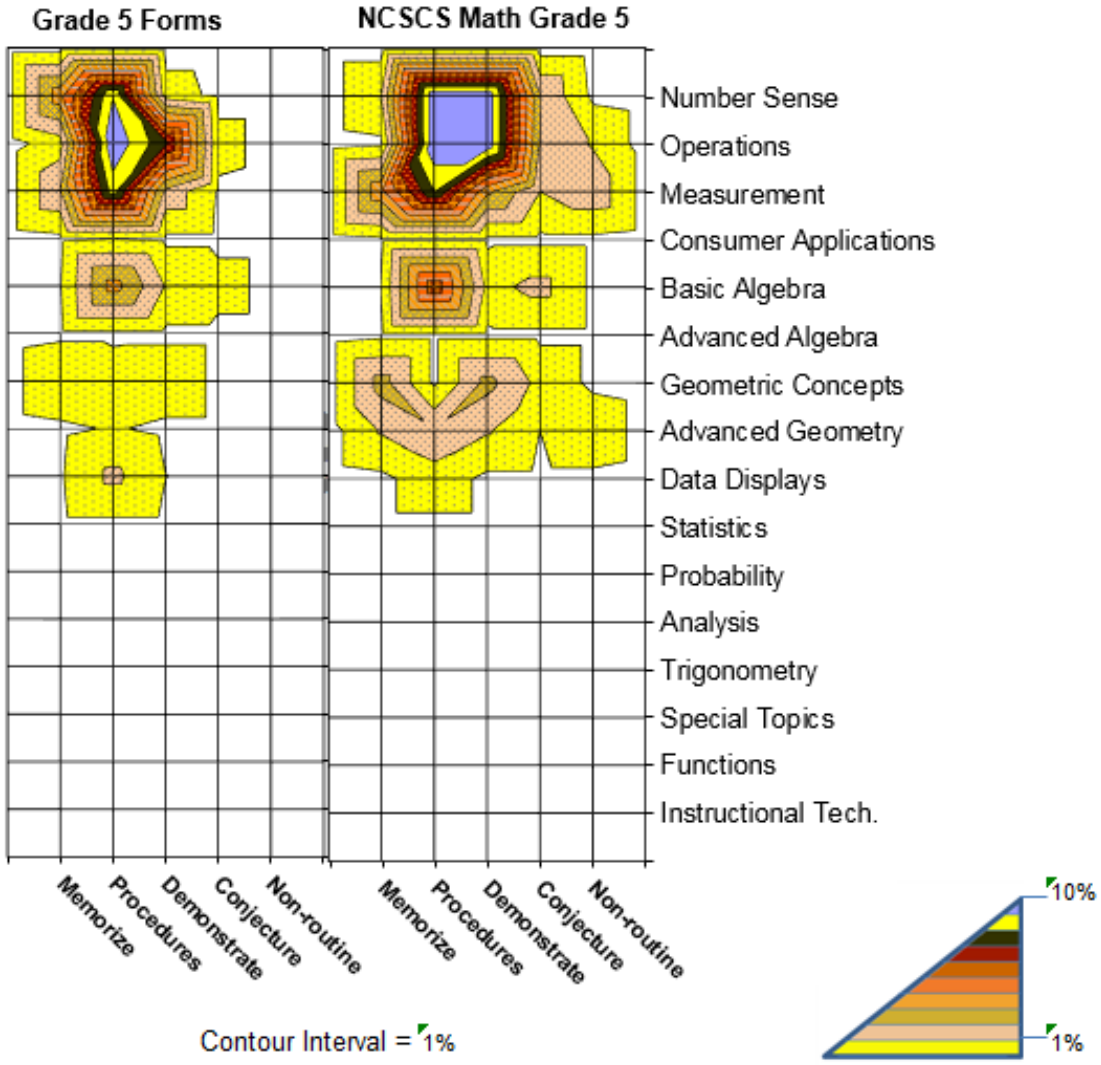


Figure 10.11 EOG Grade 6 Assessment and Standard content map

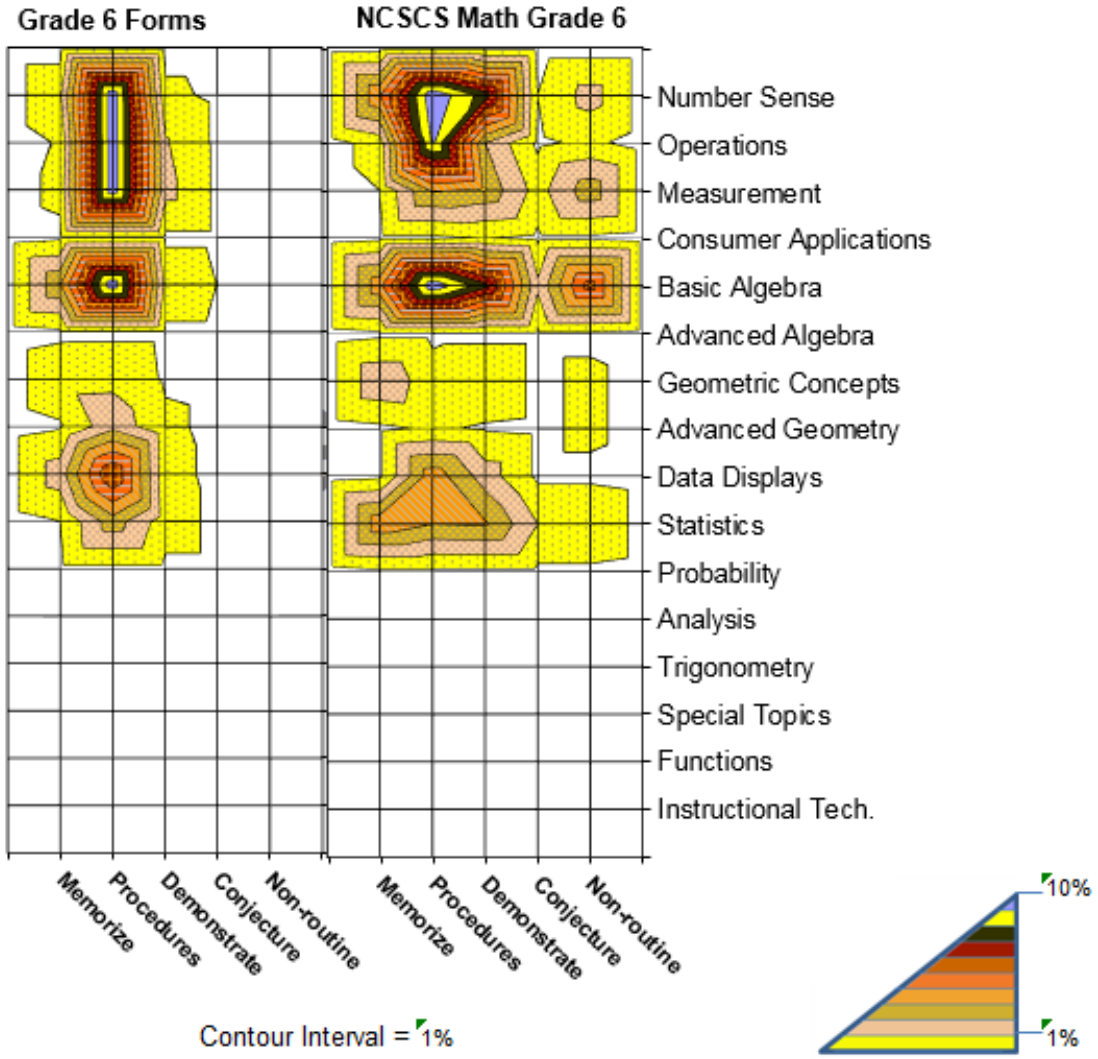


Figure 10.12 EOG Grade 7 Assessment and Standard content map

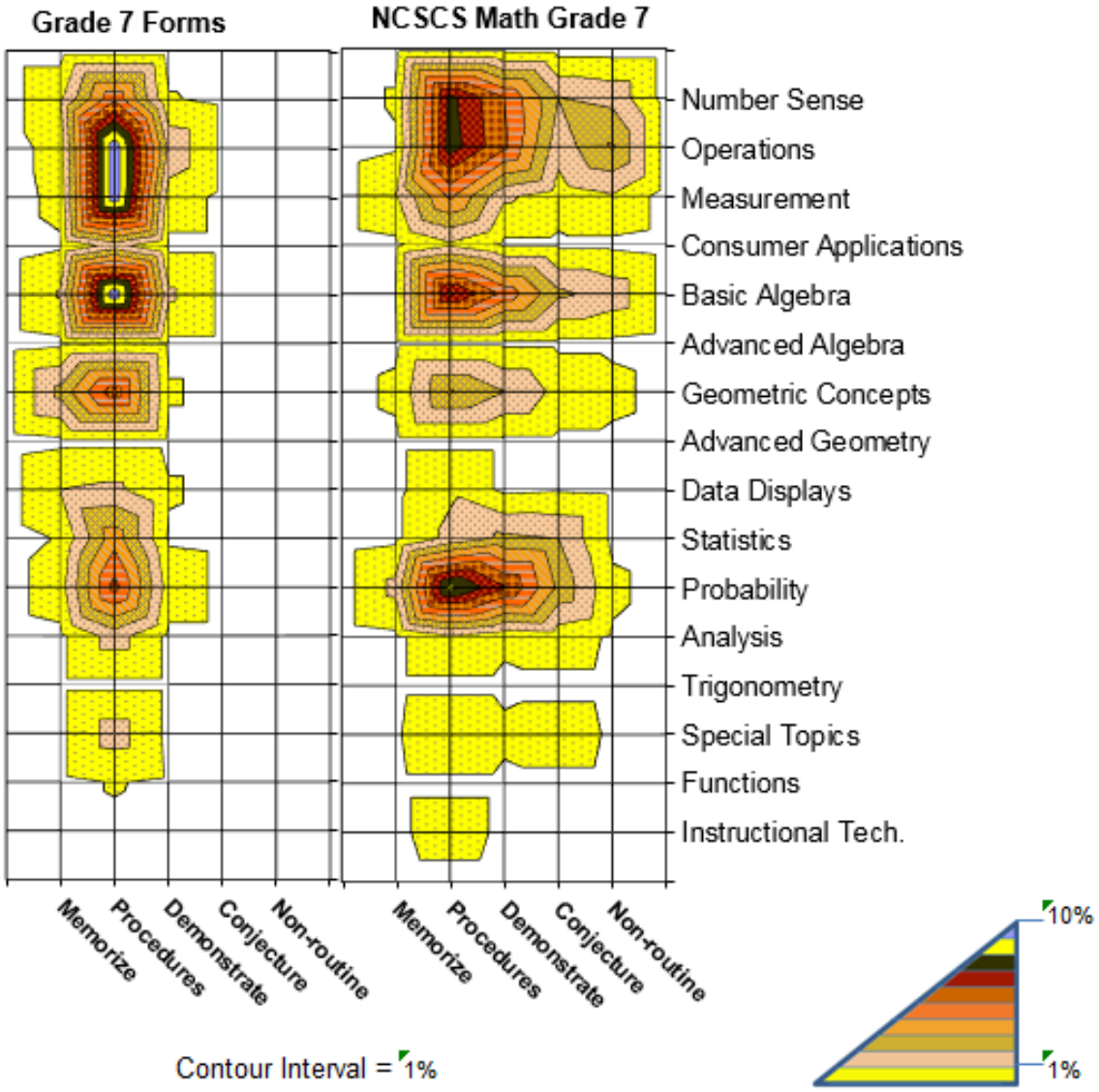


Figure 10.13 EOG Grade 8 Assessment and Standard content map

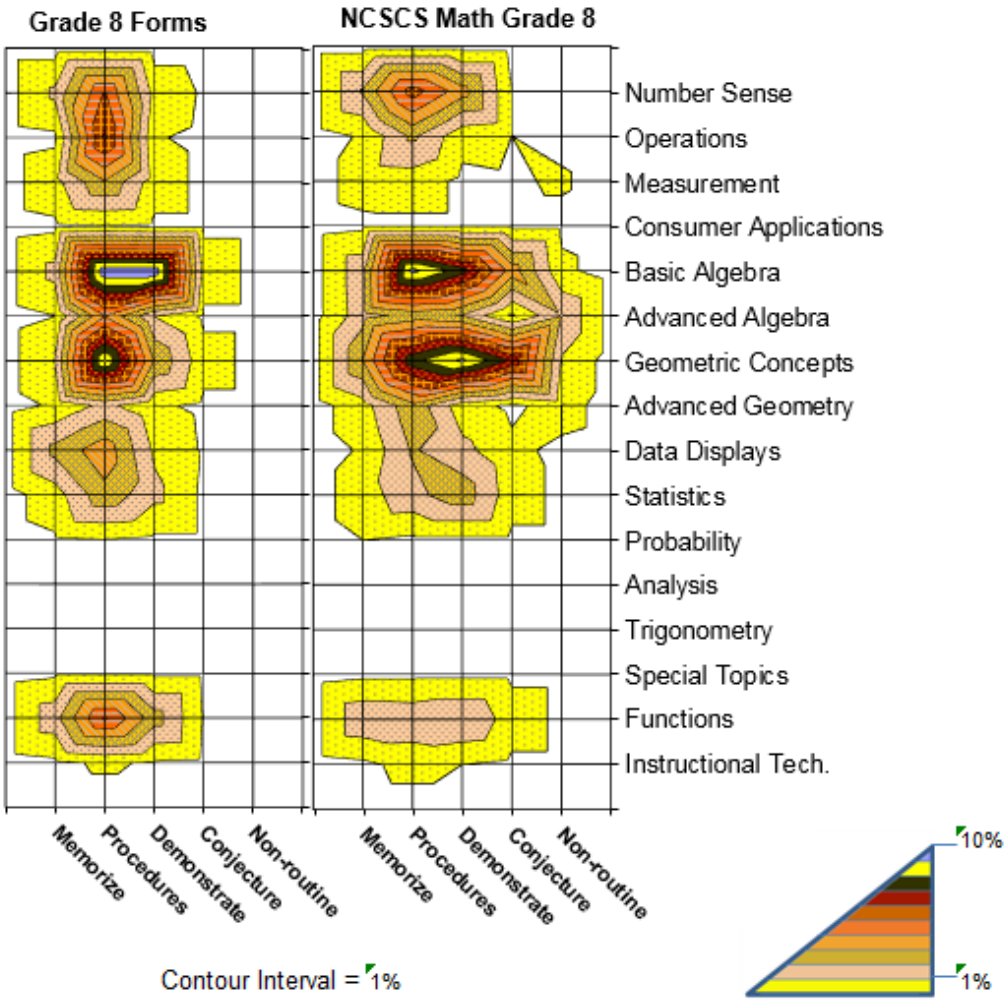
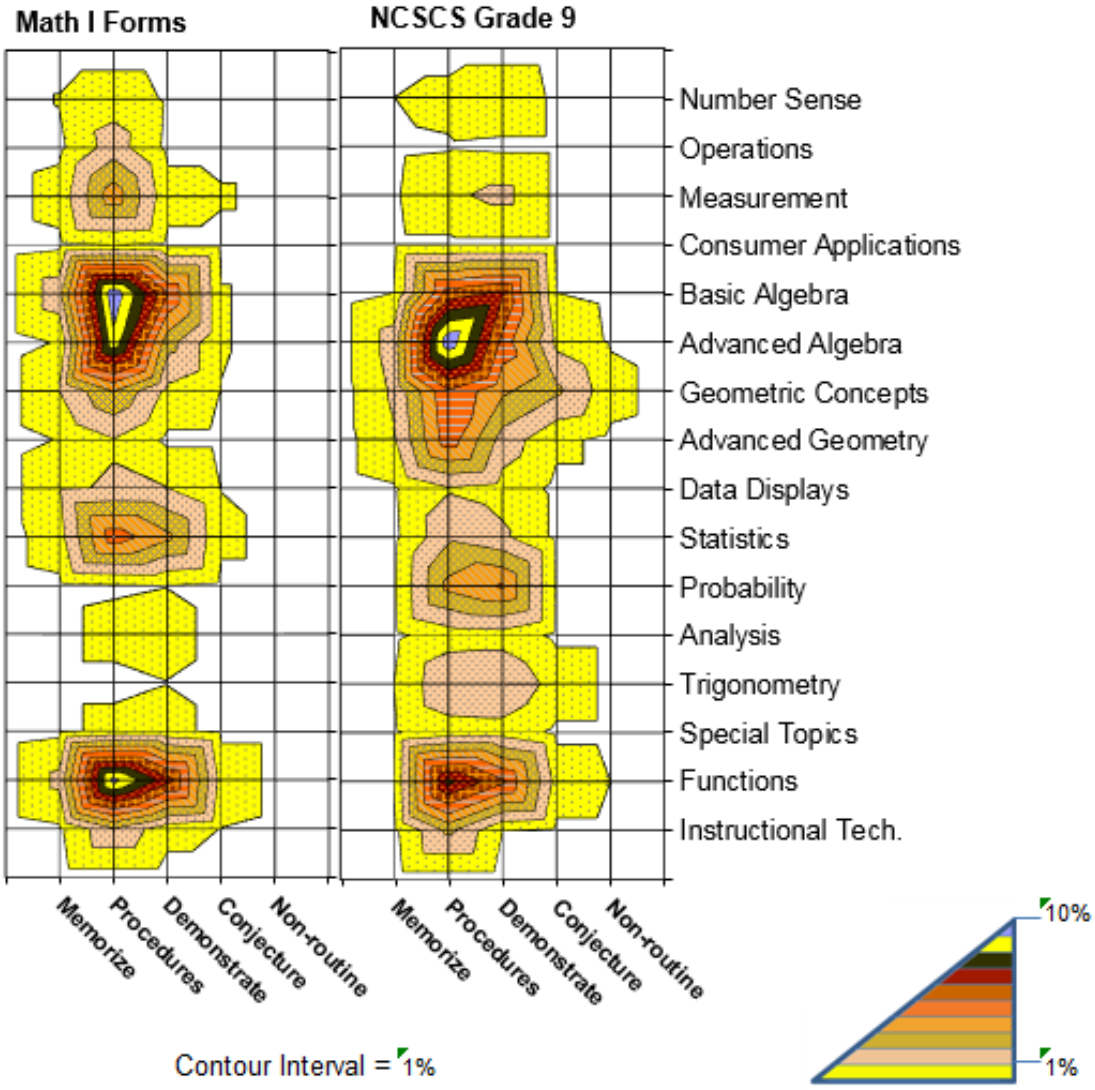


Figure 10.14 EOC Math I Assessment and Standard content map



10.5.10 Discussion of Findings

As indicated by the results presented above, with the exception of grades 3 and 7 math, the EOG and EOC state assessments in mathematics show strong levels of alignment. The results make clear that the design of the assessments attends 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 a 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 (discussed in more detail below). Moreover, while the threshold measure provides a convenient benchmark against which to compare results, it is, at the end of the day, a measure selected by convention, and the reader would be well-advised to use these measures as indicators of alignment that must be considered within the real-world contexts of assessment validity and economic feasibility.

In mathematics, all assessments were held to the full span of mathematics content, regardless of whether a particular content area was actually targeted as part of the assessment program for a given grade level. This sets a more challenging alignment criterion for the grade-specific mathematics assessments. Nonetheless, in only three of twenty-one instances did the indicator results dip below the 0.50 threshold. Relatively weak alignment measures are noted for the grades 3 and 7 overall alignment indices (OAI), the most sensitive and demanding of the alignment indicators, as well as the performance expectation (PE) indicator for grade 3. All other indicators for mathematics at all other grades exceeded the 0.50 measure.

Fine-grain results summarized using content maps presented in *Figure 10.8* through *Figure 10.14*, indicate weak alignment in grades 3 and 7 are related to the PE targets for mathematics topics covered in the assessments. Fine grain results indicate that alignment would be improved with a shift in performance expectations from memorize to procedures and demonstrate.

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 the 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 to be provided in Phase II along with results of student performance on the assessment examined here 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 others with valuable information regarding the curriculum and assessment strategies employed in classrooms around the state and their impact on student learning.

Conclusion

This study collected and examined a comprehensive set of content descriptions covering the full span of the assessment instruments for mathematics in grades 3 through 8, as well as one end of course assessment for high school Math I. The resulting content descriptions provide a unique set of visual displays depicting assessed content and provide the NC Department of Public Instruction a rich descriptive resource for reviewing and reflecting upon the assessment program being implemented throughout the state.

Alignment analyses indicated that the mathematics assessments administered by the state are for the most part very well aligned. Marginally low alignment measures were noted for grades 3 and 7.

10.6 Evidence Regarding Relationships with External Variables

One of the primary intended uses of the EOG and EOC math assessments is to provide data to measure students' achievement and progress relative to readiness as defined by College-and Career-Readiness standards. For the math assessments to provide evidence of this type of achievement, it is important to appropriately match students with materials at a level where the student has the background knowledge necessary to be ready for instruction on the new mathematical skills and concepts. To examine the mathematics achievement levels that can be matched with math skills and concepts based on the NC READY EOG math/EOC Math I assessments, NCDPI commissioned MetaMetrics, Inc. to examine the relationship of the math assessments to the Quantile Framework for math (Contract No. NC10025818 dated December 17, 2012).

The primary purpose of this study was to provide tools (Math@Home, Quantile Teacher Assistant, and Math Skills Database) and information that can be used to answer questions related to standards, student-level accountability, test score interpretation, and test validation; to create conversion tables for determining Quantile measures from the scores on the NC READY EOG Mathematics/EOC Math I assessments; and to produce a report that describes the linking analysis procedures. This section summarizes important evidence from the report. The full report may be found in Appendix 10-A Quantile Linking Technical Report 2014.

10.6.1 The Quantile Framework for Mathematics

The Quantile Framework was developed to assist teachers, parents, and students in identifying strengths and weaknesses in mathematics and forecast growth in overall mathematical achievement. Items and mathematical content are calibrated using the Rasch IRT model. The Quantile scale ranges from "EM" (Emerging Mathematician, 0Q and below) to above 1600Q. The Quantile Framework was developed to assess how well a student (1) understands the natural language of mathematics, (2) knows how to read mathematical expressions and employ algorithms to solve decontextualized problems, and (3) knows why conceptual and procedural knowledge is important and how and when

to apply it. The Quantile Framework Item Bank consists of multiple-choice items aligned with first grade content through Geometry, Algebra II, and Pre-calculus content and was field tested with a national sample of students during the winter of 2004.

For the Quantile Framework, which measures student understanding of mathematical skills and concepts, the most important aspect of validity that should be examined is construct validity. MetaMetrics, Inc. has collected a good amount of validity evidence to show how well Quantile measures relate to other measures of mathematics: (1) standardization set of items used with PASeries Mathematic, (2) relationship of Quantile Measures to other Measures of Mathematical Ability, (3) quantile Framework Linked to other Measures of Mathematics Understanding, and (4) multidimensionality of the Quantile Framework.

10.6.2 Linking the Quantile Framework to the NC Assessments

The Quantile Linking Test was constructed by aligning the items from the NC READY EOG and EOC Mathematics assessments for grades 3, 4, 6, and 8, and Math I with the Quantile Framework taxonomy of Quantile Skills and Concepts (QSCs). Based upon these target test reviews, previously tested items were used to develop each grade-level linking test. Each Quantile Linking Test reflects comparable material that is tested at each identified grade level of the NC READY EOG Mathematics. The comparability of the material, includes the number of operational items per test, the distribution of the content strands (which are closely matched to the distribution of the domains from the North Carolina Core Standards), and the difficulty of the items. The linking study was conducted using linear equating. Separate linking functions were developed for each grade since they are not on a vertical scale. Because the original design for the NC READY mathematics assessments was to report results using a vertical scale across grades, no Quantile data were collected for Grades 5 and 7. During the calibration of the NC READY mathematics items for Grades 3 through 8 it was determined that a vertical scale could not be fitted. Consequently, the Quantile measure equations needed to be estimated for these two grades. Details of the linking are provided in the full report (see Appendix 10-A).

Table 10.9 presents the achievement level cut scores on the NC READY EOG math/EOC Math I assessments and the associated Quantile measures. The North Carolina Department of Public Instruction established four achievement levels: Level 1, Level 2, Level 3, and Level 4 (NCDPI, 2013b) and later revised to five achievement levels for 2014 and beyond (see Chapter 8). The values in the table are the cut scores associated with the bottom score of proficiency levels (3, 4, and 5) for each category.

Table 10.9 NC READY EOG Math/EOC Math I Performance Levels Cut Scores and the Associated Quantile Measures.

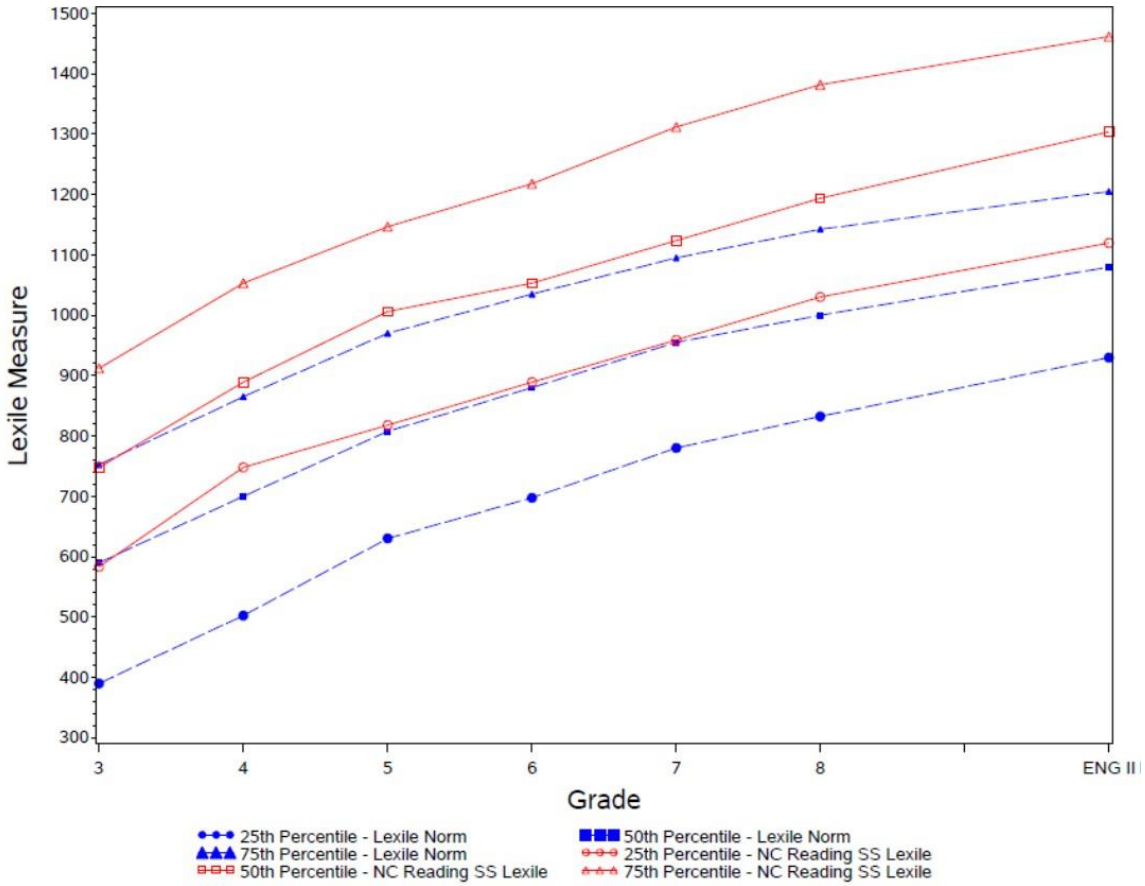
Grade/ Course	Level 3[m]		Level 4		Level 5	
	EOG/EOC Scale Score	Quantile Measure	EOG/EOC Scale Score	Quantile Measure	EOG/EOC Scale Score	Quantile Measure
3	448	610Q	451	680Q	460	885Q
4	449	725Q	451	765Q	460	950Q
5	449	775Q	451	820Q	460	1010Q
6	451	910Q	453	950Q	461	1125Q
7	451	960Q	453	1000Q	461	1165Q
8	452	1095Q	454	1140Q	463	1335Q
Math I	250	1020Q	253	1080Q	264	1310Q

Figure 10.15 shows the Quantile measures for the NC READY EOG and EOC math assessments from the final sample and the Quantile norms. These norms were created based on linking studies conducted with the Quantile Framework. The sample's distribution of scores from this study was similar to the distribution of scores on norm-referenced assessments and other standardized measures of mathematics achievement. The results compared favorably with other mathematics measures which reinforced MetaMetrics' confidence in the Quantile norms. As can be seen in *Figure 10.15*, the Quantile measures for the EOG and EOC math assessments are higher than

[m] Table is different from that presented in original report. This version was updated to reflect the current five achievement level cuts currently used by NCDPI.

the Quantile measure norms. This indicates that the final sample in this study is more able than the samples used for the Quantile norms.

Figure 10.15 Selected Percentiles (25th, 50th, and 75th) plotted for the NC READY EOG Reading/EOC English II Quantile measure against the Quantile measure norms.



10.6.3 The Quantile Framework and College- and Career-Readiness

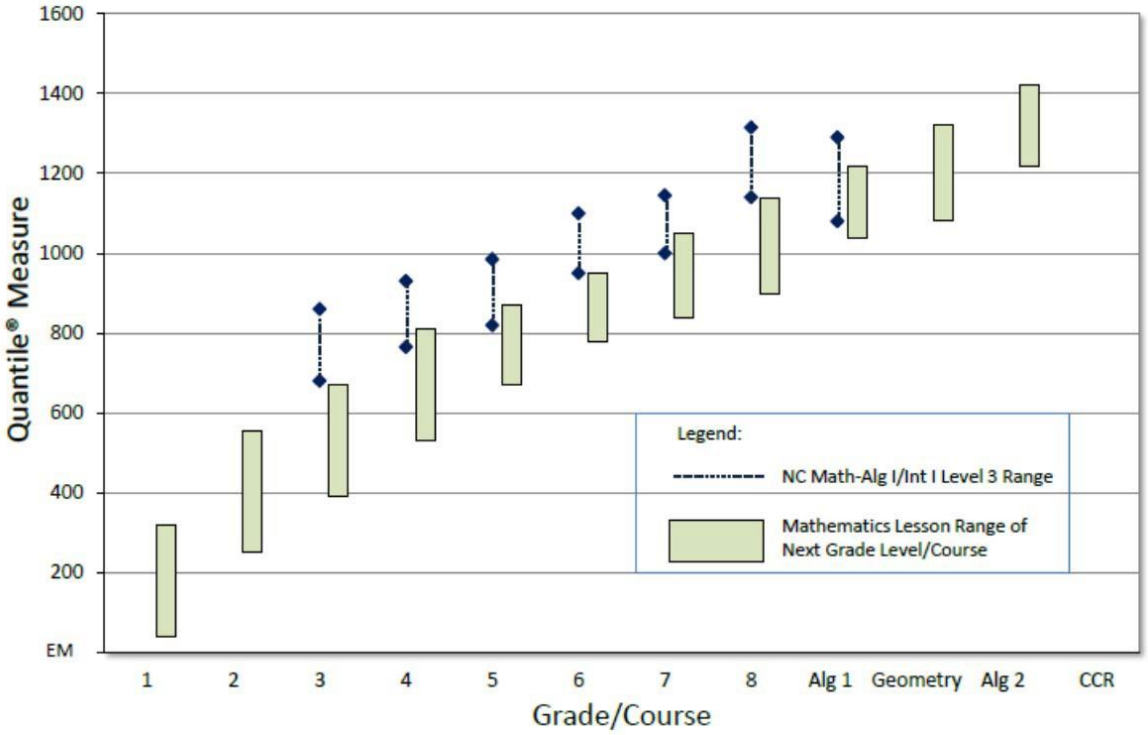
As noted above, one purpose of this study was to examine the mathematics level associated with the NC READY EOG Math/EOC Math I Assessments. If these assessments are to provide information about college- and career-readiness, then the math level of the assessments must be an appropriate measure of college- and career-readiness. It would undermine the credibility of the NC assessments to measure college- and career-readiness if the math levels of the mathematics assessments were, say, below grade level.

If, however, they align to Quantile measures associated with college- and career-readiness, then this is evidence supporting the use of the NC assessments.

> MetaMetrics has calibrated more than 41,000 instructional materials (e.g., textbook lessons and instructional resources) across the K–12 mathematics curriculum (Smith and Turner, 2012) to create a continuum of calibrated textbook lessons from Kindergarten through Pre-calculus. The median of the distribution for Pre-calculus is 1350Q. The range between the first quartile and the median of the first three chapters of Pre-calculus textbooks is from 1200Q to 1350Q. This range describes an initial estimate of the mathematical achievement level needed to be ready for mathematical instruction corresponding to the "college- and career-readiness" standard in the Common Core State Standards for Mathematics.

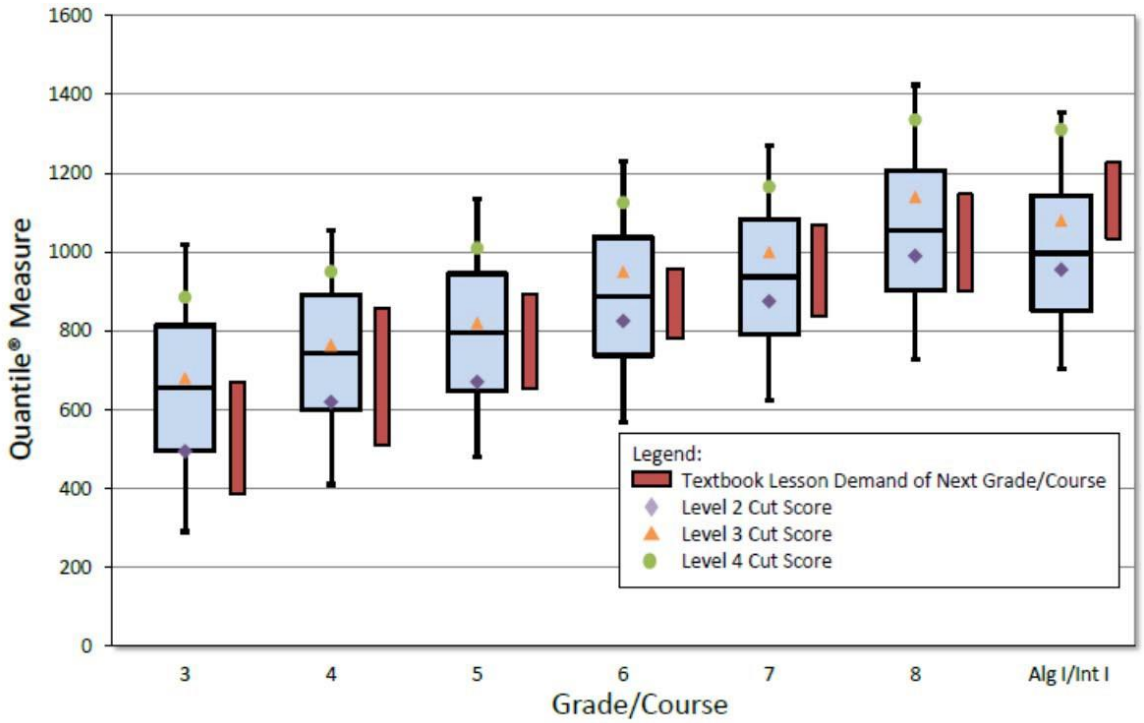
This information describing college- and career-readiness in mathematics can be used to interpret the NC READY EOG Mathematics/EOC Math I performance standards. For each grade the "proficient" (Level 3 or current Level 4) range of Quantile measures as defined by the EOG and EOC math assessments is compared to the mathematical demands in the next grade/course. As can be seen in Figure 10.16 almost all students scoring at the "proficient" level should be prepared for the mathematical demands of the next grade/course. The Math I (Alg I) students at the proficient level are less ready for the next course work.

Figure 10.16 NC READY EOG Mathematics/EOC Math I "proficient" ranges Compared with the mathematical demands of the next grade/course.



shows that the spring 2013 student performance on the EOG and EOC math assessments at each grade/course level is "on track" for college- and career-readiness in Grades 3 through 8. In comparing the performance of students in EOC Math I (Alg I), some students will need encouragement with supplemental materials at the next course. Students can be matched with mathematics materials that are at or above the recommendations in the Common Core State Standards for each grade/course.

Figure 10.17 NC READY EOG Mathematics/EOC Algebra I/Integrated I 2012–2013 student performance expressed as Quantile measures.



In 2009, MetaMetrics and the North Carolina Department of Public Instruction conducted a study to relink the NCEOG/EOC Mathematics Tests with the Quantile scale (MetaMetrics Inc., 2010). The minimum score considered "Proficient" (Level 3 or current Level 4) at each grade level on the NCEOG/EOC math is presented in *Table 10.10*. In 2013, NCDPI transitioned their assessment program to the NC READY EOG and EOC math assessment to align with the Common Core State Standards in Mathematics and to describe student mathematics performance in relation to college- and career-readiness. One outcome of this change was to set the performance standards for NC READY EOG and EOC math at a higher level. The Quantile scale can be used as an external "yardstick" to evaluate this change in the mathematical demands of the North Carolina Mathematics assessments. The information in *Table 10.10* shows that the NC READY EOG/EOC Mathematics standards are demanding more of students in terms of mathematical ability in 2013.

Table 10.10 Minimum "Level 3" Quantile measure on NC EOG/EOC Mathematics (2009) and NC READY EOG Mathematics/EOC Math I (2013).

Grade	"Proficient" Level 3 Cut Score (2009)	"Proficient[n]" Level 3 Cut Score (2013)
3	515Q	680Q
4	645Q	765Q
5	775Q	820Q
6	795Q	950Q
7	860Q	1000Q
8	900Q	1140Q
Math I	1020Q	1080Q

10.6.4 Conclusions

The NC assessments were linked to the Quantile Framework as a means of collecting evidence on the rigor of the NC assessments in relation to the demands of college- and career-readiness standards. This study showed that the math levels of the NC assessments are aligned with expectations of college- and career-readiness as measured by the Quantile Framework. In addition, this study showed that the rigor of math measured by the NC assessments has increased since the previous version of the assessments.

[n] Proficient in 2013 refers to students at Level 3 and 4 on the four Achievement Level scale. This will correspond to students at Levels 4 and 5 on the 5 Achievement Level scale beginning 2014.

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 measures what students know and are able to do as defined in the North Carolina State Content Standards. Assessment must ensure comprehensible access to the content being measured to allow students to accurately demonstrate their standing in the content assessed. In order to ensure items and assessments were developed with universal design principles, NCDPI organized a workshop named "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) has been documented throughout the item development and review, form review, and test administration sections in the report. Some of the universal design principles applied include:

- Precisely defined constructs
    - Direct match to objective being measured
- Accessible, nonbiased items[o]
    - Accommodations included 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

[o] See discussions on bias review in Chapter 4

    - 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
    - Common, everyday words
    - Purposeful graphics to clarify what is being asked
- Accommodations
    - One item per page
    - Extended time for ELL Students
    - Test in a separate room
- Computer-based Forms
    - All students receive one item per test page
    - All students may receive larger font and different background colors.

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 which are also aligned to the NCSCS. Assessments' content domain specifications and blueprints 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 thus provides them with an 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 released forms provided students, teachers, and parents with sample items and a general practice form similar to the operational assessment. These released forms also served as a resource to familiarized students with the various response formats in the new assessments.

10.7.3 Fairness in Administration

Chapter 5 of this report documents the procedures put in place by NCDPI to assure the administration that 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, and testing window and makeup testing options. Assessment guides 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. 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; 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.

In order to prepare students for gridded response items in their upcoming EOG Math grade 5 – 8 and EOC Math I assessments, Accountability Services produced practice activities for using the grids. The North Carolina Department of Public Instruction (NCDPI) requires students take the Gridded Response Practice Activity before the administration of the paper-and-pencil EOG Mathematics Grades 5–8 and EOC Math I tests. Schools must ensure that every student participating in the paper-and-pencil grades 5–8 EOG mathematics assessments complete the grade-appropriate Gridded Response Practice Activity at least one time at the school before test day. Students taking the mathematics test online should also complete the practice activity as part of instruction in the event there is a computer system crash.

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." It is imperative that when multiple forms are created from the same test blueprint, the resulting test scores from parallel forms are 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 section 4.5.3 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 class to obtain equivalent samples for calibration and scaling. This ensured that each form was administered to a random equivalent sample of students 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 located onto the same IRT scale and separate raw-to-scale tables were created to adjust for any form differences.

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, 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 anchors 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 that 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, 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 procedure and 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 on proficiency classifications.

To ensure equitable access for students taking computer-based forms, 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 a review of industry standards and usability studies and research findings conducted with other national testing programs. NCDPI device requirements for EOG and EOC computer-based assessments include:

- A minimum screen size of 9.5 inches
- A 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.
- A Pentium 4 or newer processor for Windows machines and Intel for MacBooks

In addition to the technical specification of devices NCDPI also conducts a review of each sample item across devices i.e. laptops, iPads and desktops, to make sure items are rendered as intended. Reviews also check functionalities of the test platform, such as audio files, large font, and high contrast versions.

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™ 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.

Smith III, M. & Turner, J. (2012). A mathematics problem: How to help students achieve success in mathematics through college and beyond. Durham: MetaMetrics, Inc. Retrieved [January 26, 2016] from https://lexile.com/about-lexile/Policy-Briefs/mathematics-problem/

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.

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

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

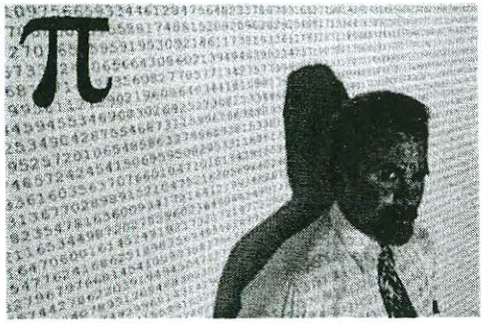
1) 190
2) 200
3) 290
4) N

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

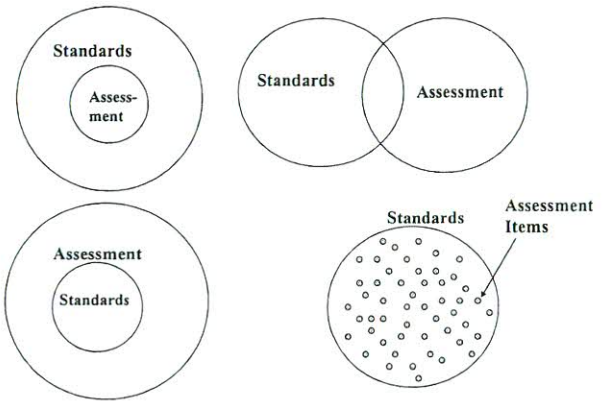
Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 Recall & Reproduction	Level 2 Skills & Concepts	Level 3 Strategic Thinking/ Reasoning	Level 4 Extended Thinking
Remember				
Understand				
Apply				
Analyze				
Evaluate				
Create				

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

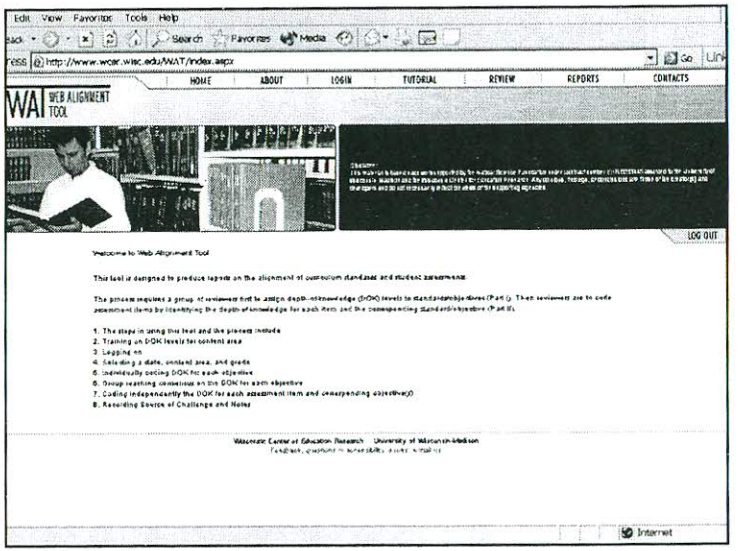
D. Balance of Representation

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



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 to show* $(2/3) \times 4 = 8/3$, *and create a story context for this equation; do the same 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; multiply fractional 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

NC Standard Course of Study (*NCSCS*) for Mathematics

End-of-Grade Grades 3–8 Math Assessments
End-of-Course Math I Assessment

North Carolina Assessment Specifications

Purpose of the Assessments

- Edition 4 grades 3–8 mathematics assessments and the Math I assessment will measure students' proficiency on the NC *Standard Course of Study* (NC*SCS*) for Mathematics, adopted by the North Carolina State Board of Education in June 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

- June 2010: North Carolina State Board of Education adoption of the NC*SCS*
- 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 assessments aligned to the NC*SCS*

Standards

- The NC*SCS* may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net
- North Carolina will teach and assess a common set of standards for the first-year high school course of mathematics, Math I.
- The eight Standards for Mathematical Practice help develop processes and proficiencies in students such as problem solving, reasoning, proof, communication, representations, and connections as well as conceptual understanding and procedural fluency. Test items that are developed for content standards may link to one or more of the Standards for Mathematical Practice.
- The End-of-Course Assessment of Math I is the only high school math EOC assessment available. All high school students are transitioning to Math I, II and III.

Prioritization of Standards

- The North Carolina Department of Public Instruction invited teachers to collaborate and develop recommendations for a prioritization of standards indicating the relative importance of each standard, the anticipated instructional time, and the appropriateness of the standard for a multiple-choice or gridded-response 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 below.
- Some content standards in the NC*SCS* will not be directly assessed in the Edition 4 test because either (1) the standard cannot be appropriately assessed during a limited time assessment using multiple-choice and/or gridded-response items or (2) the standard is better assessed through another, more inclusive standard.

Table 1: Weight Distributions for Grades 3–5

Domain	**Grade 3**	**Grade 4**	**Grade 5**
Operations and Algebraic Thinking	30–35%	12–17%	5–10%
Number and Operations in Base Ten	5–10%	22–27%	22–27%
Number and Operations—Fractions	20–25%	27–32%	47–52%
Measurement and Data	22–27%	12–17%	10–15%
Geometry	10–15%	12–17%	2–7%
Total	**100%**	**100%**	**100%**

Table 2: Weight Distributions for Grades 6–8

Domain	**Grade 6**	**Grade 7**	**Grade 8**
Ratios and Proportional Relationships	12–17%	22–27%	NA
The Number System	27–32%	7–12%	2–7%
Expressions and Equations	27–32%	22–27%	27–32%
Functions	NA	NA	22–27%
Geometry	12–17%	22–27%	20–25%
Statistics and Probability	7–12%	12–17%	15–20%
Total	**100%**	**100%**	**100%**

Table 3: Weight Distributions for Math I

Conceptual Category	**Math I**
Number and Quantity	5–10%
Algebra	25–31%
Functions	35–40%
Geometry	10–15%
Statistics and Probability	15–20%
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 NC*SCS* for Mathematics. These items will require students to not only recall information, but also apply concepts and skills and make decisions.

Types of Items

- Grades 3 and 4 mathematics assessments will consist of four-response-option multiple-choice items. Multiple-choice items will be worth one point each.

- The grades 5–8 mathematics assessments and the Math I assessment will consist of four-response-option multiple-choice items and about twenty percent gridded- response items requiring numerical responses. All items will be worth one point each.

- All NC*SCS* mathematics assessments will include both calculator-active and calculator-inactive sections. One-third to one-half of the grades 3–8 assessments will be comprised of calculator-inactive items; approximately one-third of the high school assessments will be calculator inactive.

- The ***NCEXTEND1*** mathematics alternate assessments will consist of fifteen performance-based, multiple-choice items. All items will be worth one point each.

- Appendices A-G show the number of operational items for each standard administered on the assessments. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*.

Delivery Mode and Translation

- Grades 3–8 mathematics assessments will be designed for paper/pencil administrations. The grade 7 mathematics assessment will be available for online administration effective with the 2014–15 spring administration. The grade 8 mathematics assessment will be available for online administration effective with the 2015–16 spring administration.

- The Math I assessment will be available for online and paper/pencil administrations.

- ***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*** mathematics 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 3 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 3 Math	Number of Operational Items Per Standard*
Operations and Algebraic Thinking 3.OA.1	–
3.OA.2	–
3.OA.3	2
3.OA.4	2
3.OA.5	3
3.OA.6	–
3.OA.7	–
3.OA.8	4
3.OA.9	3
Number and Operations in Base Ten 3.NBT.1	1
3.NBT.2	2
3.NBT.3	1
Number and Operations-Fractions 3.NF.1	3
3.NF.2	4
3.NF.3	3
Measurement and Data 3.MD.1	1
3.MD.2	1
3.MD.3	2
3.MD.4	1
3.MD.5	–
3.MD.6	–
3.MD.7	3
3.MD.8	3
Geometry 3.G.1	2
3.G.2	3

* Some standards 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 4 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 4 Math	Number of Operational Items Per Standard*
Operations and Algebraic Thinking 4.OA.1	–
4.OA.2	3
4.OA.3	2
4.OA.4	1
4.OA.5	1
Number and Operations in Base Ten 4.NBT.1	–
4.NBT.2	2
4.NBT.3	3
4.NBT.4	2
4.NBT.5	2
4.NBT.6	2
Number and Operations-Fractions 4.NF.1	3
4.NF.2	1
4.NF.3	3
4.NF.4	3
4.NF.5	1
4.NF.6	1
4.NF.7	1
Measurement and Data 4.MD.1	2
4.MD.2	1
4.MD.3	1
4.MD.4	1
4.MD.5	–
4.MD.6	1
4.MD.7	1
Geometry 4.G.1	2
4.G.2	2
4.G.3	2

* Some standards 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
Grade 5 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 5 Math	Number of Operational Items Per Standard*
Operations and Algebraic Thinking 5.OA.1	1
5.OA.2	1
5.OA.3	1
Number and Operations in Base Ten 5.NBT.1	–
5.NBT.2	1
5.NBT.3	1
5.NBT.4	1
5.NBT.5	1
5.NBT.6	3
5.NBT.7	4
Number and Operations-Fractions 5.NF.1	3
5.NF.2	4
5.NF.3	3
5.NF.4	5
5.NF.5	–
5.NF.6	3
5.NF.7	4
Measurement and Data 5.MD.1	2
5.MD.2	1
5.MD.3	–
5.MD.4	–
5.MD.5	3
Geometry 5.G.1	–
5.G.2	1
5.G.3	–
5.G.4	1

* Some standards 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 D
Grade 6 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 6 Math	Number of Operational Items Per Standard*
Ratios and Proportional Relationships 6.RP.1	–
6.RP.2	–
6.RP.3	7
The Number System 6.NS.1	3
6.NS.2	–
6.NS.3	5-6
6.NS.4	1
6.NS.5	–
6.NS.6	1
6.NS.7	2-3
6.NS.8	2
Expressions and Equations 6.EE.1	2
6.EE.2	2
6.EE.3	4
6.EE.4	–
6.EE.5	–
6.EE.6	2
6.EE.7	3
6.EE.8	1
6.EE.9	1
Geometry 6.G.1	2
6.G.2	2
6.G.3	2
6.G.4	2
Statistics and Probability 6.SP.1	–
6.SP.2	–
6.SP.3	–
6.SP.4	2
6.SP.5	3

* Some standards 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 E
Grade 7 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 7 Math	Number of Operational Items Per Standard*
Ratios and Proportional Relationships 7.RP.1	3
7.RP.2	5
7.RP.3	5
The Number System 7.NS.1	–
7.NS.2	–
7.NS.3	5
Expressions and Equations 7.EE.1	3
7.EE.2	–
7.EE.3	4
7.EE.4	6
Geometry 7.G.1	2
7.G.2	1
7.G.3	1
7.G.4	3
7.G.5	2
7.G.6	3
Statistics and Probability 7.SP.1	1
7.SP.2	–
7.SP.3	–
7.SP.4	3
7.SP.5	–
7.SP.6	–
7.SP.7	1
7.SP.8	2

* Some standards 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 F
Grade 8 Math
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Grade 8 Math	Number of Operational Items Per Standard*
The Number System 8.NS.1	1
8.NS.2	2
Expressions and Equations 8.EE.1	1
8.EE.2	1
8.EE.3	1
8.EE.4	1
8.EE.5	4
8.EE.6	2
8.EE.7	3
8.EE.8	3
Functions 8.F.1	1
8.F.2	3
8.F.3	2
8.F.4	4
8.F.5	2
Geometry 8.G.1	–
8.G.2	–
8.G.3	2
8.G.4	–
8.G.5	2
8.G.6	–
8.G.7	3
8.G.8	2
8.G.9	2
Statistics and Probability 8.SP.1	2
8.SP.2	3
8.SP.3	2
8.SP.4	1

* Some standards 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 G
Math I
Number of Operational Items by Standard

The following table shows the number of operational items for each standard. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1-3*. Some standards 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 standards may be reviewed by visiting the North Carolina DPI K-12 Mathematics wiki site at http://maccss.ncdpi.wikispaces.net.

Math I	Number of Operational Items Per Standard*
The Real Number System N-RN.1	–
N-RN.2	2
Quantities N-Q.1	1
N-Q.2	–
N-Q.3	–
Seeing Structure in Expressions A-SEE.1	–
A-SEE.2	1
A-SEE.3	0-1
Arithmetic with Polynomials & Rational Expressions A-APR.1	1
Creating Equations A-CED.1	4
A-CED.2	2
A-CED.3	2
A-CED.4	1-2
Reasoning with Equations & Inequalities A-REI.1	–
A-REI.3	–
A-REI.5	–
A-REI.6	1
A-REI.10	–
A-REI.11	1
A-REI.12	1

* Some standards 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 G (continued)
Math I
Number of Operational Items by Standard

Math I	Number of Operational Items Per Standard*
Interpreting Functions F-IF.1	–
F-IF.2	1-2
F-IF.3	–
F-IF.4	1-2
F-IF.5	0-1
F-IF.6	1-2
F-IF.7	1
F-IF.8	2-3
F-IF.9	1
Building Functions F-BF.1	2-3
F-BF.2	0-1
F-BF.3	1
Linear, Quadratic, & Exponential Models F-LE.1	1-2
F-LE.2	1
F-LE.3	1
F-LE.5	1
Congruence G-CO.1	–
Expressing Geometric Properties with Equations G-GPE.4	1
G-GPE.5	1
G-GPE.6	1
G-GPE.7	1
Geometric Measurement & Dimension G-GMD.1	–
G-GMD.3	1
Interpreting Categorical & Quantitative Data S-ID.1	–
S-ID.2	1
S-ID.3	1-2
S-ID.5	2
S-ID.6	1
S-ID.7	1
S-ID.8	1-2
S-ID.9	–

* Some standards 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.

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

>>> 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

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:

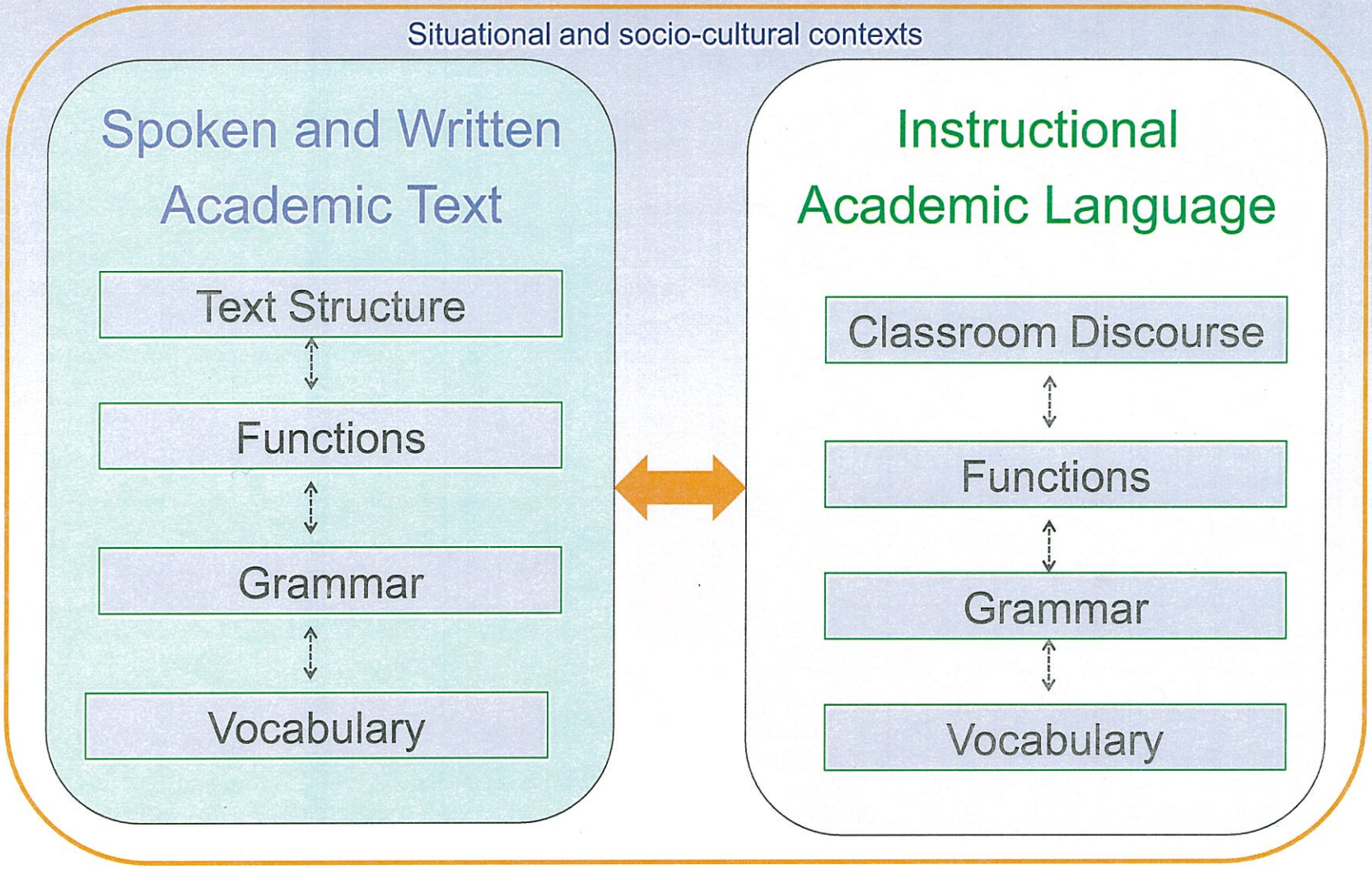
- to help analyze the content and language in standards, assessment tasks, and instructional materials;
- to help make explicit the expectations (cognitive, language) of students;
- to help inform instructional planning and practice so that they are intentional and appropriate in supporting students' 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).

Language for Achievement: Overview



Additional considerations include: receptive (listening, reading) and productive (speaking, writing) language; language complexity

Contact: Dr. Edynn Sato at esato@wested.org

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

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

NCEE 2009-4079
U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

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?

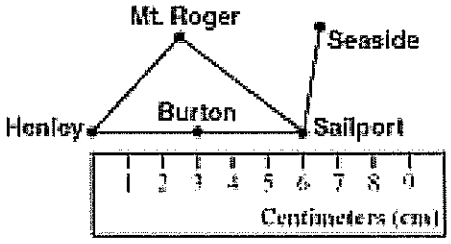
A. 2 days

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?

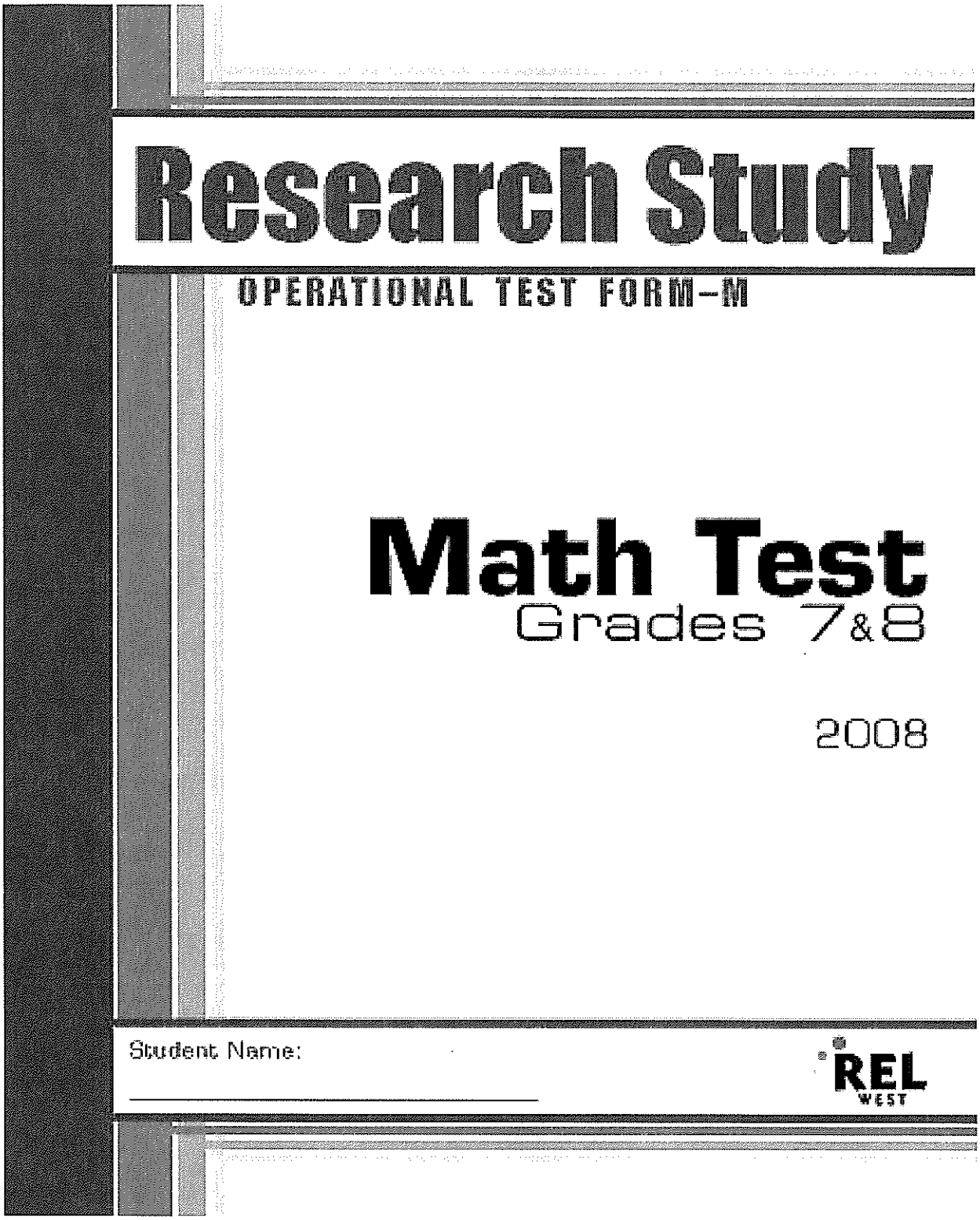
A. 1 cm = 6 km

B. 1 cm = 12 km

C. 1 cm = 15 km

D. 1 cm = 20 km

Appendix O. Operational item set—linguistically modified



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. $|-16|$

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?

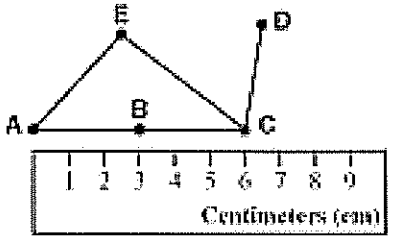
A. 2 days

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:

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?

A 3

B 10

C 12

D 25

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?

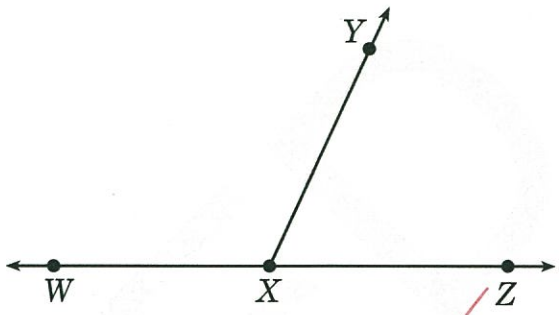
A $\frac{1}{12}$

B $\frac{1}{6}$

C $\frac{6}{12}$

D $\frac{5}{6}$

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?

A 11 angles, 8 right angles

B 11 angles, 9 right angles

C 12 angles, 8 right angles

D 12 angles, 9 right angles

19. Cara used this multiplication table to help her find the quotient for $112 \div 14$.

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

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

21. Sallie baked 4 apple pies and cut each of them into sixths. If she served $3\frac{1}{2}$ pies, 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?

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 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 $45 \times 3 \times 5$

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

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?

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

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?

A 6

B 18

C 62

D 72

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

5. Dean had 1,062 pennies in his bank. Shawn had 889. How many more pennies did Dean have than Shawn?

A 173

B 223

C 227

D 283

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?

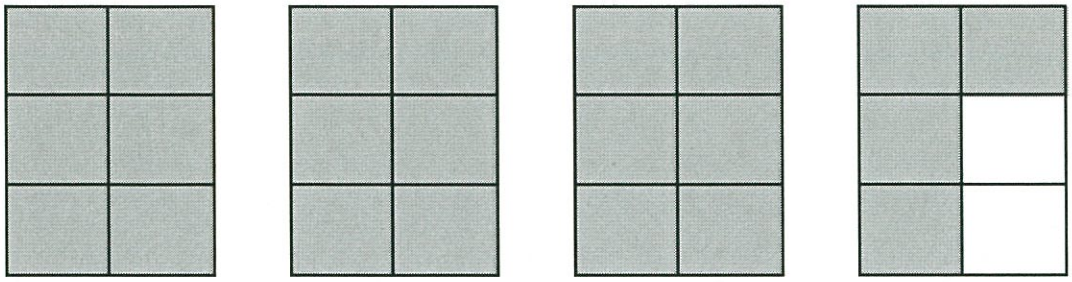
A between 50 and 60 pounds

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 $3\frac{4}{6}$

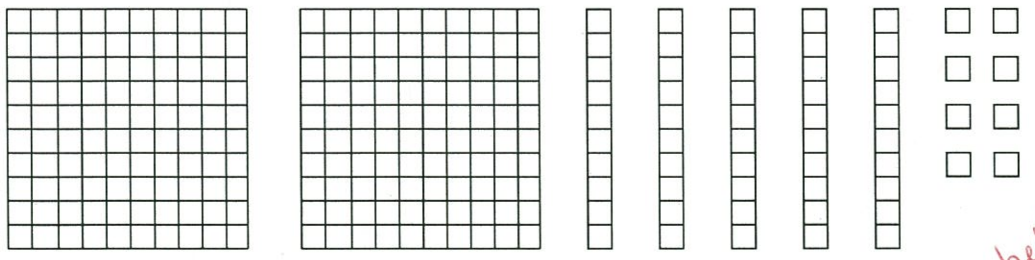
C $4\frac{2}{6}$

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

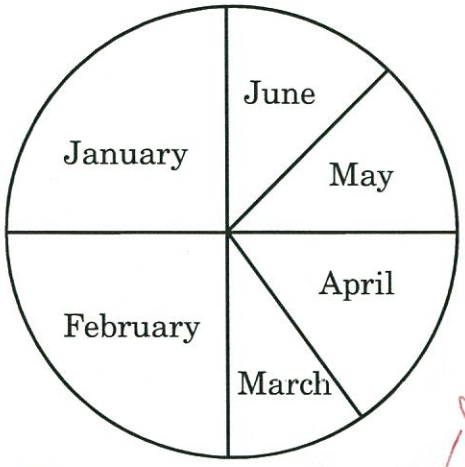
C 358

D 385

Is there better placement for the stem?

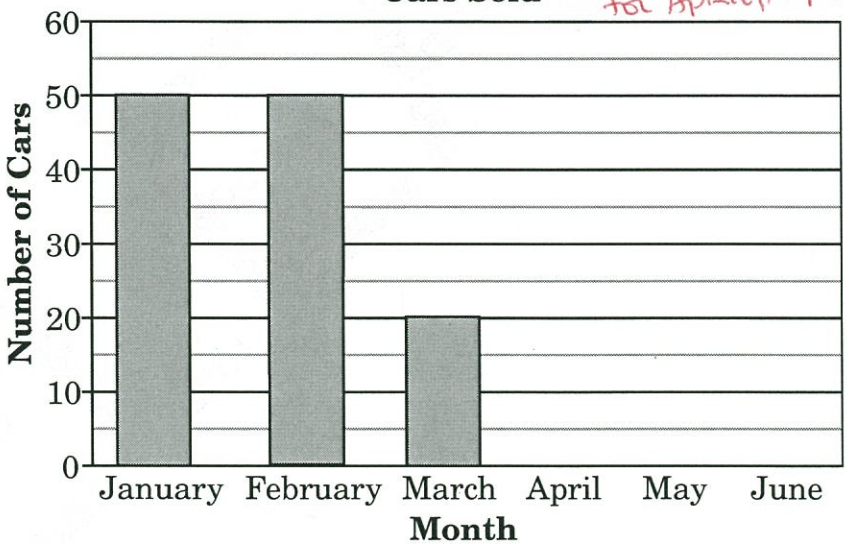
30. A dealership sold 200 cars in a six-month period. The circle graph below displays the distribution of sales by month.

Distribution of Car Sales



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.

Cars Sold



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).

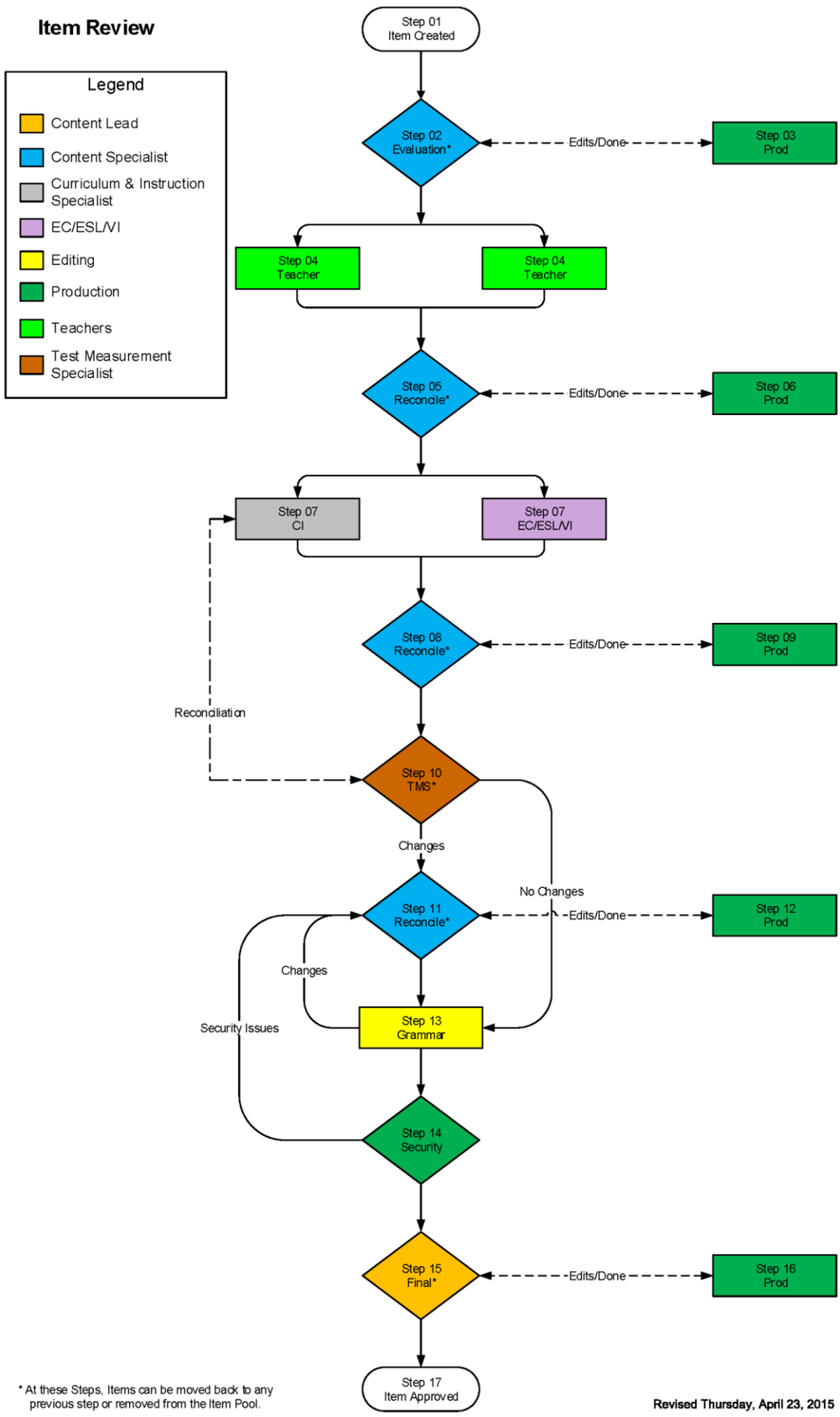
Step 14: Security Check
Production staff checks to make sure no duplicate copy of the item exists in the test development databases. If there is a duplicate copy of the item or a requested revision was not made, then the item is flagged and sent back to **Step 11**.

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
Legend
Content Lead
Content Specialist
Curriculum & Instruction Specialist
EC/ESL/VI
Editing
Production
Teachers
Test Measurement Specialist
Step 01 Item Created
Step 02 Evaluation*
Edits/Done
Step 03 Prod
Step 04 Teacher
Step 04 Teacher
Step 05 Reconcile*
Edits/Done
Step 06 Prod
Step 07 CI
Step 07 EC/ESL/VI
Step 08 Reconcile*
Edits/Done
Step 09 Prod
Reconciliation
Step 10 TMS*
Changes
No Changes
Step 11 Reconcile*
Edits/Done
Step 12 Prod
Changes
Security Issues
Step 13 Grammar
Step 14 Security
Step 15 Final*
Edits/Done
Step 16 Prod
Step 17 Item Approved
* 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.

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.

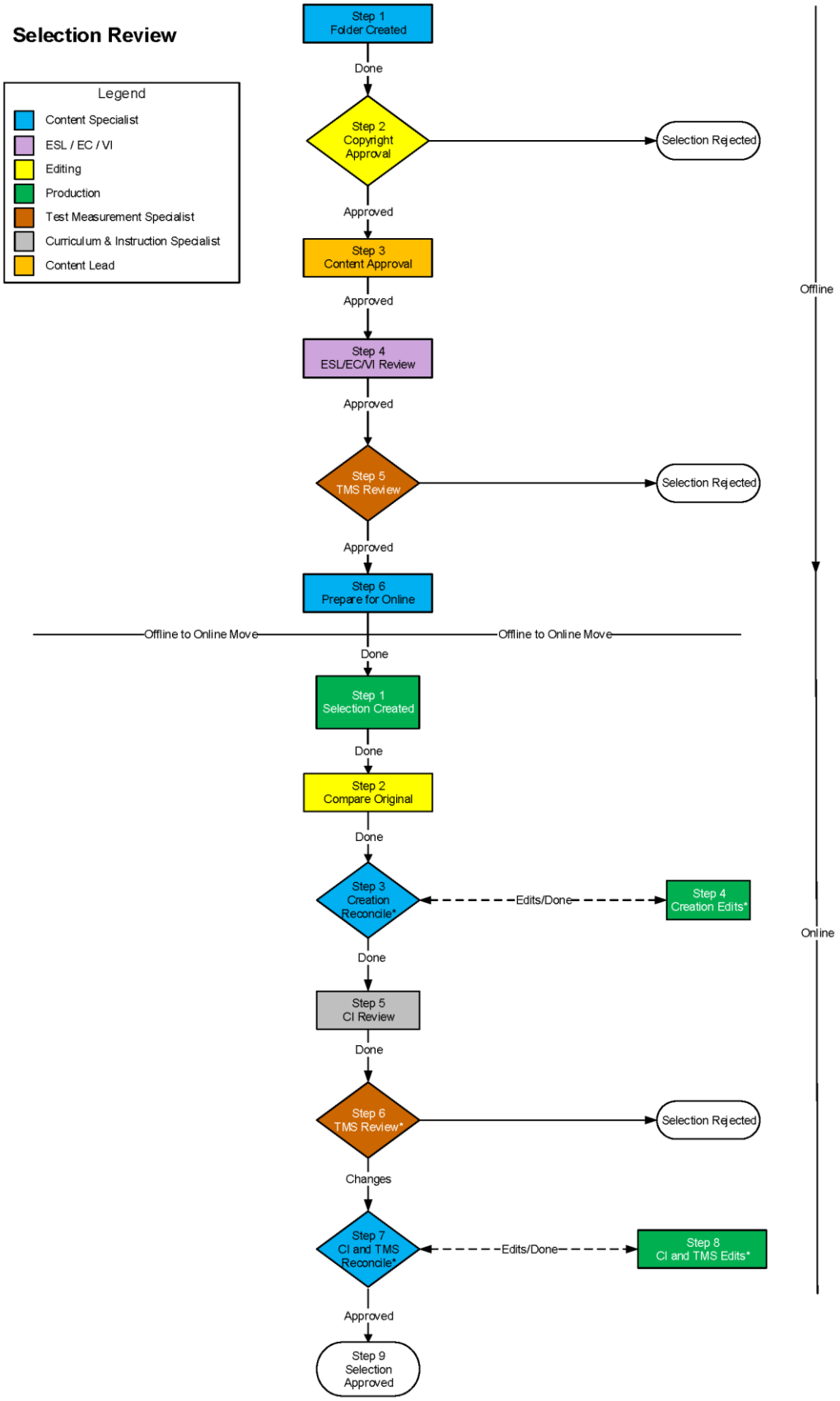
Step 7: Reconcile Curriculum and Instruction Review and Test and Measurement Specialist Review
A Content Specialist reviews any comments/changes requested by Curriculum and Instruction or by the Test and Measurement Specialist, and sends changes to **Step 8** (Production) to be made if necessary. Once 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.

Revised Thursday, April 23, 2015 4:16:28 PM

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.

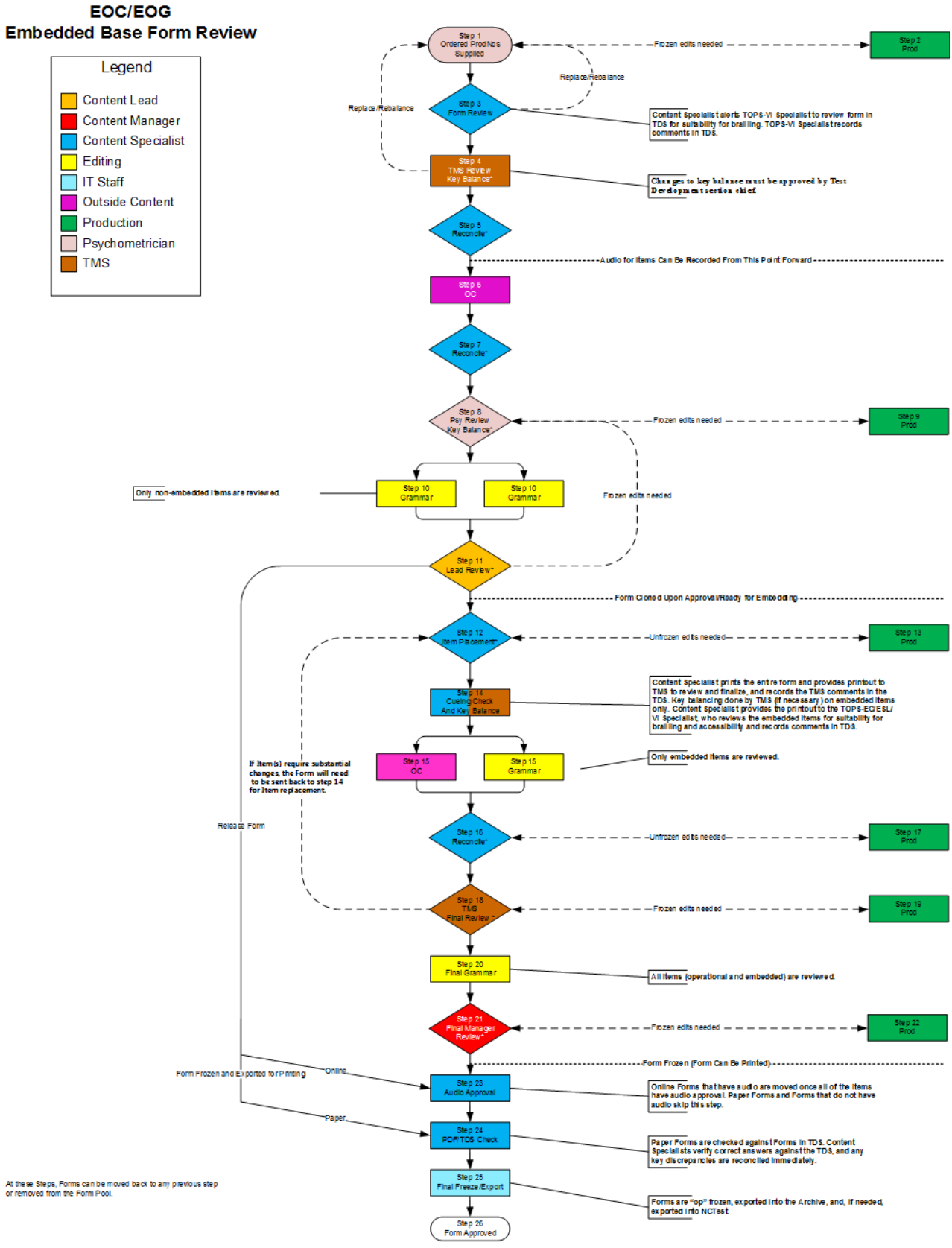

EOC/EOG
Embedded Base Form Review
Legend
Content Lead
Content Manager
Content Specialist
Editing
IT Staff
Outside Content
Production
Psychometrician
TMS
Step 1 Ordered Prod Nos Supplied
Frozen edits needed
Step 2 Prod
Replace/Rebalance
Step 3 Form Review
Content Specialist alerts TOPS-VI Specialist to 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*
Step 9 Prod
Only non-embedded items are reviewed.
Step 10 Grammar
Step 10 Grammar
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-ECE/SL/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.

Figure 1. EOG Math Grade 3 Test Information with Associated Standard Errors

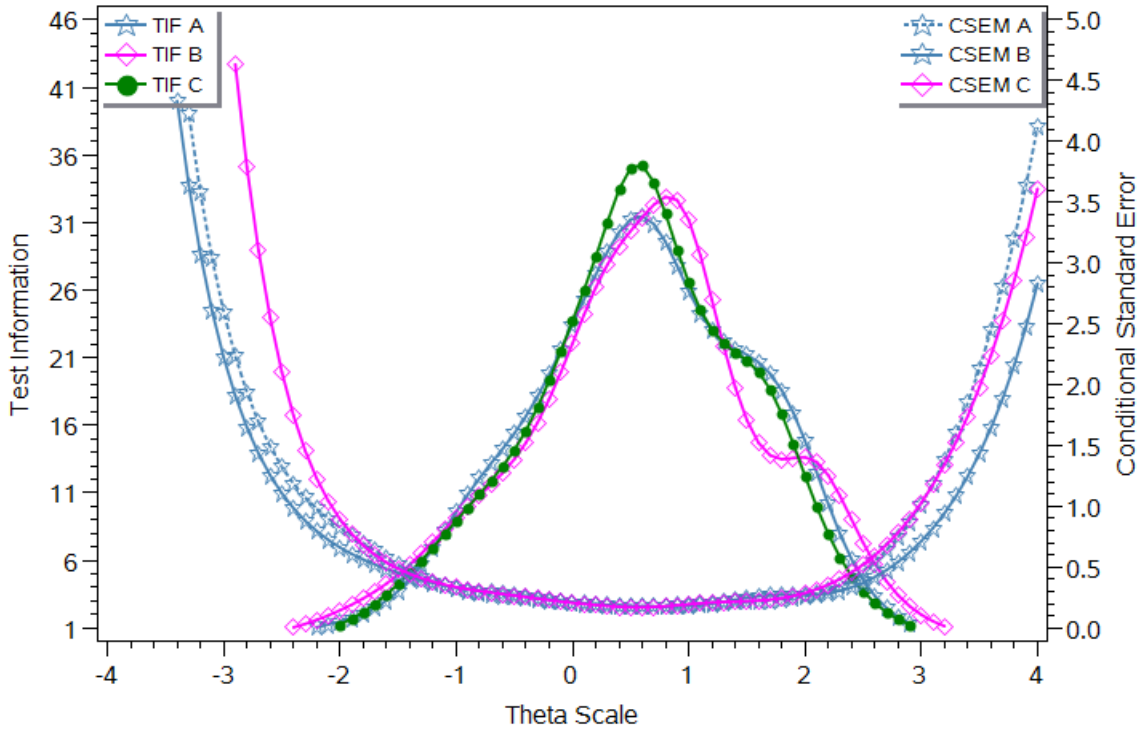


Figure 2. EOG Math Grade 4 Test Information with Associated Standard Errors

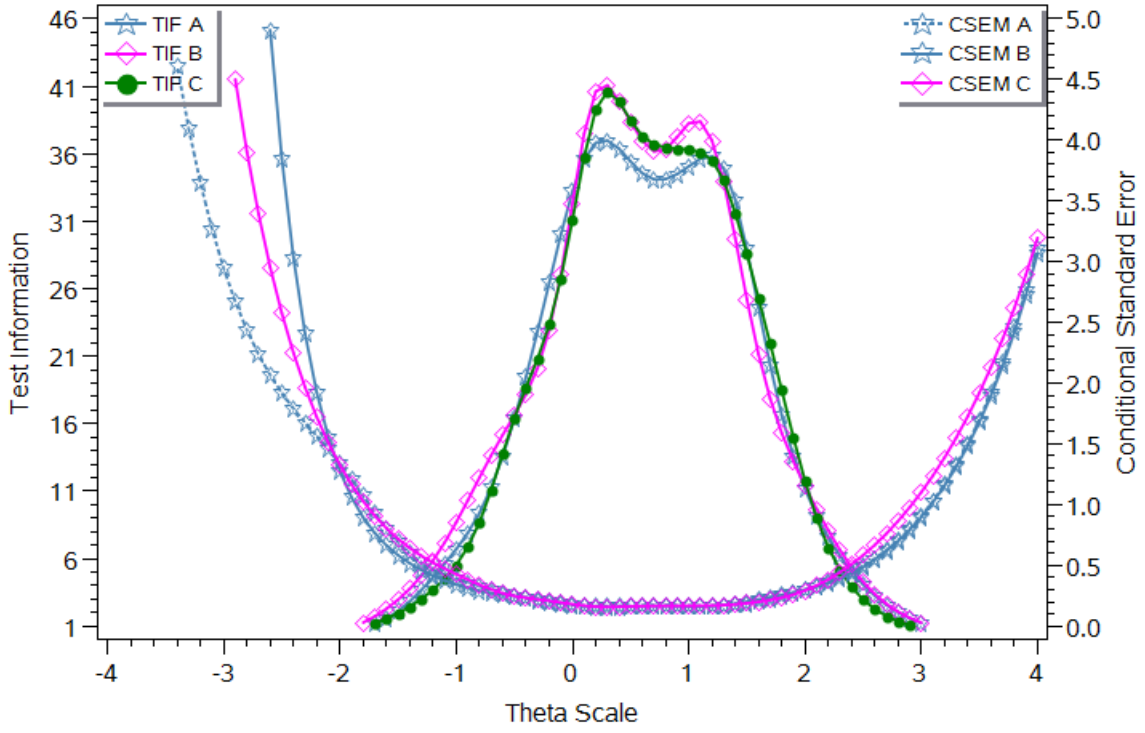


Figure 3. EOG Math Grade 5 Test Information with Associated Standard Errors

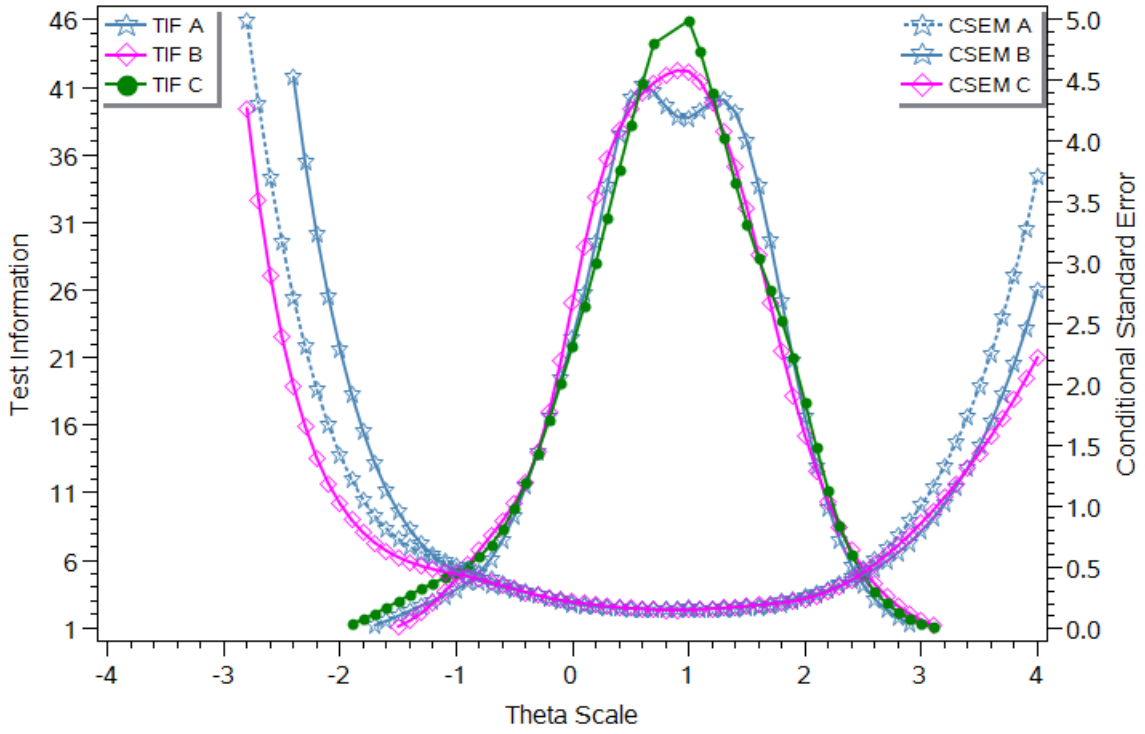


Figure 4. EOG Math Grade 6 Test Information with Associated Standard Errors

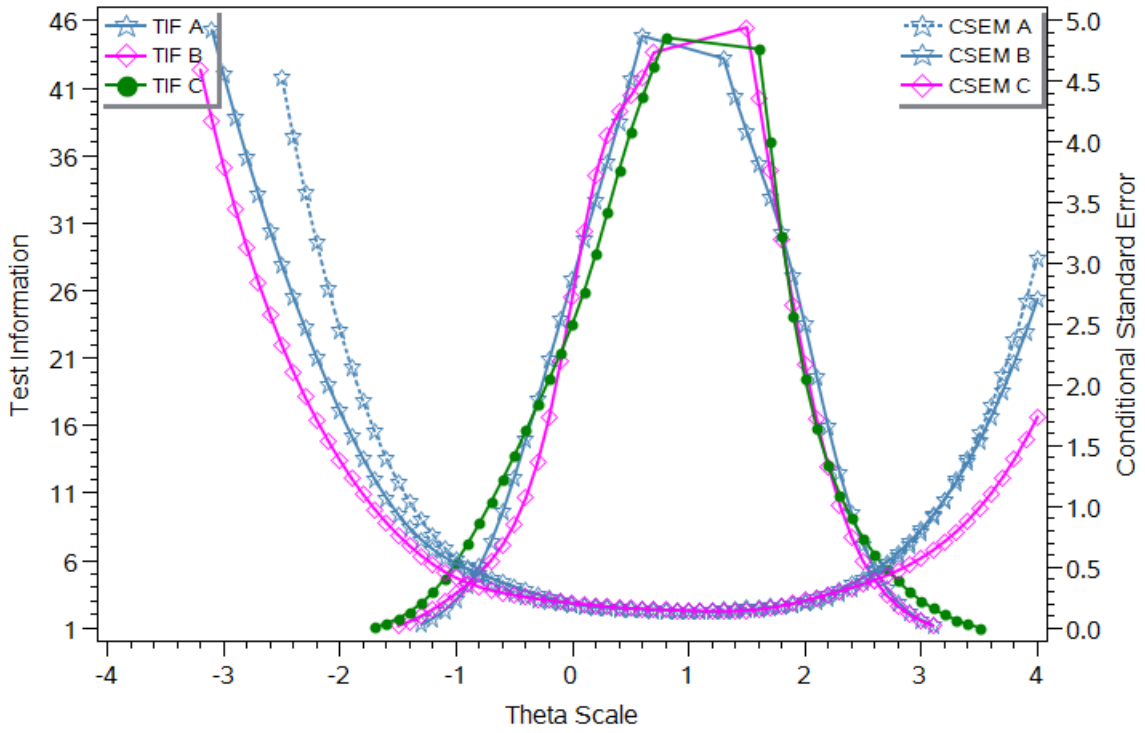


Figure 5. EOG Math Grade 7 Test Information with Associated Standard Errors

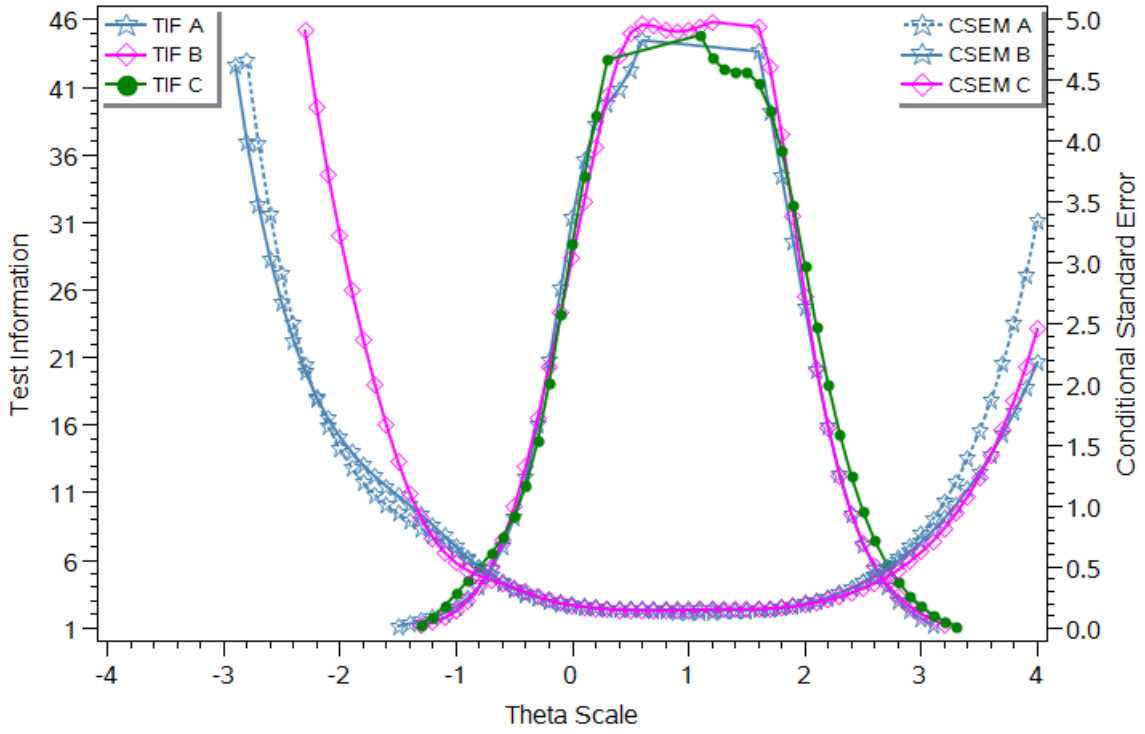


Figure 6. EOG Math Grade 8 Test Information with Associated Standard Errors

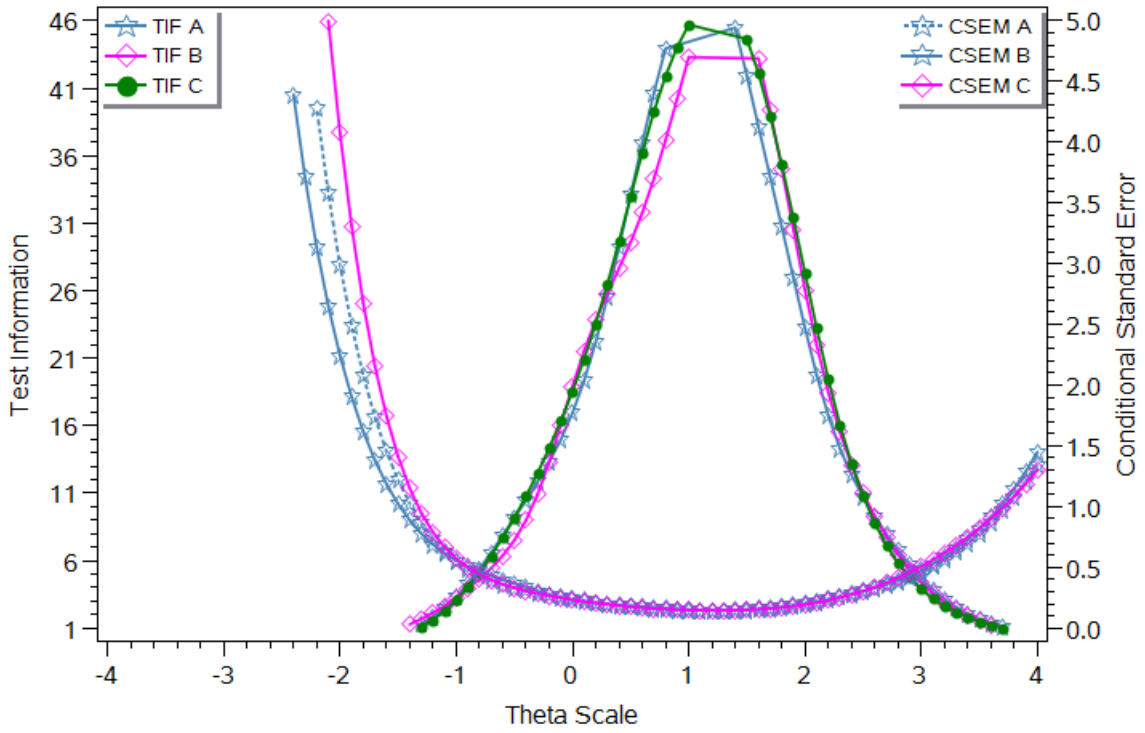
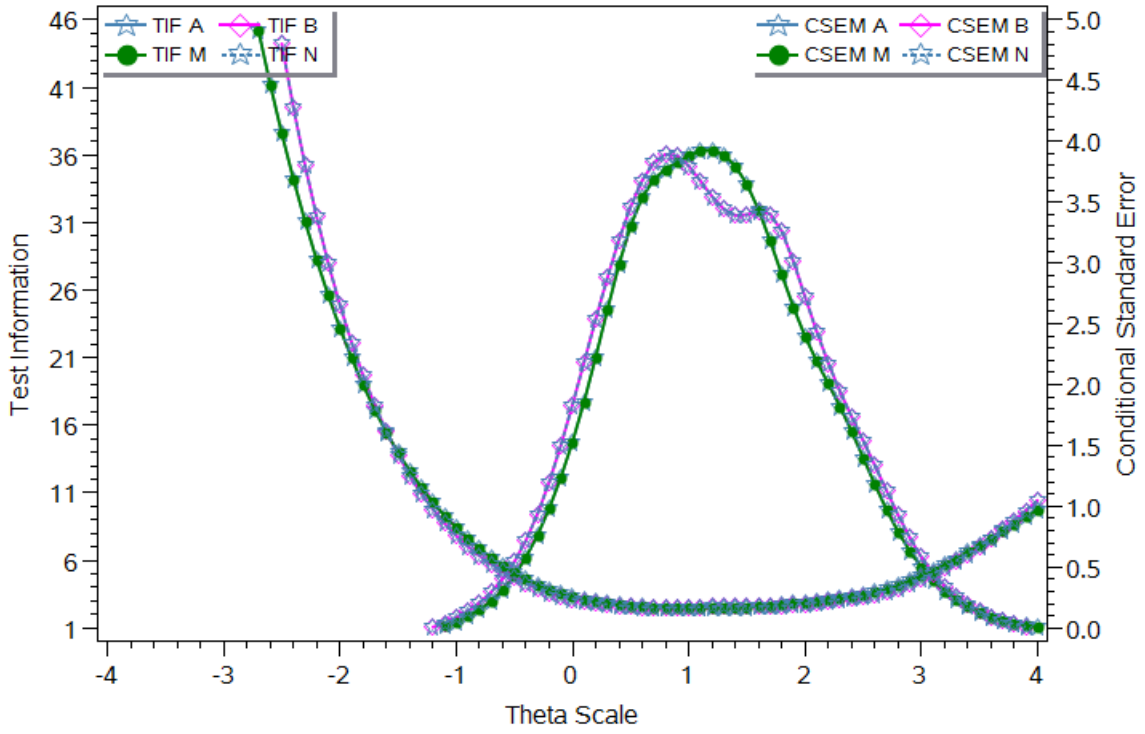


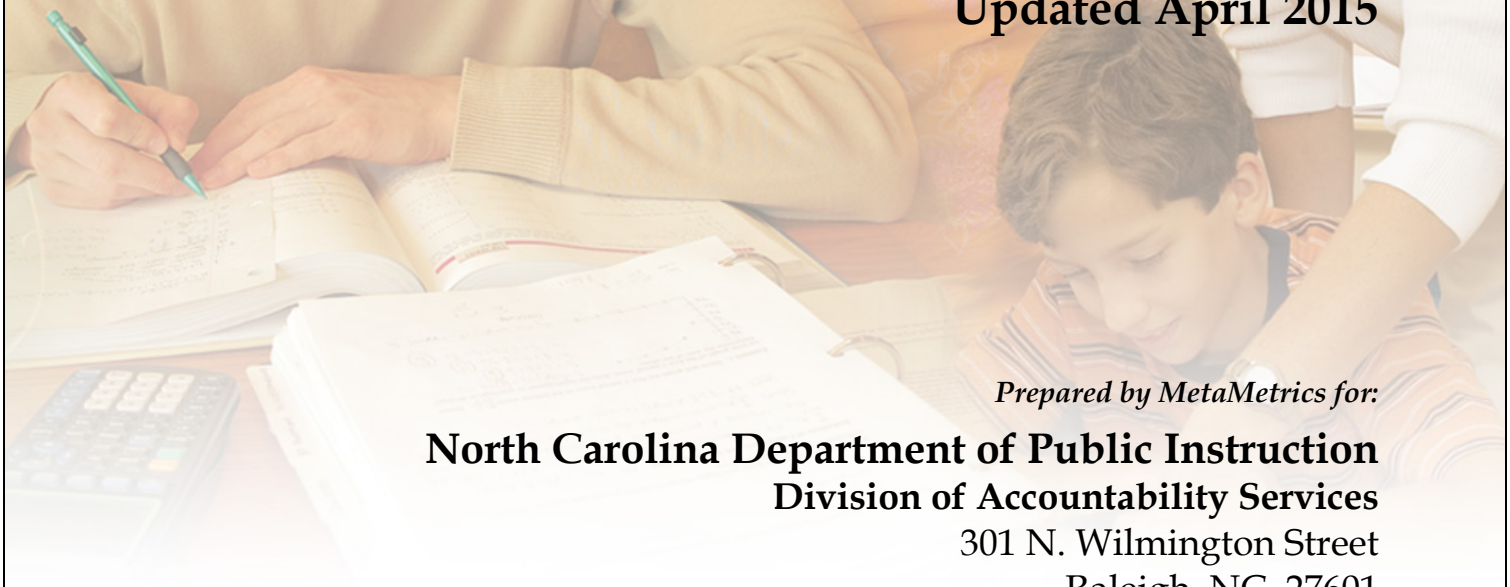
Figure 7. EOC Math I Test Information with Associated Standard Errors



Linking the NC READY EOG Math/EOC Algebra I/Integrated I with the Quantile® Framework

A Study to Link the North Carolina READY EOG Math/EOC Algebra I/Integrated I with The Quantile® Framework for Mathematics

January 2014
Updated April 2015



Prepared by MetaMetrics for:

North Carolina Department of Public Instruction
Division of Accountability Services
301 N. Wilmington Street
Raleigh, NC 27601

MetaMetrics

1000 Park Forty Plaza Drive, Suite 120
Durham, North Carolina 27713
www.MetaMetricsInc.com
www.Quantiles.com

Preface
Quantile Framework/Scale Enhancements

The Quantile® Framework for Mathematics is a scientific approach to measuring mathematics achievement and concept/application solvability. The Quantile Framework consists of a Quantile measure and the Quantile scale. A Quantile measure represents the difficulty of a mathematical skill, concept, or application. A Quantile measure also describes a student's understanding of the Quantile Skills and Concepts (QSCs) in the areas of geometry, measurement, numbers and operations, algebra, and data analysis and probability.

Quantile measures are expressed as numeric measures followed by a "Q" (e.g., 850Q), and are placed on the Quantile scale. (There is no space between the measure and the "Q.") The Quantile Framework spans the developmental continuum from prekindergarten mathematics through the content typically taught in Algebra II, Geometry, Trigonometry and Pre-Calculus -- from below 0Q (Emerging Mathematician) to above 1600Q. Quantile measures of one thousand or greater are reported without a comma (e.g., 1050Q). All Quantile measures are rounded to the nearest 5Q. If the Quantile measure is xxx2.5 or higher or xxx7.5 or higher, it is rounded up to the next highest 5Q; below those points should be rounded down. For example, if a computed Quantile measure is 772.51, it should be reported as 775Q. If the computed Quantile measure is 777.42, it should be reported as 775Q.

Prior to May 1, 2014, all Quantile measures at or below 0Q were reported as EM (Emerging Mathematician). Starting in spring 2014, Quantile measures below 0Q can be reported with a more specific measure. These EM measures are shown as "EMxxxQ." For example, a Quantile measure of -150 is reported as EM150Q where "EM" stands for "Emerging Mathematician" and replaces the negative sign in the number. The Quantile scale is like a thermometer, with numbers below zero indicating decreasing mathematical demand or achievement as the number moves away from zero. The smaller the number following the EM code, the more advanced the student is or the more demanding the skill or concept. For example, an EM150Q student is more advanced than an EM200Q student. Above 0Q, measures indicate increasing mathematical achievement as the numbers increase. For example, a 200Q QSC is more demanding than a 150Q QSC.

Quantile measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is research (e.g., to measure growth at the student, grade, school, district, or state level), then actual measures should be used at all score points, rounded to the nearest integer. If the purpose is instructional, then the Quantile measures should be capped at the upper bound of measurement error

(e.g., at the 95th percentile of the national Quantile norms) to ensure developmental appropriateness of the material. MetaMetrics expresses these measures used for instructional purposes as "Reported Quantile Measures" and recommends that they be used on individual score reports. In an instructional environment, all scores below 0Q should be reported as "EMxxxQ" (Emerging Mathematician); no student should receive a negative Quantile measure. As with any test score, uncertainty is present in the form of measurement error. The lowest reported value below 0Q is EM400Q.

Table i. Maximum reported Quantile measures by grade.

Grade	uantile Cap
	600
1	675
2	725
3	975
4	1075
5	1125
6	1200
7	1325
8	1450
9	1475
10	1500
11	1575
12	1600

Some assessments report a Quantile range for each student, which is 50Q above and 50Q below the student's actual Quantile measure. This range represents the limits within which instruction should be focused to ensure that the student understands the prerequisite skills and concepts associated with a specific QSC. Once a student's Quantile measure and grade are known, mathematical concepts, topics, materials, and resources can be identified within the same Quantile range.

The Quantile Framework has been aligned more closely with the Common Core State Standards for Mathematics. This was done by:

1. Moving from 5 to 6 strands, and
2. Adding approximately 70 QSCs.

Text on the following pages in the Technical Report has been updated to correspond with the language of the enhanced Quantile Framework/scale.

Table of Contents

Introduction

Often it is desirable to convey more information about test performance than can be incorporated into a single primary score scale. When two score scales are linked, the linkage can be used to provide a context for understanding the results of one of the assessments. It is often hard to explain what mathematical skills and concepts a student actually understands based on the results of a mathematics test. Parents typically ask the question, "Based on my child's test results, what math problems can he or she understand and how well?" Once a linkage is established with an assessment that is reported as specific concepts and skills, then the results of the assessment can be explained and interpreted in the context of the specific concepts and skills that a student will likely understand.

Auxiliary score scales can be used to "convey additional normative information, test-content information, and information that is jointly normative and content based" (Petersen, Kolen, and Hoover, 1989, p. 222). One such auxiliary scale is The Quantile® Framework for Mathematics, which was developed to appropriately match students with materials at a level where the student has the background knowledge necessary to be ready for instruction on the new mathematical skills and concepts.

The Quantile Framework for Mathematics takes the guesswork out of mathematics instruction. It serves as a hands-on tool which demonstrates which mathematics skills a learner has likely learned and which ones require further instruction. Teachers can also use the Quantile Framework to determine a student's readiness to learn more advanced skills. Because the Quantile Framework uses a common, developmental scale to measure both student mathematical achievement and task difficulty, educators can also determine how well a student is likely to be able to solve more complex problems (if provided with targeted instruction). The Quantile Framework includes the Quantile® measure and the Quantile® scale. The Quantile Framework targets instruction, forecasts understanding, and helps improve mathematics instruction and achievement by placing the mathematics curriculum, the materials to teach mathematics, and the students themselves on the same scale.

The Quantile Framework for Mathematics can be used to:

- Monitor student mathematics progress.
- Forecast student performance on end-of-year assessments.
- Match students with appropriate materials at their level.
- Determine if a student is ready for a new mathematics skill or concept.
- Link big mathematical concepts with state curriculum objectives.
- Identify student strengths and weaknesses.

- Understand the prerequisite skills needed to learn more advanced concepts in mathematics.
- Adapt instructional methods in the classroom to ensure a greater level of understanding and application.

The Quantile Framework for Mathematics is a unique resource for accurately estimating a student's ability to think mathematically and matching him/her with appropriate mathematical content. With this valuable information in the hands of educators, instruction can be more accurately tailored to the mathematical achievement of individual students. The structure of the Quantile Framework is organized around two principles—(1) mathematics and mathematical achievement are developmental in nature and (2) mathematics is a content area.

Linking assessment results with the Quantile Framework provides a mechanism for matching each student with materials on a common scale. It serves as an anchor to which resources, concepts, skills, and assessments can be connected allowing parents, teachers, and administrators to speak the same language. By using the Quantile Framework, the same metric is applied to the materials the children use, the tests they take, and the results that are reported. Parents often ask questions like the following:

- How can I help my child become better at mathematics?
- How do I challenge my child to think mathematically?

Questions like these can be challenging for parents and educators. By linking the North Carolina READY EOG Mathematics/EOC Algebra I/Integrated I Math scales with the Quantile Framework, educators and parents will be able to answer these questions and will be better able to use the results from the tests to improve instruction and to develop each student's level of mathematics understanding.

This research study was designed to determine mathematics achievement levels that can be matched with mathematical skills and concepts based on test results on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments. The study was conducted by MetaMetrics, Inc. in collaboration with the North Carolina Department of Public Instruction (NCDPI) (Contract No. NC10025818 dated December 17, 2012). The primary purposes of this study were to:

- provide tools (Math@Home, Quantile Teacher Assistant, and Math Skills Database) and information that can be used to answer questions related to standards, student-level accountability, test score interpretation, and test validation;
- create conversion tables for determining Quantile measures from the scores on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments; and
- produce a report that describes the linking analysis procedures.

The Quantile Framework for Mathematics

Just as for reading, there are dozens of tests of mathematics ability measuring a common construct and all reporting the results in proprietary, non-exchangeable metrics. The benefits of having a common supplemental metric to describe mathematics ability include the following:

(1) Individual growth trajectories spanning the educational experience can be developed because the Quantile scale is developmental in nature and spans this range.
(2) Various state definitions of grade-level proficiency can be compared by re-expressing scores on a common scale.
(3) Textbook publishers can build links between mathematics curricula and major mathematics tests.
(4) Test publishers can develop classroom/interim assessments that can link to the major mathematics tests and forecast how likely the student is to meet the state performance standards.
(5) The classroom teacher can link his or her day-to-day instructional needs to the year-to-year needs of a state-level accountability system.

The Quantile Framework consists of a common supplemental metric—the Quantile—that is employed to scientifically measure a student's ability to think mathematically and his or her mathematics achievement and to locate the student in a taxonomy of mathematical skills, concepts, and applications. In order to develop the Quantile Framework, several tasks were undertaken: (1) develop a structure of mathematics that spans the developmental continuum from first grade content through Algebra I, Geometry, and Algebra II content, (2) develop a bank of items that have been field tested, (3) develop the Quantile scale (multiplier and anchor point) based on the calibrations of the field-test items, and (4) validate the measurement of mathematics ability as defined by the Quantile Framework.

Structure of the Quantile Framework

In order to develop a framework of mathematical ability, first a structure needs to be established. The structure of the Quantile Framework is organized around two principles—(1) mathematics and mathematical ability are developmental in nature and (2) mathematics is a content area.

Developmental Nature of Mathematics. The developmental nature of mathematics over time describes the increase in sophistication of the problems that can be addressed and the increase in the integration of skills and content to address these problems. The

National Research Council (2001, 2002) described mathematical proficiency as "…five intertwined strands: (1) understanding mathematics; (2) computing fluently; (3) applying concepts to solve problems; (4) reasoning logically; and (5) engaging with mathematics, seeing it as sensible, useful, and doable" (p. 1). Geary and Hamson (2002) observed that much of mathematics can be understood as an interlocking triad of competencies: conceptual competence, procedural competence, and utilization competence. In short, these competencies refer, respectively, to (1) understanding the natural language of mathematics, (2) knowing how to read mathematical expressions and employ algorithms to solve decontextualized problems, and, finally, (3) knowing why the conceptual and procedural knowledge is important and how and when to apply it. The descriptions of these three competencies follow.

A. *Vocabulary of Mathematics*. This aspect concerns the recognition of a concept either verbally or pictorially. Concepts are drawn from the mathematical content (e.g., alternate interior angles, mean, tangent) and the mathematical process (e.g., compare, estimate, etc.) strands of the National Council of Teachers of Mathematics (NCTM) framework, and include contexts (e.g., sales tax, commission) and measurement concepts (e.g., time, weight). The NCTM Standards describe this as the language of mathematics.

B. *Procedures of Mathematics*. This aspect concerns being able to apply mathematical procedures in a controlled environment (decontextualized). Procedural items ask the student to perform operations and can include graphics. For example, (1) simplifying $(3x + 2)(4x – 8)$; or (2) identifying which three angles could form a triangle knowing that the sum of the angles of a triangle equals 180°. Procedures of mathematics can also be described as algorithmic, symbolic computation, and skills.

C. *Applications of Mathematics*. This aspect involves being able to apply a mathematical procedure to solve a problem (contextualized). Application items ask the student to apply operations and concepts and can include graphics. For example, (1) determining how many cars are needed to transport the class to the museum knowing that each car can hold four students; or (2) determining how much soil is needed for a garden plot that is 3 feet wide, 6 feet long, and 8 inches deep. Applications of mathematics can also be described as problem solving, reasoning, projects, and experiences.

MetaMetrics recognizes that in order to adequately address the scope and complexity of mathematics, multiple proficiencies/competencies must be utilized. Just as the "math wars" have brought to the forefront the various aspects of mathematics instruction, we must also address these same issues. On the issue of the "math wars," Richard Riley stated "We are suffering here from an 'either-or' mentality. As any good K-12 teacher will tell you, to get a student enthused about learning, you need a mix of information

and styles of providing that information. You need to provide traditional basics, along with more challenging concepts, as well as the ability to problem-solve, and to apply concepts in real world settings" (Starr, 2002). The Quantile Framework is an effort to recognize and define a basis for this "mix of information and styles" in the developmental context of mathematics instruction.

Content of Mathematics. A strand is a major subdivision of mathematical content. The strands describe what students should know and be able to do. The five strands of the Quantile Framework are based on the five Content Standards in the National Council of Teachers of Mathematics framework (NCTM, 2000), which are as follows:

1. *Number and Operations*. The development of number sense. Students with number sense naturally decompose numbers, use particular numbers as referents, solve problems using the relationships among operations and knowledge about the base-ten system, estimate a reasonable result for a problem, and have a disposition to make sense of numbers, problems, and results. Includes computational fluency.

   Instructional programs should enable all students to—
   - Understand numbers, ways of representing numbers, relationships among numbers, and number systems;
   - Understand meanings of operations and how they relate to one another;
   - Compute fluently and make reasonable estimates.

2. *Geometry*. The study of geometric shapes and structures; specifying their characteristics and relationships. A means to interpret and reflect on our physical environment and serve as tools for the study of other topics.

   Instructional programs should enable all students to—
   - Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships;
   - Specify locations and describe spatial relationships using coordinate geometry and other mathematical systems;
   - Apply transformations and use symmetry to analyze mathematical situations;
   - Use visualization, spatial reasoning, and geometric modeling to solve problems.

3. *Algebra/Patterns and Functions*. The relationships among quantities, the use of symbols, the modeling of phenomena, and the mathematical study of change. Instructional programs should enable all students to—
   - Understand patterns, relations, and functions;

- Represent and analyze mathematical situations and structures using algebraic symbols;
- Use mathematical models to represent and understand quantitative relationships;
- Analyze change in various contexts.

4. *Data Analysis and Probability*. The collection, analysis, and interpretation of data.

   Instructional programs should enable students to—
   - Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;
   - Select and use appropriate statistical methods to analyze data;
   - Develop and evaluate inferences and predications that are based on data;
   - Understand and apply basic concepts of probability.

5. *Measurement*. The assignment of a numerical value to an attribute of an object.

   Instructional programs should enable students to—
   - Understand measurable attributes of objects and the units, systems, and processes of measurement;
   - Apply appropriate techniques, tools, and formulas to determine measurements.

The Quantile Skills and Concepts. Within the Quantile Framework, a "Quantile Skill or Concept" (QSC) describes a specific mathematical skill and is used to annotate the Quantile scale. For example, a QSC under the Numbers and Operations strand is "Model and identify the place value of each digit in a multi-digit numeral to the hundredths place;" and a QSC under the Geometry strand is "Identify and distinguish among similar, congruent, and symmetric figures; name corresponding parts." The content taxonomy of QSCs used with the Quantile Framework was developed during the spring of 2003 for grades 1 through 8, Algebra I, and Geometry. The framework was extended to Algebra II and revised during the summer and fall of 2003. The first step in developing a content taxonomy was to review the curricular frameworks from the following sources:

- National Council of Teachers of Mathematics (NCTM).

- National Assessment of Educational Progress: 2005 Pre-Publication Edition.

- North Carolina Standard Course of Study (Revised in 2003 for grades kindergarten through high school).

- California Mathematics Framework and state assessment blueprints: *Mathematics Framework for California Public Schools: Kindergarten through Grade Twelve* (2000 Revised Edition); *Mathematics Content Standards for California Public Schools: Kindergarten through Grade Twelve* (December 1997); Blueprints document for the Star Program California Standards Tests: Mathematics (California Department of Education, adopted by SBE 10/9/02), and sample items for the California Mathematics Standards Tests (California Department of Education, January 2002).

- Florida Sunshine State Standards: Sunshine State Standards Grade Level Expectations for Mathematics, grade 2 through Mathematics. The Sunshine State Standards "are the centerpiece of a reform effort in Florida to align curriculum, instruction, and assessment." They identify what students should know and be able to do for the 21st century. Publishers are required to correlate instructional materials submitted for state adoption to the standards.

- Illinois: Illinois teachers for Illinois schools developed The Illinois Learning Standards for Mathematics. Their Goals 6 through 10 emphasize the following: numbers and operations, measurement, algebra, geometry, and data analysis and statistics—*Mathematics Performance Descriptors, Grades 1-5 and Grades 6-12* (2002).

- Texas Essential Knowledge and Skills: Texas Essential Knowledge and Skills for Mathematics (TEKS) were adopted by the Texas State Board of Education and became effective on September 1, 1998. The Texas Essential Knowledge and Skills (TEKS), the state-mandated curriculum, was "specifically designed to help students to make progress … by emphasizing the knowledge and skills most critical for student learning" (TEA, 2002b, p. 4).

  The Texas Assessment of Knowledge and Skills (TAKS) was mandated by the 76th Texas Legislature in 1999 and was administered for the first time during the 2002-2003 school year (TEA, 2002a). The TAKS was developed to assess the TEKS and ask questions in more authentic ways. The TAKS test objectives, " 'umbrella statements' generated by TEA staff with input from educators," were used to develop the items (p. 2). These statements serve as headings under which the TAKS are meaningfully grouped. The TAKS measures the statewide curriculum in reading at grades 3-9; in writing at grades 4 and 7; in English Language Arts at grades 10 and 11; in mathematics at grades 3-11; in science at grades 5, 10, and 11; and in social studies at grades 8, 10, and 11. The Spanish TAKS is administered at grades 3 through 6. Satisfactory performance on the TAKS at Grade 11 is prerequisite to a high school diploma.

The review of the content frameworks resulted in the development of a list of QSCs spanning the content typically taught in kindergarten through Algebra I, Geometry and

Algebra II. Each QSC is aligned with one of the five content strands. The QSCs can be viewed and searched at www.Quantiles.com. Each QSC consists of a description of the content, a content identification number (C_ID), the grade at which it typically first appears (Grade), and the strand it is associated with (1 = Numbers and Operations, 2 = Geometry, 3 = Algebra/Patterns & Functions, 4 = Data Analysis & Probability, and 5 = Measurement).

Although states have developed their own individual curriculum standards for years, recently there has been an unprecedented focus on developing common curriculum standards for use throughout the United States of America. Guided and supported by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA), departments of education in the states, the United States territories and the District of Columbia have collaborated to identify common standards in English/language arts, mathematics and other content areas. Educators, researchers and educational policy makers were involved extensively in the effort to identify, catalog, review and adopt standards that would lead to students being "college and career ready" by the end of high school. The Common Core State Standards (CCSS) are the culmination of this work. They were released in June 2010 by the CCSSO and the NGA Center for Best Practices. Currently, forty-five states have adopted the CCSS for Mathematics. The standards may be viewed at http://www.corestandards.org/ (NGA Center & CCSSO, 2010a, 2010b). Additional information about the development of the CCSS may be found at the CCSSO website (http://www.ccsso.org/) and the website of the NGA (http://www.nga.org/). The Quantile Framework's QSCs have been aligned with the CCSS for mathematics and, where necessary, QSCs were revised to more closely align (e.g., specifically mentioning number and word problems should be addressed by a QSC) and additional QSCs were added (e.g., margin of error, residuals of a distribution). The alignment may be viewed and searched at www.Quantiles.com.

The Quantile Framework map (Appendix A) presents a picture of the construct of mathematics ability. The map is organized by the five strands and describes the development of mathematics from basic skills to sophisticated problem solving. Exemplar QSCs and problems are used to annotate the Quantile scale and the strands. QSCs are located on the Quantile scale at the point corresponding to the mean of the ensemble of items addressing that QSC from two large, national studies (Quantile Framework field study and *PASeries* Math field study described later in this document). Items are located on the Quantile scale corresponding to their Quantile measure based on the Quantile Framework field study.

Quantile Item Bank Development

The second step in the process of developing The Quantile Framework of Mathematics was to develop and field test a bank of items that could be used in future linking

studies. Item bank development for the Quantile Framework went through several stages—content specification, item writing and review, field-testing and analyses, and final evaluation.

Item Specification and Development. Based on the list of QSCs aligned to the five strands, QSCs were identified as typically being taught at a particular grade level. The curricular frameworks from Florida, North Carolina, Texas, and California were synthesized to identify the QSCs instructed and/or assessed at each grade level. If a QSC was included in any state framework it was included in the list of QSCs for which items were to be developed for use with the Quantile Framework field study.

During the summer and fall of 2003, over 1,400 items were developed to assess the QSCs associated with content in grades 1 through Algebra II. The items were written and reviewed by mathematics educators trained to develop multiple-choice items (Haladyna, 1994). The items for the pool were specified by both strand and QSC. At least three items were written for each QSC within each grade.

With the current increased focus on authentic assessment and solving problems in context using real-world applications, mathematics items now tend to require more reading. While the vocabulary specific to mathematical content is used (e.g., congruent), every attempt is made to have the non-content vocabulary below the grade level.

Item Writer Training. Item writers were experienced teachers and item-development specialists who had experience with the everyday mathematical ability of students at various levels. The use of individuals with these types of experiences helped to ensure that the items were valid measures of mathematics. Item writers were provided with training materials concerning the development of multiple-choice items and the Quantile Framework. The item writing materials also contained incorrect and ineffective items that illustrated the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training was a short practice session with three items.

Item writers were also given additional training related to "sensitivity" issues. Part of the item writing materials addressed these issues and identify areas to avoid when selecting passages and developing items. These materials were developed based on material published on universal design and fair access—equal treatment of the sexes, fair representation of minority groups, and the fair representation of disabled individuals.

Items were reviewed and edited by a group of specialists that represented various perspectives—test developers, editors, and curriculum specialists. These individuals examined each item for sensitivity issues and for the quality of the response options.

During the second stage of the item review process, items were either "approved," "approved with edits," or "deleted."

Linking- and Field-Test Design. Tests were developed for ten levels: Levels 2 through 8 were aligned with the typical content taught in grades 2 through 8, Level 9 was aligned with the typical content taught in Algebra I, Level 10 was aligned with the typical content taught in Geometry, and Level 11 was aligned with the typical content taught in Algebra II. For each level, three forms were developed with each form containing 30 items.

First, each form consisted of 22 unique items that were targeted specifically for the grade level. Across the three grade-level forms, 66 unique items were identified. These items were selected from a pool of items that covered the content of a particular grade level. For grades 2 through 8, 22 items were from Strand 1—Numbers and Operations and 11 items were from each of the other four strands (Strand 2—Geometry, Strand 3—Algebra/Patterns & Functions, Strand 4—Data Analysis & Probability, and Strand 5—Measurement). For Algebra I and Algebra II, the primary focus of the 66 items was Strand 3—Algebra/Patterns & Functions (33 items, 50%) with the remaining items evenly distributed across the other four strands; and for Geometry, the primary focus of the 66 items was Strand 2—Geometry (33 items, 50%) with the remaining items evenly distributed across the other four strands.

Next, for each grade level, 12 of the 66 grade-level items were designated "linking" items. For each grade level set, 4 items were from Strand 1—Numbers and Operations and 2 items were from each of the other four strands (Strand 2—Geometry, Strand 3—Algebra/Patterns & Functions, Strand 4—Data Analysis & Probability, and Strand 5—Measurement). For Algebra I and Algebra II, 6 items (50%) were from Strand 3—Algebra/Patterns & Functions with the remaining six items randomly selected from the other four strands. For Geometry, 6 items (50%) were from Strand 2—Geometry with the remaining six items randomly selected from the other four strands. For Grade 1, only the "linking" set of items was included in the field-test item pool.

The linking set of items for a grade level was used to link (1) the field-test forms within the grade, (2) the field-test forms from the grade below, and (3) the field-test forms from the grade above. The final field tests were comprised of 658 unique items. Two grade 10 forms only had 29 items (one on-grade level item was dropped from each of two forms due to graphics problems).

A common-item test design was employed to vertically link the test levels. In this design, multiple tests are given to non-random groups, and a set of common items is included in the test administration to allow some statistical adjustments for possible sample-selection bias. This design is most advantageous where the number of items to be tested (treatments) is large and the consideration of cost (in terms of time) forces the

experiment to be smaller than is desired (Cochran and Cox, 1957). The multiple test forms were developed using a domain-sampling model where items were randomly assigned within QSC to a test form.

Quantile Framework Field Study—Sample. The Quantile Framework field study was conducted in February 2004. Thirty-seven schools from 14 districts across six states (California, Indiana, Massachusetts, North Carolina, Utah, and Wisconsin) agreed to participate in the study. Data were received from 34 of the schools (two elementary and one middle-school did not return data). A total of 9,847 students in grades 2 through 12 were tested. The number of students per school ranged from 74 to 920. The schools were diverse in terms of geographic location, size, and type of community (e.g., suburban; small town, city, or rural communities; and urban). *Table 1* provides information about the sample at each grade level and by gender.

Table 1. Field-study participation by grade and gender.

Grade Level	*N*	Percent Female (*N*)	Percent Male (*N*)
2	1,283	48.1 (562)	51.9 (606)
3	1,354	51.9 (667)	48.1 (617)
4	1,454	47.7 (644)	52.3 (705)
5	1,344	48.9 (622)	51.1 (650)
6	976	47.7 (423)	52.3 (463)
7	1,250	49.8 (618)	50.2 (622)
8	1,015	51.9 (518)	48.1 (481)
9	489	52.0 (252)	48.0 (233)
10	259	48.6 (125)	51.4 (132)
11	206	49.3 (101)	50.7 (104)
12	143	51.7 (74)	48.3 (69)
Missing	74	39.1 (9)	60.9 (14)
Total	9,847	49.6 (4,615)	50.4 (4,696)

Students given Levels 2 through 11 were provided with rulers and students given Levels 3 through 11 were provided with protractors. For students given taking Levels 5 through 8 and 10 and 11, formulas were provided on the back of the test booklet. Administration time was approximately 45 minutes at each level. Students given Level 2 could have the test read aloud and mark in the test booklet if that was typical of instruction.

Table 2. Test-form administration by level.

Test Level	*N*	Missing	Form 1	Form 2	Form 3
2	1,283	4	453	430	397
3	1,354	7	561	387	399
4	1,454	17	616	419	402
5	1,344	3	470	448	423
6	917	13	322	293	289
7	1,309	6	463	429	411
8	1,181	16	387	391	387
9	415	4	141	136	134
10	226	5	73	77	71
11	313	10	102	101	100
Missing	51	31	9	8	3
Total	9,847	116	3,596	3,119	3,016

Table 2 shows the number of students by level and form. The final sample included 9,678 students with complete data. Data were deleted if test level or test form was not indicated or the answer sheet was blank.

Field-Test Analyses. The field-test data were analyzed using both the classical measurement model and the Rasch (one-parameter logistic item response theory) model. Item statistics and descriptive information (item number, field test form and item number, QSC, and answer key) were printed for each item and attached to the item record. The item record contained the statistical, descriptive, and historical information for an item; a copy of the item itself as it was field-tested; any comments by reviewers; and the psychometric notations. Each item had a separate item record.

Field-Test Analyses – Classical Measurement. For each item, the *p*-value (percent correct) and the point-biserial correlation between the item score (correct response) and the total test score were computed. Point-biserial correlations were also computed between each of the incorrect responses and the total score. In addition, frequency distributions of the response choices (including omits) were tabulated (both actual counts and percents). Items with point-biserial correlations less than 0.10 were removed from the item bank. *Table 3* displays the summary item statistics.

Table 3. Summary item statistics from the Quantile Framework field study (February 2004).

Level	Number of Items Tested	Mean P-value (Range)	Mean Correct Response Point-Biserial Correlation (Range)	Mean Incorrect Responses Point-Biserial Correlation (Range)
2	90	0.583 (0.12 0.95)	0.322 (-0.15 0.56)	-0.209 (-0.43 0.12)
3	90	0.532 (0.11 0.93)	0.256 (-0.08 0.52)	-0.221 (-0.54 0.02)
4	90	0.552 (0.12 0.92)	0.242 (-0.21 0.50)	-0.222 (-0.48 0.12)
5	90	0.535 (0.12 0.95)	0.279 (-0.05 0.50)	-0.225 (-0.45 0.05)
6	90	0.515 (0.04 0.86)	0.244 (-0.08 0.45)	-0.218 (-0.46 0.09)
7	90	0.438 (0.10 0.77)	0.294 (-0.12 0.56)	-0.207 (-0.46 0.25)
8	90	0.433 (0.10 - 0.81)	0.257 (-0.15 0.50)	-0.201 (-0.45 0.13)
9	90	0.396 (0.10 0.79)	0.208 (-0.19 0.52)	-0.193 (-0.53 0.22)
10	88	0.511 (0.01 0.97)	0.193 (-0.26 0.53)	-0.205 (-0.55 0.18)
11	90	0.527 (0.09 0.98)	0.255 (-0.09 0.51)	-0.223 (-0.52 0.07)

Field-Test Analyses – Bias. Differential item functioning (DIF) examines the relationship between the score on an item and group membership while controlling for ability. The Mantel-Haenszel procedure has become "the most widely used methodology [to examine differential item functioning] and is recognized as the testing industry standard" (Roussos, Schnipke, and Pashley, 1999, p. 293). The Mantel-Haenszel procedure examines DIF by examining *j* 2 × 2 contingency tables, where *j* is the number

of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the group of interest and the reference group serves as a basis for comparison for the focal group (Dorans and Holland, 1993; Camilli and Shepherd, 1994).

The Mantel-Haenszel chi-square statistic tests the alternative hypothesis that there is a linear association between the row variable (score on the item) and the column variable (group membership). The χ^2 distribution has 1 degree of freedom and is determined as

$$Q_{MH} = (n-1)r^2 \qquad \text{(Equation 1)}$$

where *r* is the Pearson correlation between the row variable and the column variable (SAS Institute, 1985).

The Mantel-Haenszel (MH) Log Odds Ratio statistic is used to determine the direction of differential item functioning (SAS Institute Inc., 1985). This measure is obtained by combining the odds ratios, α_j, across levels with the formula for weighted averages (Camilli and Shepherd, 1994, p. 110):

$$\alpha_j = \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} = \frac{\Omega_{Rj}}{\Omega_{Fj}} \qquad \text{(Equation 2)}$$

For this statistic, the null hypothesis of no relationship between score and group membership, or that the odds of getting the item correct are equal for the two groups, is not rejected when the odds ratio equals 1. For odds ratios greater than 1, the interpretation is that an individual at score level *j* of the Reference Group has a greater chance of answering the item correctly than an individual at score level *j* of the Focal Group. Conversely, for odds ratios less than 1, the interpretation is that an individual at score level *j* of the Focal Group has a greater chance of answering the item correctly than an individual at score level *j* of the Reference Group. The Breslow-Day Test is used to test whether the odds ratios from the *j* levels of the score are all equal. When the null hypothesis is true, the statistic is distributed approximately as a χ^2 with *j*-1 degrees of freedom (Camilli and Shepherd, 1994; SAS Institute, 1985).

For the gender analyses, males (approximately 50.4% of the population) were defined as the reference group and females (approximately 49.6% of the population) were defined as the focal group. The results from the Quantile Framework field study were reviewed for inclusion on later linking studies. The following statistics were reviewed for each item: *p*-value, point-biserial correlation, and DIF estimates. Items that exhibited extreme statistics were removed from the item bank (47 out of 685).

From the studies conducted with the Quantile Framework item bank (Palm Beach County [FL] linking study, Mississippi linking study, DoDEA/TerraNova linking

study, and Wyoming linking study), approximately 6.9% of the items in any one study were flagged as exhibiting DIF using the Mantel-Haenszel statistic and the *t*-statistic from Winsteps. For each linking study the following steps were used to review the items: (1) flag items exhibiting DIF, (2) review items to determine if the content of the item is something that all students should know and be able to do, and (3) make decision to retain or delete the item.

Field-Test Analyses – Rasch Item Response Theory. Classical test theory has two basic shortcomings: (1) the use of item indices whose values depend on the particular group of examinees from which they were obtained, and (2) the use of examinee ability estimates that depend on the particular choice of items selected for a test. The basic premises of item response theory (IRT) overcome these shortcomings by predicting the performance of an examinee on a test item based on a set of underlying abilities (Hambleton and Swaminathan, 1985). The relationship between an examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases.

The conversion of observations into measures can be accomplished using the Rasch (1980) model, which states a requirement for the way that item calibrations and observations (count of correct items) interact in a probability model to produce measures. The Rasch IRT model expresses the probability that a person (n) answers a certain item (i) correctly by the following relationship:

$$P_{ni} = \frac{e^{b_n - d_i}}{1 + e^{b_n - d_i}} \qquad \text{(Equation 3)}$$

where d_i is the difficulty of item i (i = 1, 2, …, number of items);

b_n is the ability of person n (n = 1, 2, …, number of persons);

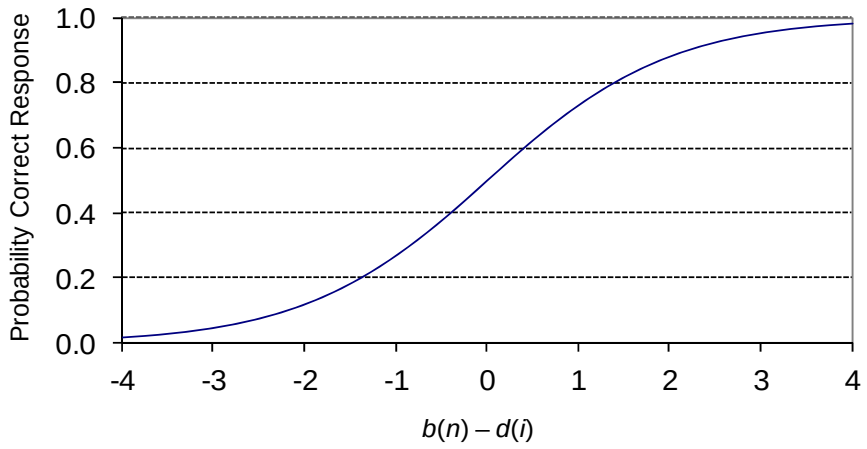
$b_n - d_i$ is the difference between the ability of person n and the difficulty of item i; and

P_{ni} is the probability that examinee n responds correctly to item i
(Hambleton and Swaminathan, 1985; Wright and Linacre, 1994).

This measurement model assumes that item difficulty is the only item characteristic that influences the examinee's performance such that all items are equally discriminating in their ability to identify low-achieving persons and high achieving persons (Bond and Fox, 2001; and Hambleton, Swaminathan, and Rogers, 1991). In addition, the lower asymptote is zero, which specifies that examinees of very low ability have zero probability of correctly answering the item. The Rasch model has the following assumptions: (1) unidimensionality—only one ability is assessed by the set of items; and (2) local independence—when abilities influencing test performance are held constant,

an examinee's responses to any pair of items are statistically independent (conditional independence, i.e., the only reason an examinee scores similarly on several items is because of his or her ability, not because the items are correlated). The Rasch model is based on fairly restrictive assumptions, but it is appropriate for criterion-referenced assessments. *Figure 1* graphically shows the probability that a person will respond correctly to an item as a function of the difference between a person's ability and an item's difficulty.

Figure 1. The Rasch Model—the probability person *n* responds correctly to item *i*.



An assumption of the Rasch model is that the probability of a response to an item is governed by the difference between the item calibration (d_i) and the person's measure (b_n). From an examination of the graph in *Figure 1*, when the ability of the person matches the difficulty of the item ($b_n - d_i = 0$), then the person has a 50% probability of responding to the item correctly.

The number of correct responses for a person is the probability of a correct response summed over the number of items. When the measure of a person greatly exceeds the calibration (difficulties) of the items ($b_n - d_i > 0$), then the expected probabilities will be high and the sum of these probabilities will yield an expectation of a high "number correct." Conversely, when the item calibrations generally exceed the person measure ($b_n - d_i < 0$), the modeled probabilities of a correct response will be low and the expectation will be a low "number correct."

Thus, Equation 3 can be rewritten in terms of the number of correct responses of a person on a test

$$O_p = \sum_{i=1}^{L} \frac{e^{b_n - d_i}}{1 + e^{b_n - d_i}} \qquad \text{(Equation 4)}$$

where O_p is the number of correct responses of person p and L is the number of items on the test.

When the sum of the correct responses and the item calibrations (d_i) is known, an iterative procedure can be used to find the person measure (b_n) that will make the sum of the modeled probabilities most similar to the number of correct responses. One of the key features of the Rasch IRT model is its ability to place both persons and items on the same scale. It is possible to predict the odds of two individuals being successful on an item based on knowledge of the relationship between the abilities of the two individuals. If one person has an ability measure that is twice as high as that of another person (as measured by b—the ability scale), then he or she has twice the odds of successfully answering the item.

Equation 4 possesses several distinguishing characteristics:

- The key terms from the definition of measurement are placed in a precise relationship to one another.

- The individual responses of a person to each item on an instrument are absent from the equation. The only information that appears is the "count correct" (O_p), thus confirming that the raw score (i.e., number of correct responses) is "sufficient" for estimating the measure.

For any set of items the possible raw scores are known. When it is possible to know the item calibrations (either theoretically or empirically from field studies), the only parameter that must be estimated in Equation 4 is the person measure that corresponds to each observable count correct. Thus, when the calibrations (d_i) are known, a correspondence table linking observation and measure can be constructed without reference to data on other individuals.

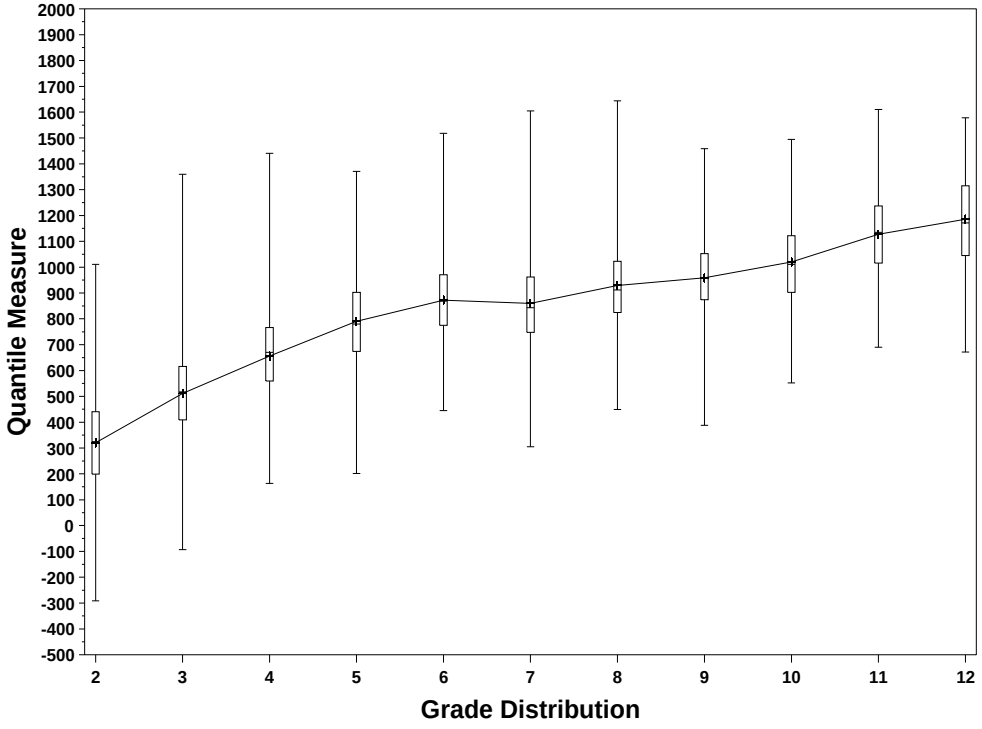
All students and items were submitted to a Winsteps analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.001. Items that a student skipped were treated as missing, rather than being treated as incorrect. Only students who responded to at least 20 items were included in the analyses (22 students were omitted, 0.22%). *Table 4* shows the mean and median Quantile measures for all students with complete data at each grade level. While there is not a monotonically increasing trend in the mean and median Quantile measures in Grades 6 and 7, the measures are not significantly different. Results from other studies (e.g., *PASeries* Math described beginning on page 25 exhibit a monotonically increasing function).

Table 4. Mean and median Quantile measures for students with complete data (*N* = 9,656).

Grade Level	*N*	Mean uantile measure (SD)	Median uantile measure
2	1,275	320.68 (189.11)	323
3	1,339	511.41 (157.69)	516
4	1,427	655.45 (157.50)	667
5	1,337	790.06 (167.71)	771
6	959	871.82 (153.02)	865
7	1,244	860.52 (174.16)	841
8	1,004	929.01 (157.63)	910
9	482	958.69 (152.81)	953
10	251	1019.97 (162.87)	1005
11	200	1127.34 (178.57)	1131
12	138	1185.90 (189.19)	1164

Figure 2 shows the relationship between grade level and Quantile measure. The following box and whisker plots (*Figures 2, 3,* and *4*) show the progression of the *y*-axis scores from grade to grade (the *x*-axis). For each grade, the box refers to the inter-quartile range. The line within the box indicates the median and the + indicates the mean. The end of each whisker shows the minimum and maximum values of the *y*-axis which is the Quantile measure. Across all students, the correlation between grade and Quantile measure was 0.76.

Figure 2. Box and whisker plot of the Rasch ability estimates of all students with complete data (*N* = 9,656).



All students with outfit mean square statistics greater than or equal to 1.8 were removed from further analyses. A total of 480 students (4.97%) were removed from further analyses. The number of students removed ranged from 8.47% (108) in grade 2 to 2.29% (22) in grade 6 with a mean percent decrease of 4.45% per grade.

All remaining students (9,176) and all items were submitted to a Winsteps analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.001. Items that a student skipped were treated as missing, rather than being treated as incorrect. Only students who responded to at least 20 items were included in the analyses. *Table 5* shows the mean and median Quantile measures for the final set of students at each grade level. *Figure 3* shows the results from the final set of students. The correlation between grade level and Quantile measure was 0.78.

Table 5. Mean and median Quantile measures for the final set of students (*N* = 9,176).

Grade Level	*N*	Median Logit Value	Mean uantile measure (Median)
2	1,167	-2.800	289.03 (292)
3	1,260	-1.650	502.18 (499)
4	1,352	-0.780	652.60 (656)
5	1,289	0.000	795.25 (796)
6	937	0.430	880.77 (874)
7	1,181	0.370	877.75 (863)
8	955	0.810	951.41 (942)
9	466	1.020	982.62 (980)
10	244	1.400	1044.08 (1048)
11	191	2.070	1160.49 (1169)
12	134	2.295	1219.87 (1210)

Figure 3. Box and whisker plot of the Rasch ability estimates for the final sample of students with outfit statistics less than 1.8 (*N* = 9,176).

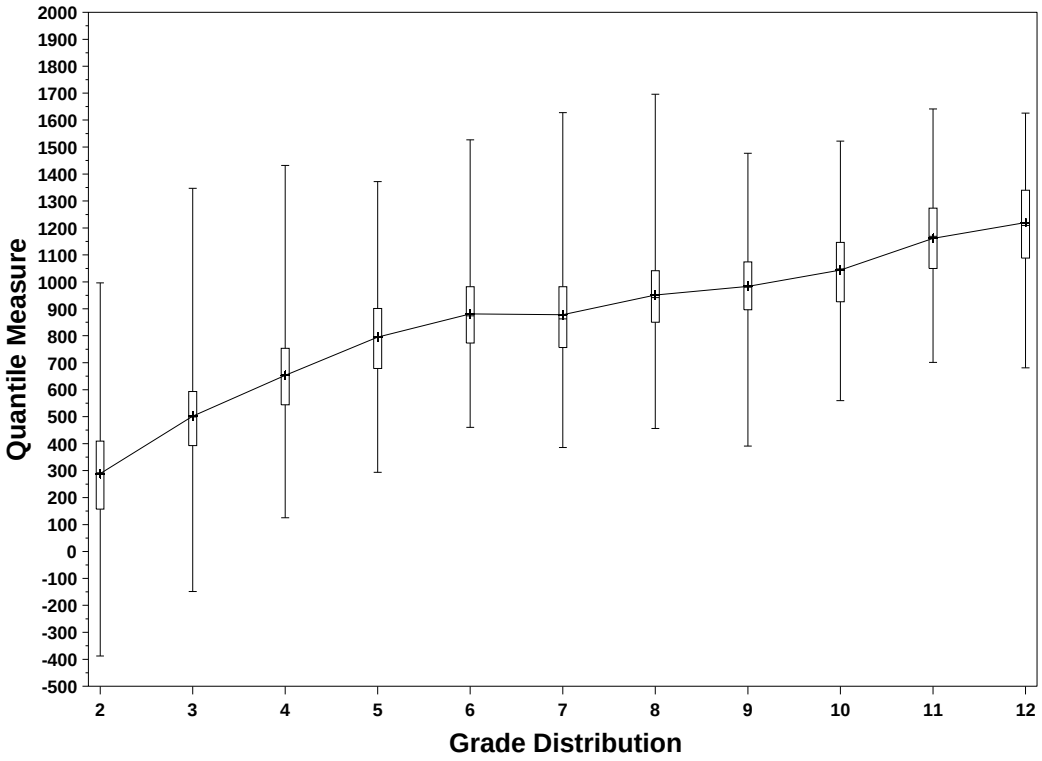
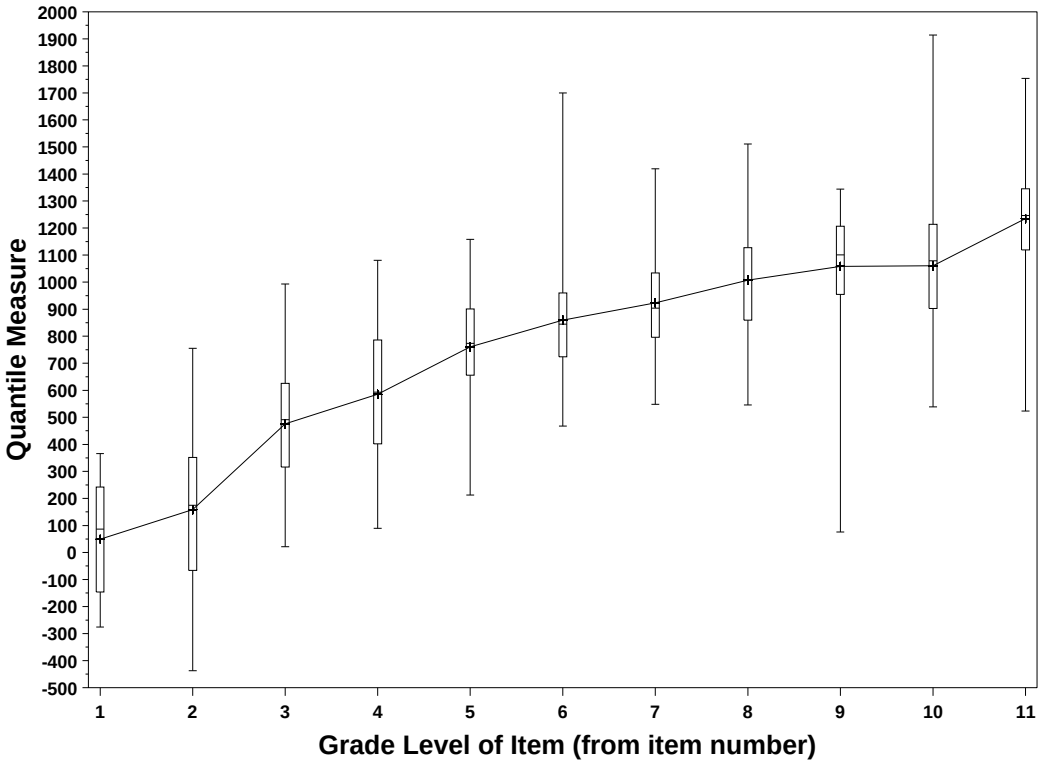


Figure 4 shows the distribution of item difficulties based on the final sample of students. For this analysis, missing data were treated as "skipped" items and not counted as wrong. There is a gradual increase in difficulty when items are sorted by level of test for which the items were written. This distribution appears to be non-linear, which is

consistent with other studies. The correlation between the grade level for which the item was written and the Quantile measure of the item was 0.80.

Figure 4. Box and whisker plot of the Rasch difficulty estimates of the 685 Quantile Framework items for the final sample of students (*N* = 9,176).



Calibration of Items on the Quantile Scale

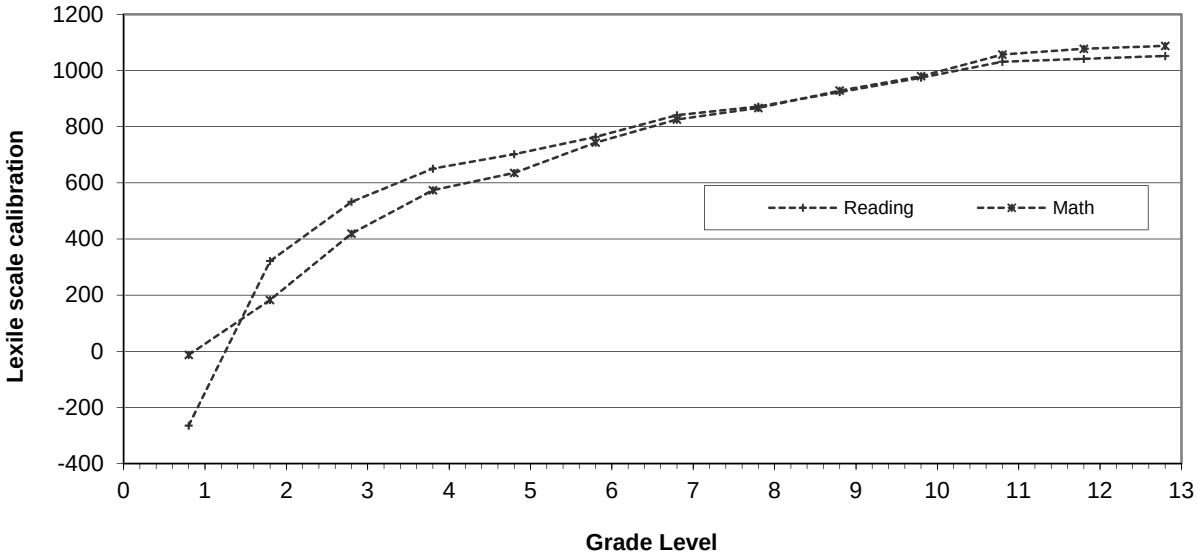
In developing the Quantile scale, two features of the scale were needed: (1) scale multiplier (conversion factor) and (2) anchor point. The Rasch item response theory model (Wright and Stone, 1979) was used to estimate the difficulties of items and the abilities of persons on the logit scale.

The calibrations of the items from the Rasch model are objective in the sense that the relative difficulties of the items will remain the same across different samples of persons (specific objectivity). When two items are administered to the same person it can be determined which item is harder and which one is easier. This ordering should hold when the same two items are administered to a second person. If two different items are administered to the second person, there is no way to know which set of items is harder and which set is easier. The problem is that the location of the scale is not known. General objectivity requires that scores obtained from different test administrations be tied to a common zero—absolute location must be sample independent (Stenner, 1990).

To achieve general objectivity, the theoretical logit difficulties must be transformed to a scale where the ambiguity regarding the location of zero is resolved.

The first step in developing the Quantile scale was to determine the conversion factor (CF) to be used to go from logits to Quantile measure. Based on prior research with reading and the Lexile scale, the decision was made to examine the relationship between reading and mathematics scales used with other assessments. The median scale score for each grade level on a norm-referenced assessment linked with the Lexile scale is plotted in *Figure 5* using the same conversion equation for both reading and mathematics.

Figure 5. Relationship between reading and mathematics scale scores on a norm-referenced assessment linked to the Lexile scale in reading.

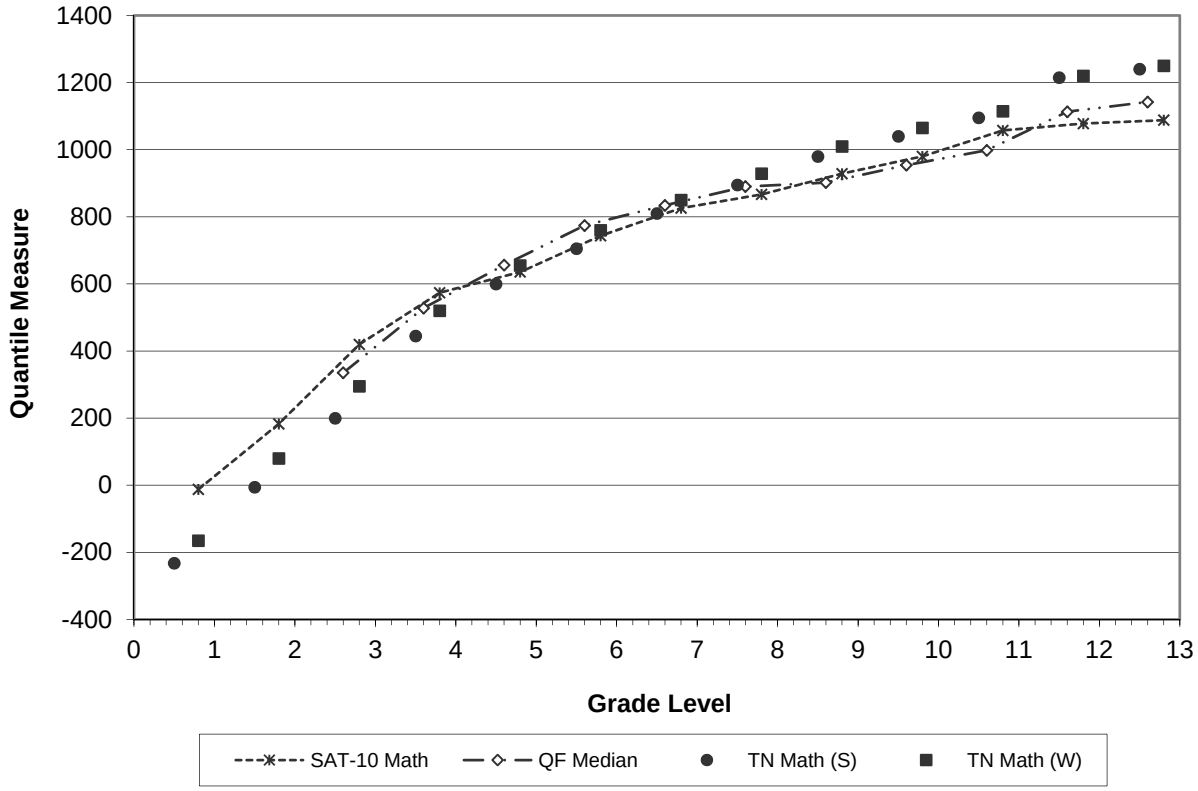


Based on an examination of *Figure 5*, it was concluded that the same conversion factor of 180 that is used with the Lexile scale could be used with the Quantile scale. Both sets of data exhibited a similar pattern across grades.

The second step in developing the Quantile scale with a fixed zero was to identify an anchor point for the scale. Given the number of students at each grade level in the field study, it was concluded that the scale should be anchored at grade 4 or 5 (middle of grade span typically tested by state assessment programs). Median performance at the end of grade 3 on the Lexile scale is 590L. The Quantile Framework field study was conducted in February and this point would correspond to six months (0.6) through the

school year. Median performance at the end of grade 4 on the Lexile scale is 700L. To determine the location of the scale, 66Q were added to the median performance at the end of grade 3 to reflect the growth of students in grade 4 prior to the field study (700 – 590 = 110; 110 × 0.6 = 66). The value of 656Q was used for the location of grade 4 median performance. The anchor point was validated with other assessment data and collateral data from the Quantile Framework field study (see *Figure 6*).

Figure 6. Relationship between grade level and mathematics performance on the Quantile Framework field study and other mathematics assessments.



Finally, a linear equation of the form

$$[(\text{Logit} - \text{Anchor Logit}) \times \text{CF}) + 656 = \text{Quantile measure} \qquad \text{(Equation 5)}$$

was developed to convert logit difficulties to Quantile calibrations where the anchor logit is the median for grade 4 in the Quantile Framework field study.

Quantile Skill and Concept (QSC) Quantile Measures

In order to use the Quantile Framework to examine the difficulty of skills and concepts and the complexity of resources, the Quantile measure of each QSC must be estimated. The Quantile measure of a QSC estimates its solvability, or a prediction of how difficult the skill or concept will be for the learner with a Quantile measure of his or her own. The QSCs fall into knowledge clusters along a content continuum.

The Quantile measures and knowledge clusters for QSCs are determined by a group of three to five subject-matter experts (SMEs). Each SME has had classroom experience at multiple developmental levels, has completed graduate-level courses in mathematics education, and understands basic psychometric concepts and assessment issues.

Knowledge Clusters. Knowledge clusters are a family of skills, like building blocks, that depend one upon the other to connect and demonstrate how skills are founded, supported, and extended along the continuum. The knowledge clusters illustrate the interconnectivity of the Quantile Framework and the natural progression of mathematical skills (content progressions) needed to solve increasingly complex problems (Hudnutt, 2012).

Each QSC was classified as having "prerequisite" and "supplemental" QSCs or as being a "foundational" QSC by the SMEs. A *prerequisite QSC* is a QSC that describes a skill or concept that provides the foundation necessary for another QSC. For example, adding single-digit numbers is a prerequisite for adding two-digit numbers. A *supplemental QSC* is a QSC which describes supplementary skills or knowledge that assists and enriches the understanding of another QSC. An *impending QSC* describes the skills and concepts that will be built from a primary QSC and helps the teacher or parent to see a trajectory of knowledge across grades and content strands. The SMEs examined each QSC to determine where the specific QSC comes in the content progression based on classroom experience, instructional resources (e.g., textbooks), and other curricular frameworks (e.g., NCTM Standards). A QSC that is classified as "foundational" means this QSC describes a skill or concept that only requires readiness to learn. Readiness is based upon the learner's cognitive experiences rather than knowledge of specific mathematical concepts. It is the basis upon which other QSCs are built.

Once the knowledge cluster for a QSC was established, the information was used when determining the Quantile measure of a QSC (described below). If necessary, knowledge clusters were reviewed and refined if the Quantile measures of the QSCs in the cluster were not monotonically increasing or there was not an instructional explanation for the pattern.

Quantile measures of QSCs. To determine the Quantile measure of a QSC, actual performance by examinees was used. While expert judgment alone could have been

used to scale the QSCs, empirical scaling was more replicable. Items and resulting data from two national field studies were used in the process:

- Quantile Framework field study (685 items, *N* = 9,647, grades 2 through Algebra II) which is described earlier in this section; and
- *PASeries* Mathematics field study (7,080 items, *N* = 27,329, grades 2 through 9/Algebra I) which is described in the *PASeries* Mathematics Technical Manual (MetaMetrics, 2005).

The items initially associated with each QSC were reviewed by SMEs and accepted for inclusion in the set of items, moved to another QSC, or not included in the set. The following criteria were used:

- Psychometric (responded to by at least 50 examinees, administered at the target grade level, point-biserial correlation greater than or equal to 0.16);
- Matched grade level of introduction of concept/skill from national review of curricular frameworks (described on pages 3 and 4); and,
- Appropriate for instruction of concept (first night's homework; from the A and B sections of the lesson problems) based on consensus of the SMEs.

Once the set of items meeting the inclusion criteria was identified, the set of items was reviewed to ensure that the curricular breadth of the QSC was covered. If the group of SMEs considered the set of items to be acceptable, then the Quantile measure of the QSC was calculated. The Quantile measure of a QSC is defined as the mean Quantile measure of items that met the criteria. The standard deviation of the item difficulties was also calculated (mean standard deviation of item difficulties across QSCs was 177.3Q). The final step in the process was to review the Quantile measure of the QSC in relationship to the Quantile measures of the QSCs identified as prerequisite and supplementary to the QSC. If the group of SMEs did not consider the set of items to be acceptable, then the Quantile measure of the QSC was estimated and assigned a Quantile zone. By assigning a Quantile zone instead of a Quantile measure to a QSC, the SMEs were able to provide a valid estimate of the skill or concept's difficulty.

QSC Quantile measures are used in the calibration of resources (e.g., textbooks, instructional materials, supplemental materials, workplace documents, everyday documents) used with the Quantile Framework.

Validation of The Quantile Framework for Mathematics

Validity is the extent to which a test measures what its authors or users claim it measures; specifically, test validity concerns the appropriateness of inferences "that can be made on the basis of observations or test results" (Salvia and Ysseldyke, 1998, p. 166).

The 1999 *Standards for Educational and Psychological Testing* (America Educational Research Association, American Psychological Association, and National Council on Measurement in Education) state that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed in the uses of tests" (p. 9). For the Quantile Framework, which measures student understanding of mathematical skills and concepts, the most important aspect of validity that should be examined is construct validity. The construct validity of The Quantile Framework for Mathematics can be evaluated by examining how well Quantile measures relate to other measures of mathematics described in the following sections.

Standardization set of items used with PASeries Mathematics. *PASeries* Mathematics is a series of classroom-based, progress monitoring assessments designed for use in the US school market in grades 3 through 8 (MetaMetrics, 2005). Each *PASeries* Mathematics assessment measures a range of mathematics skills appropriate to a specific grade. For each grade, *PASeries* Mathematics includes a screener test (pre-test) and six progress assessments designed to be administered approximately every six weeks. Each assessment contains 30 items; an assessment can be administered in one typical class period. As the school year progresses, each assessment is designed to be at a higher Quantile level, resulting in progressively more challenging tests.

For the standardization set, the items in the Quantile Framework field study that were also in the *PASeries* Mathematics field study were examined. Only items that were presented in exactly the same form in both studies were retained. A total of 213 items were identified that were administered in both studies. One item was dropped because none of the responses were correct, five items were dropped because they were too easy, and five items were dropped because there were presentation differences between the studies. The final number of items in the standardization set was 207. The test level breakdown is presented in *Table 6*.

Table 6. Number of items in the Quantile Framework standardization set by grade level of the item content.

Content Level of Items (by Grade)	Number of Items in Standardization Set
1	6
2	32
3	25
4	29
5	27
6	26
7	27
8	19
9	15
10	1

The relationship between the calibrations of the standardization set of items used in the Quantile Framework field study and on *PASeries* Mathematics (the calibration of the *PASeries* Mathematics items will be described later in this technical manual) was examined. The correlation of the Quantile measures of the 207 items was 0.92. The mean difference was -186Q and the standard deviation of the differences was 153Q. The standardization set of items is validated by consistency of measures between the two studies. Characteristics of the items in the standardization set from the two field studies are presented in *Figures 7* and *8*.

Figure 7. Comparison of the difficulty (Quantile measure) of the standardization set of items across two field studies.

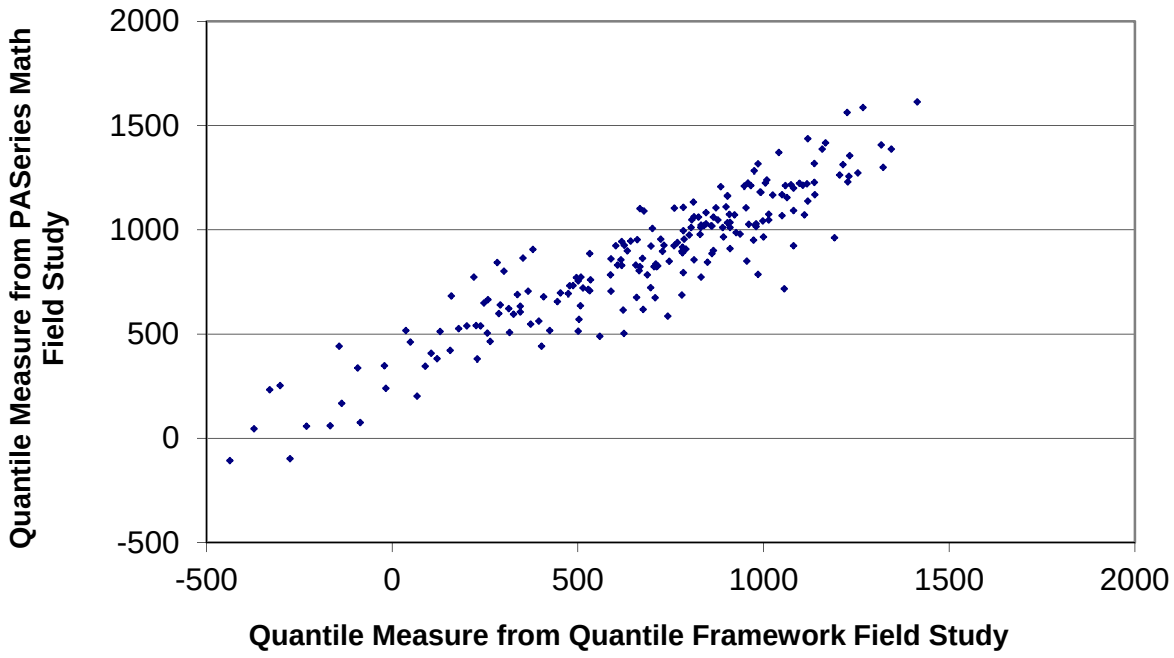
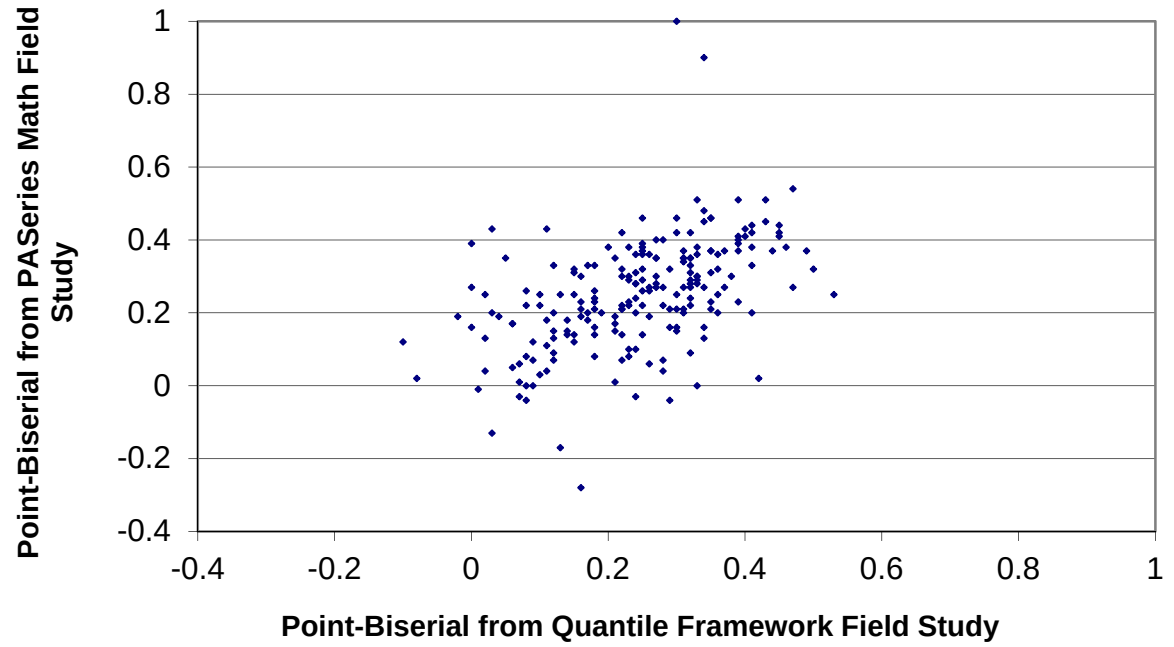


Figure 8. Comparison of the point-biserial correlations of the standardization set of items across two field studies.



The *PASeries* Math field study included 23,987 students who provided their grade level. *Table 7* shows the descriptive statistics for the sample by grade level. A monotonically increasing Quantile measure is observed across the grade levels.

Table 7. Mean and median Quantile measures for students with complete data from the *PASeries* Math field study (*N* = 23,987).

Grade Level	*N*	Mean uantile measure	Median uantile measure
3	4,703	370.46	370
4	4,478	592.29	598
5	3,871	696.54	690
6	2,813	788.32	771
7	3,555	827.24	816
8	3,481	884.81	874
9	1,086	970.24	967

Relationship of Quantile Measures to other Measures of Mathematical Ability. Scores from tests purporting to measure the same construct, for example "mathematical ability," should be moderately correlated (Anastasi, 1982). *Table 8* presents the results from field studies conducted with The Quantile Framework for Mathematics. For each of the tests listed, student mathematics scores were correlated with Quantile measures from the Quantile Framework field study.

Table 8. Results from studies conducted with The Quantile Framework for Mathematics.

Standardized Test	Grades in Study	*N*	Correlation Between Test Score and uantile measure
RIT and Measures of Academic Progress (MAP by NWEA)	4 5	94	0.69
North Carolina End-of-Grade Tests (Mathematics)	4 5	341	0.73

Quantile Framework Linked to other Measures of Mathematics Understanding. The Quantile Framework for Mathematics has been linked to several standardized tests of mathematics achievement. When assessment scales are linked, a common frame of reference can be used to interpret the test results. This frame of reference can be "used to convey additional normative information, test-content information, and information

that is jointly normative and content-based. For many test uses … [this frame of reference] conveys information that is more crucial than the information conveyed by the primary score scale" (Petersen, Kolen, and Hoover, 1989, p. 222).

Table 9 presents the results from linking studies conducted with the Quantile Framework. For each of the tests listed, student mathematics scores can also be reported as Quantile measures. This dual reporting provides a rich, criterion-related frame of reference for interpreting the standardized test scores. When a student takes one of the standardized tests, in addition to receiving her or his norm-referenced test results, s/he can receive information related to the specific QSCs that s/he is ready to receive instruction.

Table 9. Results from linking studies conducted with the Quantile Framework.

Standardized Test	Grades in Study	*N*	Correlation Between Test Score and uantile measure
Mississippi Curriculum Test, Mathematics (MCT)	2 8	7,039	0.89
TerraNova (CTB/McGraw-Hill)	3, 5, 7, 9	6,356	0.92
Texas Assessment of nowledge and Skills (TA S)	3 11	14,286	0.69 to 0.78
Proficiency Assessments for Wyoming Students (PAWS)	3, 5, 8, and 11	3,923	0.87
Progress Towards Standards (PTS3)	3-8 and 10	8,544	0.86 to 0.90
Progress in Maths (PiM GL Assessments)	1 8	3,183	0.71 to 0.81
North Carolina End-of-Grade/End-of-Course Tests (NC EOG/NC EOC)	3, 5, 7, A1, G, and A2	5,069	0.88 to 0.90
entucky Core Content Tests (CCT)	3 - 8 and 11	12,660	0.80 to 0.83
Oklahoma Core Competency Tests (OCCT)	3 8	5,649	0.81 to 0.85
Iowa Assessments	2, 4, 6, 8, and 10	7,365	0.92
ReadiStep (The College Board)	8	2,183	0.83
Virginia Standards of Learning (SOL)	3-8, A1, G, and A2	12,470	0.86 to 0.89
entucky Performance Rating for Educational Progress (-PREP)	3 - 8	6,859	0.81 to 0.85

Notes: * TAKS, PTS3, PiM, NCEOC, KCCT, OCCT, K-PREP, and SOL were not vertically scaled; separate linking equations were derived for each grade/course.

Multidimensionality of the Quantile Framework. Test dimensionality is defined as the minimum number of abilities or constructs measured by a set of test items. A construct is a theoretical representation of an underlying trait, concept, attribute, process, and/or structure that a test purports to measure (Messick, 1993). A test can be considered to measure one latent trait, construct, or ability (in which case it is called unidimensional); or a combination of abilities (in which case it is referred to as multidimensional). The dimensional structure of a test is intricately tied to the purpose and definition of the construct to be measured. It is also an important factor in many of the model(s) used in data analyses. Though many of the models assume unidimensionality, this assumption cannot be strictly met because there are always other cognitive, personality, and test-taking factors that have some level of impact on test performance (Hambleton and Swaminathan, 1985).

Study 1 – Comparison of Mathematics with Reading. The multidimensionality of the Quantile scale was examined using the Principal Components Analysis of Residuals in Winsteps (PRCOMP=S). The items were renamed with the strand number first for ease in review of the output. A three-step process was undertaken in order to examine the results and provide a context for interpreting the results.

The first step in the process was to run the Principal Components Analysis on all Quantile Framework field study items (N = 898). Next, the residual matrix was factor analyzed. *Table 10* shows the output from the analysis. The variance that is unexplained by the first factor (the Rasch measurement model) is 0.2% of the residual variance or 2.5 items of information. Based upon this set of data, it cannot be concluded that mathematics achievement as measured by the Quantile scale is multidimensional. The results supported the use of a unidimensional item response model on the items.

Table 10. Principal components analysis and distribution of variance explained by the model with the Quantile Framework field-study mathematics items (N = 685).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	1327.4	100.0	100.0
Variance Explained by Measures	642.4	48.4	49.9
Unexplained Variance (Total)	685.0	51.6	50.1
Unexplained Variance Explained by 1st Factor of the Residual Matrix	2.5	0.2	

Next, the items were ordered by factor loading. Based on an examination of the item names with strand listed first, there did not appear to be any effect of strand. Only 6 items out of the 685 unique items had loadings above 0.30 on the first residual factor. These six items were all level 10 (Geometry) items and were from both strands 2 (Geometry) and 3 (Algebra).

To better understand the values produced in the first analysis, a second analysis was undertaken. The Level 5 (Grade 5) Quantile items were analyzed separately. The results are presented in *Table 11*.

Table 11. Principal components analysis and distribution of variance explained by the model with the Grade 5 Quantile Framework field-study mathematics items (N = 65).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	118.1	100.0	100.0
Variance Explained by Measures	53.1	45.0	45.9
Unexplained Variance (Total)	65.0	55.0	54.1
Unexplained Variance Explained by 1st Factor of the Residual Matrix	1.8	1.5	

Three examples in the research literature describe the investigation of reading as a unidimensional construct: the 1940s Davis Study (Davis, 1944; Thurstone, 1946), the 1970s Anchor Study (Rentz and Bashaw, 1975, 1977; Jaeger, 1973; Loret, Seder, Bianchini, and Vale, 1974), and five 1980s and 1990s studies examining research conducted by ETS (Kirsch & Jungeblut and their colleagues, 1993, 1994; Reder, 1996; Salganik & Tal, 1989; Zwick, 1987). Other more recent examples include Harvey Goldstein's research with PISA (November 17, 2003), research on the development of the North Carolina End-of-Grade Tests (NCDPI, 1996), and research with the 2003 Maryland School Assessment—Reading. All of the studies confirm the assumption of unidimensionality of the reading assessments. Since most research concludes that reading is a unidimensional construct, for comparison purposes, a set of reading grade 5 reading items was also analyzed. The results are presented in *Table 12*.

Table 12. Principal components analysis and distribution of variance explained by the model with Grade 5 reading comprehension items (*N* = 54).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	137.1	100.0	100.0
Variance Explained by Measures	83.1	60.6	62.1
Unexplained Variance (Total)	54.0	39.4	37.9
Unexplained Variance Explained by 1st Factor of the Residual Matrix	2.0	1.5	

The Rasch model explains 60.6% of the variance in the reading comprehension items. Along with the results presented in *Tables 11* and *12*, these data are consistent with the use of a unidimensional item response theory model for each of the analyses (reading and mathematics).

Finally, items from strands 2 (geometry) and 3 (algebra) were analyzed. It was hypothesized, that if multi-dimensionality were to be evidenced in the data, this would be the most likely contrast. The Winsteps analysis using all 296 of the strand 2 and 3 items in all of the forms did not appear to have any connectivity (common item) problems.

Table 13. Principal components analysis and distribution of variance explained by the model with the Strand 2 and 3 Quantile Framework field-study mathematics items (*N* = 296).

Source	Standardized Residual Variance (in Eigenvalue units)	Empirical	Modeled
Total Variance in Observations	644.7	100.0	100.0
Variance Explained by Measures	348.7	54.1	55.5
Unexplained Variance (Total)	296.0	45.9	44.5
Unexplained Variance Explained by 1st Factor of the Residual Matrix	2.3	0.4	

Given the larger number of items in the analyses (296 in *Table 13* compared to 65 when only the Grade 5 items were examined in *Table 11*), the Rasch model explains 54.1% of the variance in the geometry (strand 2) and algebra (strand 3) items. The results presented in *Tables 10* and *11* are consistent with the interpretation of a single construct for each of the analyses (reading and mathematics).

Study 2 – Burg 2007. A study conducted by Burg (2007) analyzed the dimensional structure of mathematical achievement tests aligned to the NCTM content standards. Since there is no consensus within the measurement community on a single method to determine dimensionality, Burg employed four different methods for assessing dimensionality: (1) exploring the conditional covariances (DETECT), (2) assessment of essential unidimensionality (DIMTEST), (3) item factor analysis (NOHARM), and (4) principal component analysis (WINSTEPS). All four approaches have been shown to be effective indices of dimensional structure. Burg analyzed Grades 3 through 8 data from the Quantile Framework field study previously described.

Each set of on-grade items for a test form from Grades 3 through 8 were analyzed for possible sources of dimensionality related to the five mathematical content strands. The analyses were also used to compare test structures across grades. The results indicated that although mathematical achievement tests for Grades 3 through 8 are complex and exhibit some multidimensionality, the sources of dimensionality are not related to the content strands. The complexity of the data structure, along with the known overlap of mathematical skills, suggests that mathematical achievement tests could represent a fundamentally unidimensional construct. Therefore, while these sub-domains of mathematics are useful for organizing instruction, developing curricular materials such as textbooks, and describing the organization of items on assessments, they do not

describe a significant psychometric property of the test or impact the interpretation of the test results. Mathematics, as measured by the Quantile Framework, can be described as one construct with various sub-domains.

Furthermore, these findings support the NCTM Connections Standard, which states that all students (prekindergarten through Grade 12) should be able to make and use connections among mathematical ideas and see how the mathematical ideas interconnect. Mathematics can be best described as an interconnection of overlapping skills with a high degree of correlation across the mathematical topics, skills, and strands.

Study 3 – Hennings and Simpson 2012. Results from Hennings and Simpson (2012) also suggest that the mathematics assessments used in MetaMetrics' linking studies are functionally unidimensional. Data from a Quantile Framework linking study involving the end-of-grade tests from a Southeastern state was examined. Scored student responses to items on the combined Quantile Linking Test and the state end-of-grade test were used. The end-of-grade tests had three polytomous items worth two points each on the forms for Grades 3 through 8, and one polytomous item worth four points on the forms for Grades 4 through 8. The remaining items on both tests were dichotomous and scored 0/1. *Table 14* shows the number of students and the number of items, combined and by test, for each grade.

Table 14. Number of items included in analyses

Grade	N of Students	uantile Linking Test	End-of-Grade Test	Total
3	897	40	47	87
4	1,161	42	48	90
5	1,029	46	48	94
6	1,327	44	48	92
7	1,475	43	48	91
8	933	47	48	95

The polychoric item correlation matrix was analyzed for each test and grade. Because the principal components method of factor extraction in SAS does not require a positive-definite correlation matrix as input, principal component analyses were conducted instead of factor analyses.

The results support treating the data as unidimensional. The first component was dominant in all analyses. The first eigenvalue accounted for greater than 20% of the

total variance in the analyses. Ratios of first-to-second eigenvalues ranged from approximately 6 to slightly over 9 (Gorsuch, 1983; Reckase, 1979). Secondary dimensions, i.e., the second and third components, accounted for approximately 5 - 6.5% of the total variance for each grade. *Table 15* lists the eignevalues for the first five principal components by grade, *Table 16* shows the ratios of first-to-second eigenvalues, and *Table 17* shows the proportion of variance accounted for by the first five principal components for each grade.

Table 15. Eigenvalues for the first five principal components.

	Principal omponents				
rade					
3	24.152	3.463	2.411	2.253	2.011
4	23.252	3.637	2.257	1.894	1.829
5	22.770	3.222	2.407	2.239	1.935
6	21.400	3.058	2.297	2.185	1.866
7	23.919	3.922	2.442	1.744	1.648
8	24.572	2.654	2.152	2.076	1.914

Table 16. Ratio of the first-to-second eigenvalues by grade.

rade	**Ratio**
3	6.975
4	6.394
5	7.066
6	6.997
7	6.099
8	9.257

Table 17. Proportion of variance explained for the first five principal components by grade.

	Principal omponents				
rade					
3	0.278	0.040	0.028	0.026	0.023
4	0.258	0.040	0.025	0.021	0.020
5	0.242	0.034	0.026	0.024	0.021
6	0.233	0.033	0.025	0.024	0.020
7	0.263	0.043	0.027	0.019	0.018
8	0.259	0.028	0.023	0.022	0.020

The NC READY EOG Mathematics/EOC Algebra I/Integrated I - Quantile Framework Linking Process

Description of the Assessments

North Carolina READY EOG Mathematics and EOC Algebra I/Integrated I Assessments. North Carolina READY EOG Mathematics and EOC Algebra I/Integrated I Assessments measure students' proficiency based upon the Common Core State Standards for Mathematics (CCSSM) adopted by North Carolina in 2010. The EOG assessments are administered annually to students in Grades 3 through 8. The Algebra I/Integrated I assessment is administered at the end of the course to students enrolled in Algebra I or Integrated Math I. Each assessment consists of items that were written for specific content standards and demand one or more of the eight Standards for Mathematical Practice that are described in the CCSSM at every grade level (NCDPI, 2013c).

The NC Ready EOG Mathematics for Grades 3 and 4 consist of 54 items with 27 calculator inactive items and 27 calculator active items. The structure of the Grades 3 and 4 assessments consist entirely of multiple-choice items with four-response options. For the Grade 5 assessment, the calculator inactive section includes 19 multiple-choice and 8 gridded-response items and the calculator active section includes 27 multiple-choice items. For the NC Ready EOG Mathematics at Grades 6, 7, and 8, the calculator-inactive section consists of 7 multiple-choice and 11 gridded-response items that require students to insert numeric answers. The calculator-active section has 42 multiple-choice items (NCDPI, 2013e). The NC READY EOG Mathematics assessments were not vertically scaled across grades. Each test has scale scores that range from 400 to 500. These scale scores cannot be compared directly from grade to grade.

Since the CCSSM is subdivided into domains, which are large groups of related standards, the test items reflect a distinct distribution from each domain. The following table distinguishes these allocations at the identified grade levels (NCDPI, 2013c).

Table 18. Summary of the NC READY EOG Mathematics assessment blueprint targets for test development.

Domain	**rade**	**rade**	**rade**
Operations and Algebraic Thinking	30-35	12-17	5 10
Number and Operations in Base Ten	5-10	22-27	22 27
Number and Operations-Fractions	20-25	27-32	47 52
Measurement and Data	22-27	12-17	10 15
Geometry	10-15	12-17	2 7
Total			

Table 18 (continued). Summary of the NC READY EOG Mathematics assessment blueprint targets for test development.

Domain	**rade**	**rade**	**rade**
Ratios and Proportional Relationships	12-17	22 27	NA
The Number System	27-32	7 12	2-7
Expressions and Equations	27-32	22 27	27-32
Functions	NA	NA	22-27
Geometry	12-17	22 27	20-25
Statistics and Probability	7-12	12 17	15-20
Total			

The NC READY EOC Algebra I/Integrated I contains 60 items with approximately 80% four-choice multiple-choice items and 20% gridded-response items that require students to insert numeric answers (NCDPI, 2013e). Ten of the NC READY EOC Algebra I/Integrated items are embedded into the test as field-test items. Each of the remaining 50 items count as one point toward the student score. The NC READY EOC Algebra I/Integrated I scale scores range from 200 to 300, and these scale scores are on a separate scale.

At the high school course level, the CCSSM categorizes the standards by conceptual categories rather than by a set of standards for each course. As a result, states have the option to determine their own sequence of the CCSSM with the intention of completing the entire set of CCSSM standards by the end of the third year of high school study. *Table 19* shows the distribution of the high school conceptual categories for the NC READY EOC Algebra I/Integrated I assessment.

Table 19. Conceptual category distributions for Algebra I/Integrated I EOC.

onceptual ategory	**Algebra I Integrated I**
Number and uantity	5-10
Algebra	22-27
Functions	35-40
Geometry	10-15
Statistics and Probability	15-20
Total	

Assessment results will be used both for school and district accountability under the NC READY Accountability Model and for Federal reporting purposes (NCDPI, 2013c).

The Quantile Framework for Mathematics. The Quantile Framework was developed to assist teachers, parents, and students in identifying strengths and weaknesses in mathematics and forecast growth in overall mathematical achievement. Items and mathematical content are calibrated using the Rasch IRT model. The Quantile scale ranges from "EM" (Emerging Mathematician, 0Q and below) to above 1600Q. The Quantile Framework was developed to assess how well a student (1) understands the natural language of mathematics, (2) knows how to read mathematical expressions and employ algorithms to solve decontextualized problems, and, (3) knows why conceptual and procedural knowledge is important and how and when to apply it. The Quantile Framework Item Bank consists of multiple-choice items aligned with first grade content through Geometry, Algebra II, and Pre-calculus content and field tested with a national sample of students during the winter of 2004.

The Quantile Linking Test was constructed by aligning the items from the NC READY EOG Mathematics assessments for grades 3, 4, 6, and 8 with the Quantile Framework taxonomy of Quantile Skills and Concepts (QSCs). Based upon these target test reviews, previously tested items were used to develop each grade-level linking test. Each Quantile Linking Test reflects comparable material that is tested at each identified grade level of the NC READY EOG Mathematics. The Quantile Linking Tests for Grades 3 and 4 have 44 items (rather than the 54 items on the NC READY EOG Mathematics assessments) because of the 10 field-test items included in the NC READY EOG Mathematics assessments. The Quantile Linking Tests for Grades 6 and 8 have 50 items (rather that the 60 items on the NC READY EOG Mathematics assessments) because of the 10 field-test items included in the NC READY EOG Mathematics assessments.

The items used for the linking tests predominantly match the QSCs that were identified for each item in the target test. When an exact QSC match did not occur, the linking test used a different QSC that satisfied one or more of the following conditions:

1. The test item used a QSC that addressed the same North Carolina Core Standard as the target item.
2. The test item used a QSC that was a prerequisite to the matched QSC in the target test.
3. The test item was more appropriate for grade level or student expectations based on North Carolina Core Standards.

The Quantile Linking Tests for Grades 3 and 4 consisted of 44 multiple-choice items. The distribution of the content strands closely matched the distribution of the North Carolina Core domains for each grade level.

Table 20. Distributions of content strands Grades 3 and 4 Quantile Linking Tests.

ontent Strand	**rade**		**rade**	
	Percent of Items	Number of Items	Percent of Items	Number of Items
Numbers and Operations	56	25	63	28
Geometry	5	2	18	8
Algebra/Patterns Functions	14	6	5	2
Data Analysis Probability	9	4	5	2
Measurement	16	7	9	4
Total				

The Grade 3 Quantile Linking Test consisted of 9 calculator-inactive items and 35 calculator-active items. The Grade 4 test consisted of 11 calculator-inactive items and 33 calculator-active items.

The content of these tests did not require a reference sheet with formulas. In addition, no ancillary materials such as rulers or protractors were necessary. Calculators that are suggested for student use on this test were a four-function calculator that did ***not*** include the fraction key. Calculators were provided by the student or the school district for this assessment administration.

The Quantile Linking Tests for Grades 6 and 8 consisted of 50 multiple-choice items. The distribution of the content strands closely matched the distribution of the domains from the North Carolina Core standards.

Table 21. Distributions of content strands for Grades 6 and 8 Quantile Linking Tests.

ontent Strand	**rade**		**rade**	
	Percent of Items	Number of Items	Percent of Items	Number of Items
Numbers and Operations	52	26	16	8
Geometry	4	2	20	10
Algebra/Patterns Functions	18	9	48	24
Data Analysis Probability	14	7	12	6
Measurement	12	6	4	2
Total				

None of the items on the Grades 6 and 8 Quantile Linking Tests required ancillary materials or tools such as protractors, rulers, or compasses. These Quantile Linking Tests did include a formula sheet as a reference point for students to determine the formula necessary to solve a problem. Calculators were to be used only during the calculator-active sections of the linking tests. Grade 6 students could use a four-function or scientific calculator; and it was advisable to use the calculators they were accustom to

using during instruction, but use must abide by the North Carolina restrictions for calculators. Grade 8 students could use a graphing calculator that is within the North Carolina calculator requirements. Calculators were provided by the students or by the school district.

The Algebra I/Integrated I Quantile Linking Test consisted of 50 items. The distribution of the content strands closely matched the distribution of the Conceptual Categories distribution based upon the alignment study of the NC Ready EOC Algebra I/Integrated I with the Quantile Framework taxonomy.

Table 22. Distributions of content strands Quantile Linking Test Algebra I/Integrated I.

ontent Strands	**Algebra I Integrated I**	**Number o Items**
Numbers and Operations	10	5
Geometry	6	3
Algebra/Patterns Functions	62	31
Data Analysis Probability	16	8
Measurement	6	3
Total		

The Grade 3 linking test had 5 items in common with the Grade 4 linking test. The Grade 4 linking test had 12 items in common with one or more grade levels of Quantile Linking Tests. The Grades 6 and 8 linking tests each had approximately 12 items linked to one or more grade levels of the Quantile Linking Tests. The Algebra I/Integrated I EOC assessment had 11 items linked to Grade 8 and one of those items was also linked to Grade 6. These linked items were used to develop a continuum in the vertical scale for measuring student growth.

Each Quantile Linking Test had a mean Quantile measure that aligned with the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments content (Grade 3, 408Q; Grade 4, 626Q; Grade 6, 783Q; Grade 8, 965Q; and Algebra I/Integrated I, 1047Q). To the extent possible, the grade level at which each item on the Quantile Linking Test was initially calibrated matched the grade level of the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments. An exception to this guideline occurred when an item was to be used as an across-grade linking item and was selected from a higher or lower grade level.

Evaluation of the Quantile Linking Tests. After administration, the Quantile Linking Tests items were reviewed. The raw score descriptive statistics for all items and all students that took the Quantile Linking Tests are presented in *Table 23*.

Table 23. Descriptive statistics for the Quantile Linking Tests raw scores.

Grade/Course	*N**	Raw Score Mean (SD)	Minimum Score		Maximum Score	
			Observed	Possible	Observed	Possible
3	2,109	30.92 (8.1)	0	0	44	44
4	2,201	25.72 (7.9)	0	0	44	44
6	2,310	28.58 (9.6)	0	0	48	49
8	1,916	27.56 (8.7)	3	0	49	50
Alg I/Int I	2,538	24.88 (9.5)	1	0	49	50
Total	11,074					

* N-size reflects the removal of 142 students for missing, unusable, or duplicate students.
** One item was removed from Grade 6.

Based on the item examination, one item was removed from the Grade 6 analysis, because of a printing error in the test booklet. Selected item statistics for the Quantile Linking Tests are presented in *Table 24*. While some items retained on the tests had low point-biserial correlations, the items performed adequately (average ability measure for the correct answer was highest compared to the average ability measures of the three distractors from Winsteps analyses).

Table 24. Item statistics from the development of the Quantile Linking Tests.

Grade/Course	*N** (Persons)	*N*** (Items)	Percent Correct Mean (Range)	Point-Biserial Range	Coefficient Alpha
3	2,109	44	70 (35 - 96)	0.17 - 0.61	0.900
4	2,201	44	58 (10 - 95)	0.10 - 0.55	0.882
6	2,310	49	58 (18 - 94)	0.11 - 0.57	0.905
8	1,916	50	55 (14 - 91)	0.03 - 0.49	0.875
Alg I/Int I	2,538	50	50 (14 - 84)	0.13 - 0.50	0.898
Total	11,074				

* N-size reflects the removal of 142 students for missing, unusable, or duplicate students.
** One item was removed from Grade 6.

Coefficient Alphas for each of the five Quantile Linking Tests, one for each grade/course, ranged from 0.875 to 0.905. These values indicate strong internal consistency reliability for each of the five tests and high consistency across the five tests.

Study Design

A single-group/common person design was chosen for this study (Kolen and Brennen, 2004). This design is most useful "when (1) administering two forms to examinees is operationally possible, (2) differential order effects are not expected to occur, and (3) it is difficult to obtain participation of a sufficient number of examinees in an equating study that uses the random groups design" (pp. 16–17). The Quantile Linking Tests were administered between April 29, 2013 and May 15, 2013, within two weeks of the administration of the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments.

Analysis of the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment/Quantile Linking Test Sample

The sample of students for the study was identified by the North Carolina Department of Public Instruction. The participating schools were located from across North Carolina with a total of 120 schools from 61 districts participating in the linking study.

Table 25 presents the number of students tested in the linking study and the percentage of students with complete data (both a NC READY EOG Mathematics/EOC Algebra I/Integrated I scale score and a Quantile Linking Test Quantile measure). A total of 10,903 students (Grades 3, 4, 6, 8, and Algebra I/Integrated I), or 98.9%, had both test scores. This sample will be referred to as the matched sample.

Table 25. Number of students sampled and number of students in the matched sample.

Grade/Course	NC READY EOG Math/EOC *N* Received	uantile Linking Test *N*	Matched *N*	Matched Percent
3	104,035	2,090	2,069	99.0
4	111,463	2,197	2,181	99.3
6	112,688	2,308	2,283	98.9
8	109,639	1,901	1,868	98.3
Alg I/Int I	119,717	2,531	2,502	98.9
Total	557,542	11,027	10,903	98.9

All students and items were submitted to a Winsteps (Linacre, 2011) analysis using a logit convergence criterion of 0.0001 and a residual convergence criterion of 0.003.

To account for individual differences in motivation when responding to the two assessments, the sample set was trimmed. By grade, test scores from each of the assessments were rank ordered and then converted to percentiles. For each student, the difference in percentiles between the two assessments was examined. A screen of a 25-percentile-point difference was selected for all tests. This helped to minimize the number of students removed from the sample and maintain the characteristics of the distribution, while at the same time removing students that were obvious outliers on one or both of the assessments.

For the final sample of students used in the study, students in the matched sample with the following score patterns were removed:

- Accommodations that effect the construct being measured
  - AssistiveTechnology
  - Cranmer Abacus
- 100% correct on the Quantile Linking Test,
- Missing total score on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment,
- Misfit to the Rasch model, or
- Showed greater than a 25-percentile-rank difference between the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and Quantile Linking Test Quantile measures within grade.

Table 26 shows, for each grade, the number of students (*N*) in the final sample and the percent each grade *N*-count represents of the original matched sample. Of the 10,903

students in the matched sample, 8,720 (80%) remained in the final sample. The table also summarizes the number of students (by grade) removed from analysis, and the reason for their removal.

Table 26. Comparison of matched sample and final sample and the reason for student removal.

Matched Sample		*N* Removed by Reason				Final Sample	
Grade/ Course	*N*	Accommodated Students	Misfit to Rasch	Scores	Percentile Rank Difference	*N*	Percent of Matched Sample
3	2,069	2	97	15	251	1,704	82.4
4	2,181	4	177	5	280	1,715	78.6
6	2,283	2	24	1	376	1,880	82.3
8	1,868	0	22	0	340	1,506	80.6
Alg I/Int I	2,502	0	40	0	547	1,915	76.5
Total	10,903	8	360	21	1,794	8,720	80.0

* Note: Students with a 100% correct on the linking test or with an invalid NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment score.

Table 27 presents the demographic characteristics of all students in the NC READY EOG Mathematics/EOC Algebra I/Integrated I state sample, the matched sample, and the final sample of students included in this study. The three samples are very similar.

Table 27. Percentage of students in the NC READY EOG Mathematics/EOC Algebra I/Integrated I state sample, the matched sample, and the final sample for selected demographic characteristics.

Student Characteristic	Category	State Sample *N*=557,542	Matched Sample *N*=10,903	Final Sample *N*=8,720
Grade or Course	3	18.7	19.0	19.5
	4	20.0	20.0	19.7
	6	20.2	20.9	21.6
	8	19.7	17.1	17.3
	Alg I/Int I	21.5	22.9	22.0
Gender	Female	49.3	49.6	49.8
	Male	50.6	50.4	50.2
	Unknown/not avail	0.1	0.0	0.0
Race/Ethnicity	American Indian	1.5	0.9	1.0
	Asian	2.7	2.4	2.4
	Black	25.5	28.6	27.9
	Hispanic	13.9	14.9	14.6
	Pacific Islander	0.1	0.1	0.1
	White	52.6	49.5	50.5
	Two or more	3.7	3.5	3.6
	N/A	0.1	0.1	0.1
LEP Status	Currently identified	6.2	6.7	6.8
	Exit by committee	0.0	0.0	0.0
	Exits LEP	5.1	5.7	5.7
	Never identified	88.5	87.5	87.4
	No Status	0.1	0.1	0.1
	Parental refusal of IPT testing	0.0	0.0	0.0
Student/Disability	Exited within 2 years	2.0	1.6	1.5
	Yes	9.6	9.7	9.8
	No	88.5	88.8	88.7

Student Characteristic	Category	State Sample *N*=557,542	Matched Sample *N*=10,903	Final Sample *N*=8,720
EC Code	Autism	0.5	0.4	0.4
	Deaf-Blindness	0.0	0.0	0.0
	Deafness	0.0	0.0	0.0
	Developmental Delay	0.1	0.1	0.0
	Hearing Impairment	0.1	0.1	0.1
	Intell. Disability - Mild	0.2	0.1	0.1
	Intell. Disability - Moderate	0.0	0.0	0.0
	Intell. Disability - Severe	0.0	0.0	0.0
	Orthopedic Impairment	0.0	0.0	0.0
	Other Health Impairment	2.3	2.6	2.6
	Serious Emotional Disability	0.4	0.3	0.3
	Specific Learning Disability	5.5	5.5	5.6
	Speech or Language Impairment	2.3	2.1	2.1
	Traumatic Brain Injury	0.0	0.0	0.0
	VI	0.0	0.0	0.0
	Multiple Disabilities	0.0	0.0	0.0
	Not Provided	88.5	88.8	88.7
Plan 504	Yes	1.2	1.0	1.0
	No	98.8	99.0	99.0
Word To Word Bilingual	Yes	0.0	0.0	0.0
	No	100.0	100.0	100.0
Acad/Intell Gifted - Reading	Yes	11.9	12.0	12.8
	No	88.1	88.0	87.2

Table 28 presents the descriptive statistics for the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale score matched sample as well as the matched sample Quantile Linking Test Quantile measure. Evaluating the Quantile measures on the NC

READY EOG Mathematics/EOC Algebra I/Integrated I assessments and the Quantile Linking Tests show very comparable results. The correlations between the matched sample NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and the matched sample Quantile measures range between 0.726 and 0.815. Based upon these correlations, it can be concluded that the two tests are measuring similar mathematics constructs.

Table 28. Descriptive statistics for the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and Quantile measures and the Quantile Linking Test, matched sample (*N* = 10,903).

Grade/ Course	*N*	Matched Sample NC READY EOG Mathematics/EOC Algebra I/Integrated I Scale Score Mean (SD)	Matched Sample uantile Linking Test uantile Measure Mean (SD)	*r*
3	2,069	449.55 (9.5)	641.96 (228,6)	0.815
4	2,181	449.01 (9.4)	718.73 (203.0)	0.794
6	2,283	449.40 (9.4)	866.78 (204.9)	0.797
8	1,868	447.93 (8.4)	1003.70 (183.3)	0.777
Alg I/Int I	2,502	251.65 (9.7)	1040.62 (202.5)	0.726
Total	10,903			

Table 29 presents the descriptive statistics of the final sample NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment scale scores and the Quantile Linking Test Quantile measures. The correlations between the two scores range from 0.872 to 0.900. These correlations between the two scores are strong and higher than the matched sample.

Table 29. Descriptive statistics for the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and the Quantile Linking Test Quantile measures, final sample (*N* = 8,720).

Grade/ Course	*N*	Final Sample NC READY EOG Mathematics/EOC Algebra I/Integrated I Scale Score Mean (SD)	Final Sample uantile Linking Test uantile Measure Mean (SD)	*r*
3	1,704	449.21 (9.6)	637.16 (218.0)	0.900
4	1,715	449.74 (9.3)	738.67 (192.9)	0.890
6	1,880	449.82 (9.6)	884.12 (204.6)	0.896
8	1,506	448.36 (8.6)	1018.31 (187.0)	0.893
Alg I/Int I	1,915	251.90 (9.8)	1057.98 (205.3)	0.872
Total	8,720			

Figures 9 through *18* shows the relationship between the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and the Quantile Linking Test Quantile measures for the matched and final samples for each grade/course. The matched samples show more scatter than the final samples. In each grade/course, it can be seen that there is a linear relationship between the NC READY EOG Mathematics/ EOC Algebra I/Integrated I scale scores and the final sample Quantile measures reinforcing the use of linear equating.

Figure 9. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 3 matched sample (*N* = 2,069).

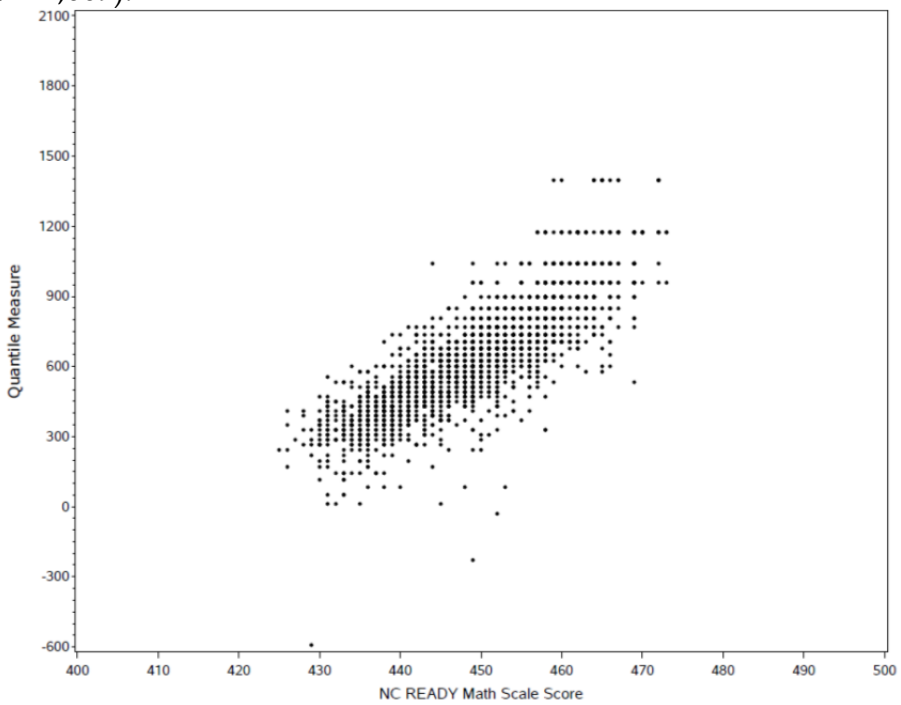


Figure 10. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 3 final sample (*N* = 1,704).

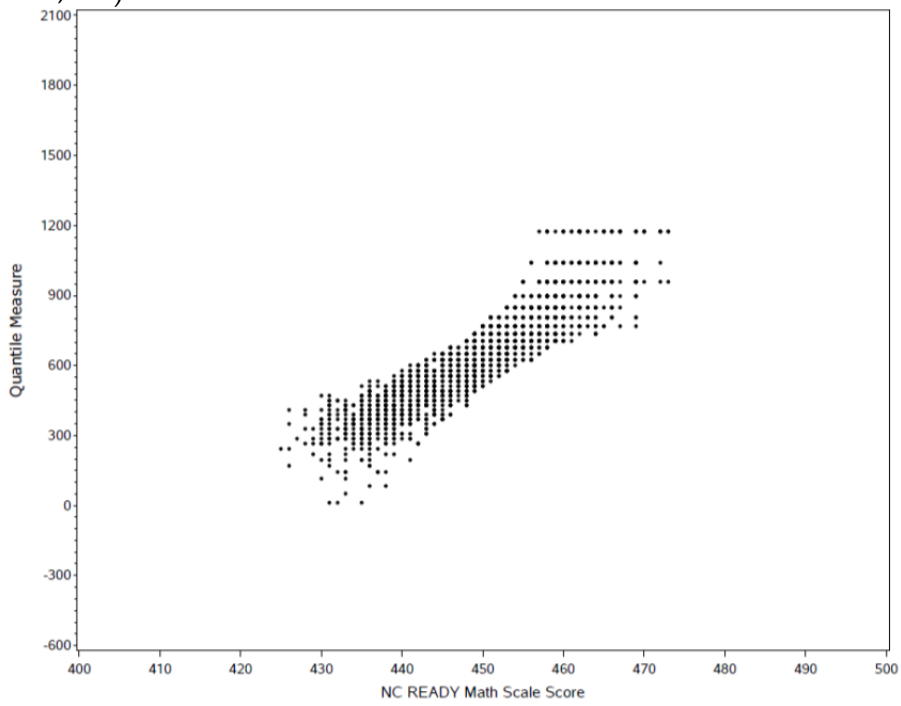


Figure 11. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 4 matched sample (*N* = 2,181).

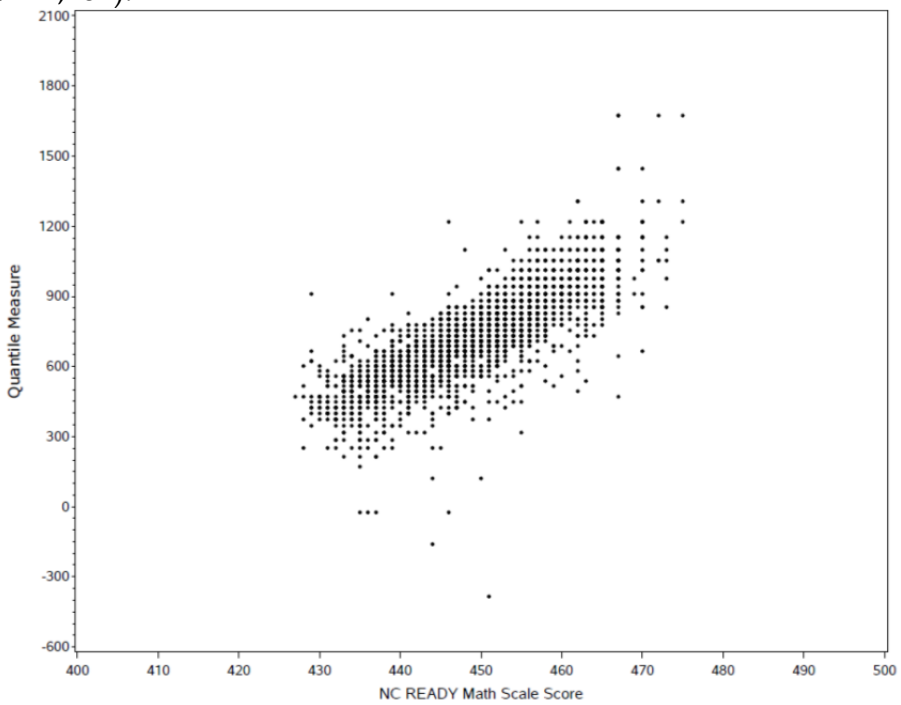


Figure 12. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 4 final sample (*N* = 1,715).

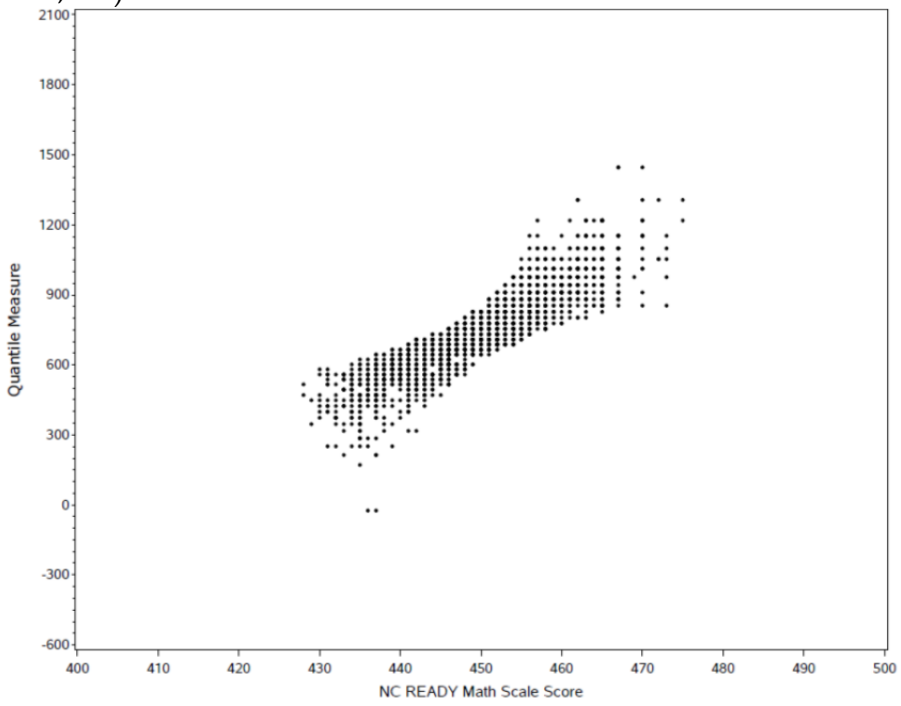


Figure 13. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 6 matched sample (*N* = 2,283).

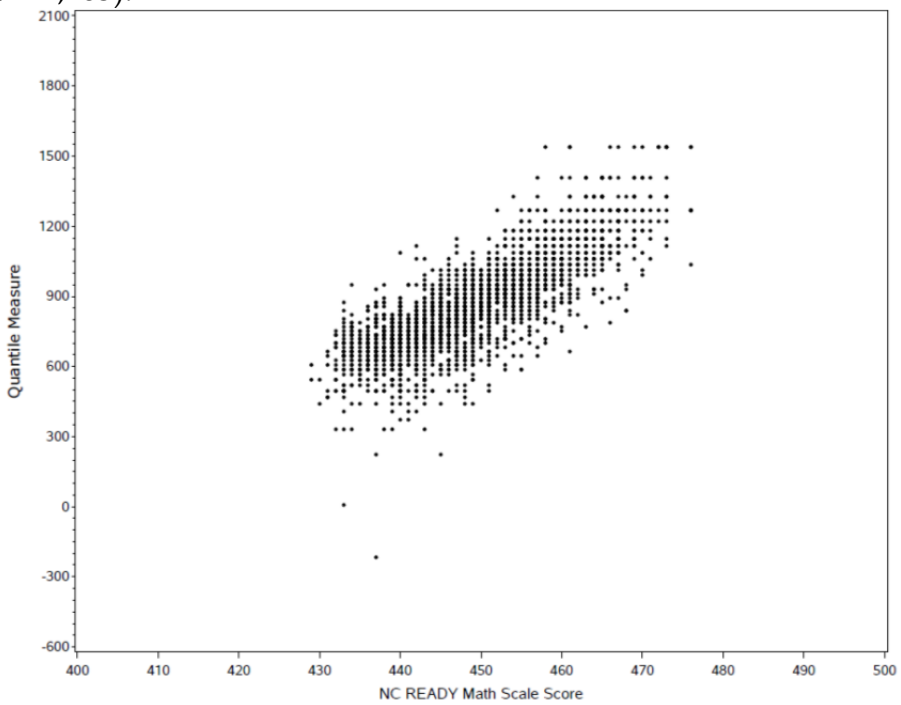


Figure 14. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 6 final sample (*N* = 1,880).

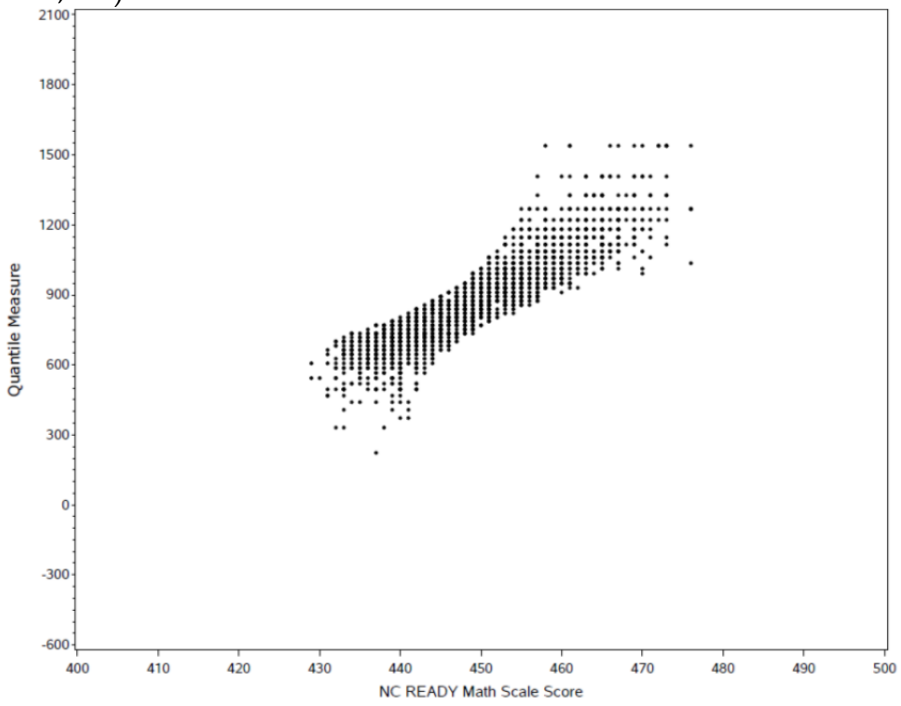


Figure 15. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 8 matched sample (*N* = 1,868).

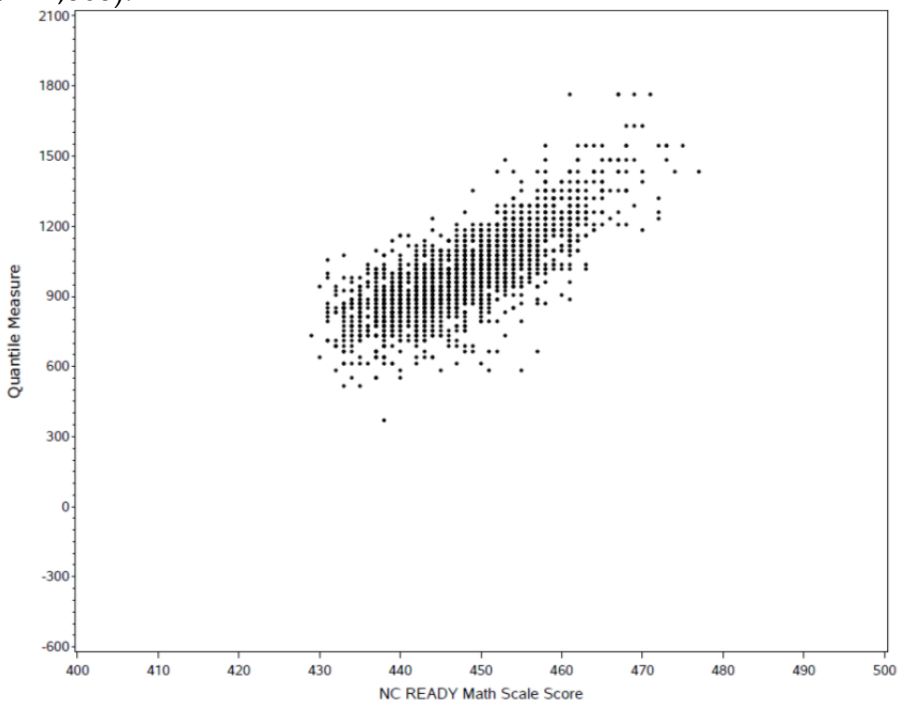


Figure 16. Scatter plot of the NC READY EOG Mathematics scale scores and the Quantile Linking Test Quantile measures for the Grade 8 final sample (*N* = 1,506).

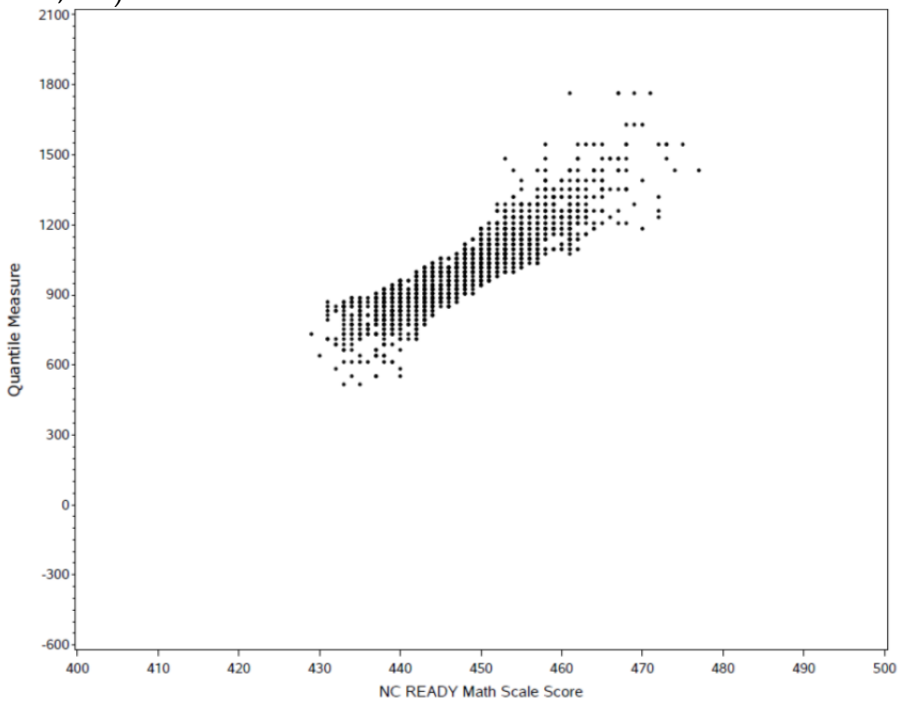


Figure 17. Scatter plot of the NC READY EOG EOC Algebra I/Integrated I scale scores and the Quantile Linking Test Quantile measures for the Algebra I/Integrated I matched sample (*N* = 2,502).

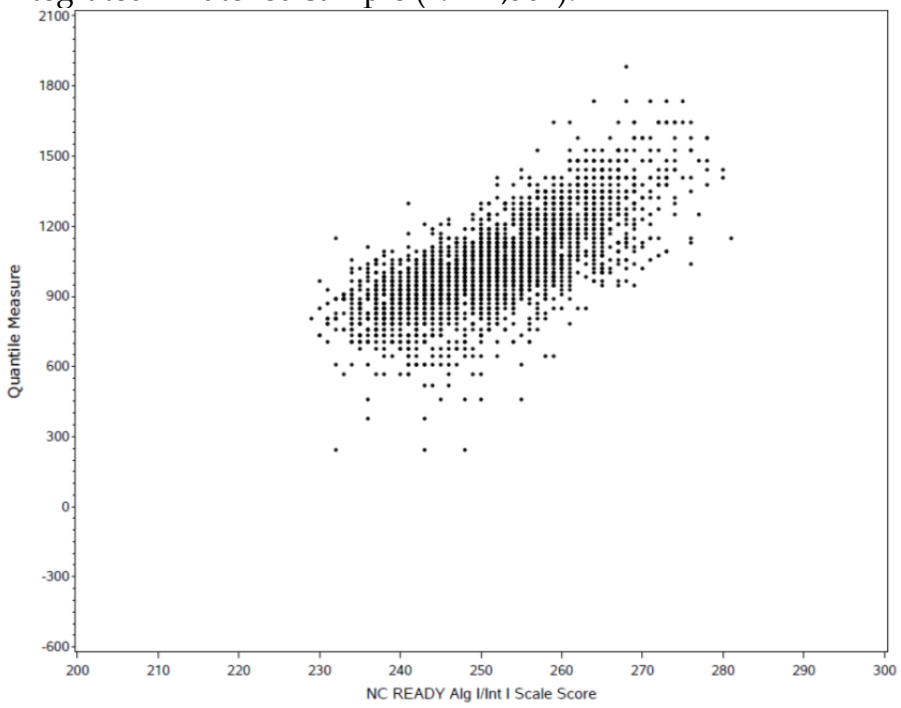
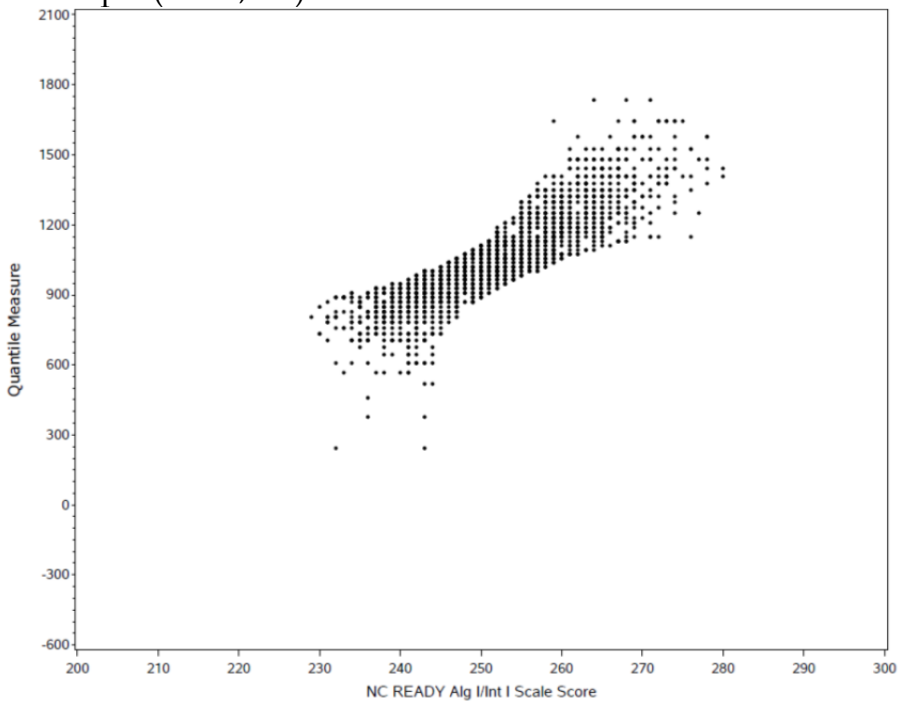


Figure 18. Scatter plot of the NC READY EOC Algebra I/Integrated I scale scores and the Quantile Linking Test Quantile measures for the Algebra I/Integrated I final sample (*N* = 1,915).



Linking the NC READY EOG Mathematics/EOC Algebra I/Integrated I Scale with the Quantile Scale

Linking in general means "putting the scores from two or more tests on the same scale" (National Research Council, 1999, p.15). This study was designed to provide information that could be used to match students' mathematical achievement with instructional resources—to identify the materials, concepts, and skills a student should be matched with for successful mathematical instruction, given their performance on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments.

Linking Analyses. Two score scales (e.g., the Quantile Scale and the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment scales) can be linked using linear equating when the underlying item response models used to develop assessments are different. The linear equating method is most appropriate when (1) sample sizes are small; (2) test forms have similar difficulties; and (3) simplicity in conversion tables or equations, in conducting analyses, and in describing procedures are desired (Kolen and Brennan, 2004).

In linear equating, a transformation is chosen such that scores on two tests are considered to be equated if they correspond to the same number of standard deviations above (or below) the mean in some group of examinees (Angoff, 1984, cited in Petersen, Kohen, and Hoover, 1989; Kolen and Brennan, 2004). Given scores x and y on tests X and Y, the linear relationship is

$$\frac{(x-\mu_X)}{\sigma_X} = \frac{(y-\mu_y)}{\sigma_y} \qquad \text{(Equation 6)}$$

and the linear transformation l_x (called the SD line in this report) used to transform scores on test Y to scores on text X is

$$x = l_x(y) = \left(\frac{\sigma_X}{\sigma_y}\right)y + \left(\mu_x - \frac{\mu_y\sigma_X}{\sigma_y}\right) \qquad \text{(Equation 7)}$$

Linear equating using an SD-line approach is preferable to linear regression because the tests are not perfectly correlated. With less than perfectly reliable tests, linear regression is dependent on which way the regression is conducted: predicting scores on test X from scores on test Y or predicting scores on test Y from scores on test X. The SD line provides the symmetric linking function that is desired.

The final linking equation between the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and the Quantile scale can be written as:

Quantile measure = Slope(NC READY EOG Mathematics/EOC Algebra I/Integrated I scale score) + Intercept (Equation 8)

where the slope is the ratio of the standard deviations of the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and Quantile Linking Test Quantile measures. These values can be found in *Table 29*.

Using the final sample data described in *Table 29*, the linear linking functions relating the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores and Quantile measures for all students in the sample are presented in *Table 30*. Separate linking functions were developed for each grade/course of the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment since they are not on a vertical scale.

Because the original design for the NC READY mathematics assessments was to report results using a vertical scale across grades, no Quantile data was collected for Grades 5 and 7. During the calibration of the NC READY mathematics items for Grades 3 through 8 it was determined that a vertical scale could not be fitted (personal communication with NCDPI). Consequently, the Quantile measure equations needed to be estimated. Using a regression analysis, the Quantile means for Grades 5 and 7 were estimated using the means from the other grades' final samples. The standard deviations for Grades 5 and 7 were calculated using a pooled variance formula of the other grade's final sample data. The NC READY EOG Mathematics Grades 5 and 7 scale score means and standard deviations were calculated using the state data. The usual SD formulas for Grades 5 and 7 were derived using the means and standard deviations determined above.

Conversion tables were developed for each grade in order to express the NC READY EOG Mathematics/EOC Algebra I/Integrated I scores in the Quantile metric and were delivered to the North Carolina Department of Public Instruction in electronic format.

Table 30. Linear linking equation coefficients used to predict Quantile measures from the NC READY EOG Mathematics/EOC Algebra I/Integrated I scale scores.

Grade/Course	Slope	Intercept
3	22.740744	-9578.224
4	20.801171	-8616.395
5	21.092335	-8694.573
6	21.357151	-8722.812
7	20.836926	-8439.688
8	21.748657	-8733.002
Alg I/Int I	20.895137	-4205.586

Table 31 contains the capped Quantile measures by grade/course. The measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is instructional, then the scores should be capped at the upper bound of measurement error (e.g., at the 95th percentile point). In an instructional environment, all scores at or below 0Q should be reported as "EM" (Emerging Mathematician); no student should receive a negative Quantile measure.

Table 31. Capped values of the Quantile measure by grade/course.

Grade/Course	Capped uantile Measure
3	975
4	1075
5	1125
6	1200
7	1325
8	1450
Alg I/Int I	1475

Validity of the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment – Quantile Link

Table 32 presents the descriptive statistics and effect size statistics of the NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile measures as well as the Quantile Linking Test Quantile measures for the final sample.

Table 32. Descriptive statistics and effect size statistics for the final sample NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile measures and the Quantile Linking Test Quantile measures.

Grade	*N*	Final Sample NC READY EOG Mathematics/EOC Algebra I/Integrated I uantile Measure Mean (SD)	Final Sample uantile Linking Test uantile Measure Mean (SD)	Effect Size
3	1,704	637.16 (218.0)	637.15 (218.0)	0.000035
4	1,715	738.67 (192.9)	738.74 (192.9)	-0.000369
6	1,880	884.12 (204.6)	884.10 (204.6)	0.000099
8	1,506	1018.31 (187.0)	1018.30 (187.0)	0.000047
Alg I/Int I	1,915	1057.98 (205.3)	1057.99 (205.3)	-0.000035
Total	8,720			

The Hedges' *g* effect size shows the relationship between two variables or, in this case, between the NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile measure and the Quantile Linking Test Quantile measure. A guideline to use for interpretation of the effect size is:

Table 33. Interpretation chart for effect size.

Small	0.20
Medium	0.50
Large	0.80

For the five comparisons in *Table 32*, effect sizes were minimal for all comparisons indicating no significant difference between the NC READY EOG Mathematics/EOC

Algebra I/Integrated I Quantile measures and the Quantile Linking Test Quantile measures. This is because each grade/course has a unique linear equation.

Table 34 contains the percentile ranks of the Quantile Linking Test Quantile measures and the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment Quantile measures (based on the final sample). The criterion of a half standard deviation (100Q) on the Quantile scale was used to determine the size of the difference. In examining the values, the measures are very similar across the distributions. This supports the use of Quantile measures on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments.

Table 34. Comparison of the Quantile measures for selected percentile ranks for the final sample Quantile Linking Test and the NC READY EOG Matematics/EOC Algebra I/Integrated I assessment.

Grade 3		
Percentile Rank	Linking Test uantile Measure	NC READY EOG Math Sample uantile Measure
1	170	200
5	308	269
10	369	337
25	470	473
50	624	655
75	806	814
90	958	928
95	1040	973
99	1174	1087

Grade 4		
Percentile Rank	Linking Test uantile Measure	NC READY EOG Math Sample uantile Measure
1	286	349
5	448	432
10	516	474
25	601	599
50	731	744
75	854	869
90	975	994
95	1053	1056
99	1219	1160

Table 34 (continued). Comparison of the Quantile measures for selected percentile ranks for the final sample Quantile Linking Test and the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment.

Grade 6		
Percentile Rank	Linking Test uantile Measure	NC READY EOG Math Sample uantile Measure
1	468	503
5	586	568
10	645	610
25	735	717
50	874	888
75	1013	1037
90	1146	1166
95	1268	1251
99	1407	1358

Grade 8		
Percentile Rank	Linking Test uantile Measure	NC READY EOG Math Sample uantile Measure
1	639	662
5	732	728
10	793	793
25	887	880
50	1017	1010
75	1138	1141
90	1259	1271
95	1353	1315
99	1545	1467

Algebra I/Integrated I		
Percentile Rank	Linking Test uantile Measure	NC READY EOC Math Sample uantile Measure
1	608	642
5	758	726
10	827	788
25	908	893
50	1038	1060
75	1188	1206
90	1348	1332
95	1408	1394
99	1577	1520

Performance standards provide a common meaning of test scores throughout a state or nation concerning what is expected at various levels of competence. The North Carolina Department of Public Instruction established four achievement levels: Level 1, Level 2,

Level 3, and Level 4. As an example, the four achievement levels for the Grade 3 NC READY EOG Mathematics Assessment are (NCDPI, 2013b):

Level 1: Students performing at this level have **limited command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at grade 3 and are likely to need intensive academic support to engage successfully in further studies in this content area.

Level 2: Students performing at this level have **partial command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at grade 3 and are likely to need additional academic support to engage successfully in further studies in this content area.

Level 3: Students performing at this level have **solid command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at grade 3 and are academically prepared to engage successfully in further studies in this content area.

Level 4: Students performing at this level have **superior command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at grade 3 and are academically well prepared to engage successfully in further studies in this content area.

The four achievement levels for NC READY EOC Algebra I/Integrated I Assessment are (NCDPI, 2013a):

Level 1: Students performing at this level have a **limited command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at the end of Math I and will need academic support to engage successfully in more rigorous studies in this content area. They will also need continued academic support to become prepared to engage successfully in credit-bearing, first-year Mathematics courses without the need for remediation.

Level 2: Students performing at this level have a **partial command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at the end of Math I and will likely need academic support to engage successfully in more rigorous studies in this content area. They will also likely need continued academic support to become prepared to engage successfully in credit-bearing, first-year Mathematics courses without the need for remediation.

Level 3: Students performing at this level have **solid command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at the end of Math I and are academically prepared to engage successfully in more rigorous studies in this content area. They are also on track to become academically prepared to engage successfully in credit-bearing, first-year Mathematics courses without the need for remediation.

Level 4: Students performing at this level have a **superior command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) for Mathematics assessed at the end of Math I and are academically well-prepared to engage successfully in more rigorous studies in this content area. They are also on-track to become academically prepared to engage successfully in credit-bearing, first-year Mathematics courses without the need for remediation.

Table 35 presents the achievement level cut scores on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments and the associated Quantile measures. The values in the table are the cut scores associated with the bottom score for each category.

Table 35. Performance level cut scores on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment and the associated Quantile measures.

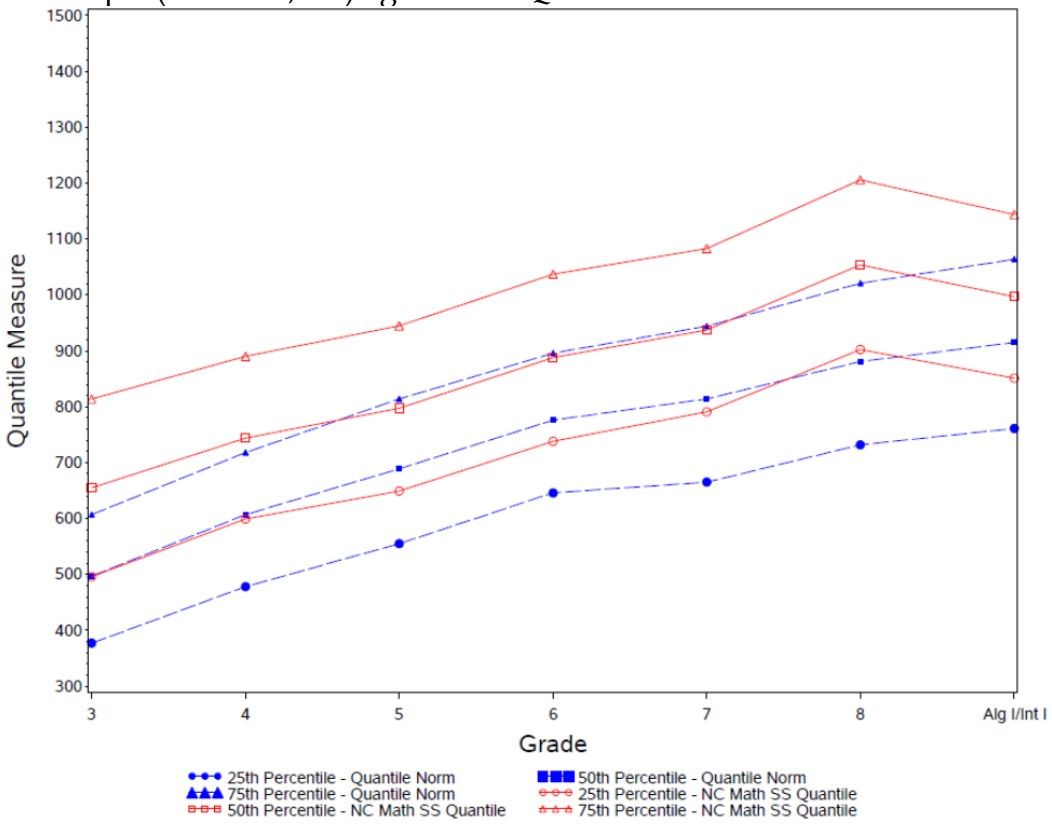
Grade/ Course	Level 2		Level 3		Level 4	
	NC READY EOG Mathematics/ EOC Algebra I/Integrated I Scale Score	uantile Measure	NC READY EOG Mathematics/ EOC Algebra I/Integrated I Scale Score	uantile Measure	NC READY EOG Mathematics/ EOC Algebra I/Integrated I Scale Score	uantile Measure
3	443	495	451	680	460	885
4	444	620	451	765	460	950
5	444	670	451	820	460	1010
6	447	825	453	950	461	1125
7	447	875	453	1000	461	1165
8	447	990	454	1140	463	1335
Alg I/Int I	247	955	253	1080	264	1310

The next graph shows the Quantile measures for the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments Quantile measures from the final sample and the Quantile norms. These norms were created based on linking studies conducted with the Quantile Framework. The sample's distribution of scores from this study was similar to the distribution of scores on norm-referenced assessments and other standardized measures of mathematics achievement. The results compared favorably with other mathematics measures which reinforced MetaMetrics' confidence in the Quantile norms.

As can be seen in *Figure 19*, the Quantile measures for the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments are higher than the Quantile

measure norms. This indicates that the final sample in this study is more able than the samples used for the Quantile norms.

Figure 19. Selected Percentiles (25th, 50th, and 75th) plotted for the NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile measures for the final sample (*N* = N=8,720) against the Quantile measure norms.



The following box and whisker plots (*Figures 20, 21, 22,* and *23*) show the progression of scores (the *y*-axis) from grade to grade (the *x*-axis). (Note: Alg I/Int I is presented as Grade 9.) For each grade, the box refers to the interquartile range. The line within the box indicates the median and the • indicates the mean. The end of each whisker shows the minimum and maximum values of the Quantile Linking Tests Quantile measures and the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments Quantile measures for each grade (the *y*-axis). The Quantile measures are on a vertical scale and *Figures 20, 21, 22,* and *23* demonstrate this by showing that as the grade increases so do the Quantile scores on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments. The pattern of Quantile measures is the same for each figure. *Figure 23* includes the performance levels of Level 2, Level 3, and Level 4 set by North Carolina.

Figure 20. Box and whisker plot of the Quantile Linking Tests Quantile measures by grade/course, final sample (*N* =8,720).

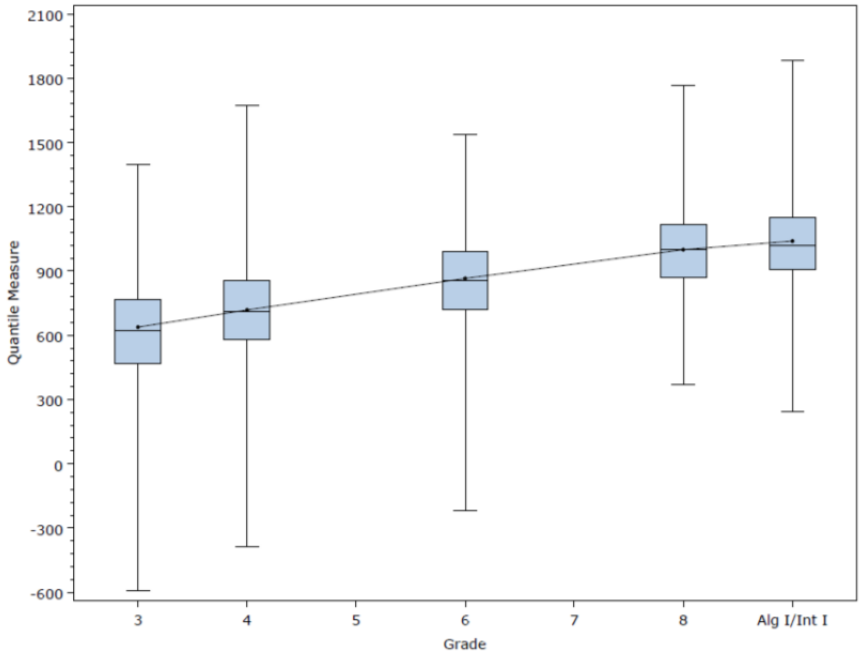


Figure 21. Box and whisker plot of the NC READY EOG Mathematics/EOC Algebra I/Integrated I Quantile measures by grade/course, matched sample (*N* = 10,903).

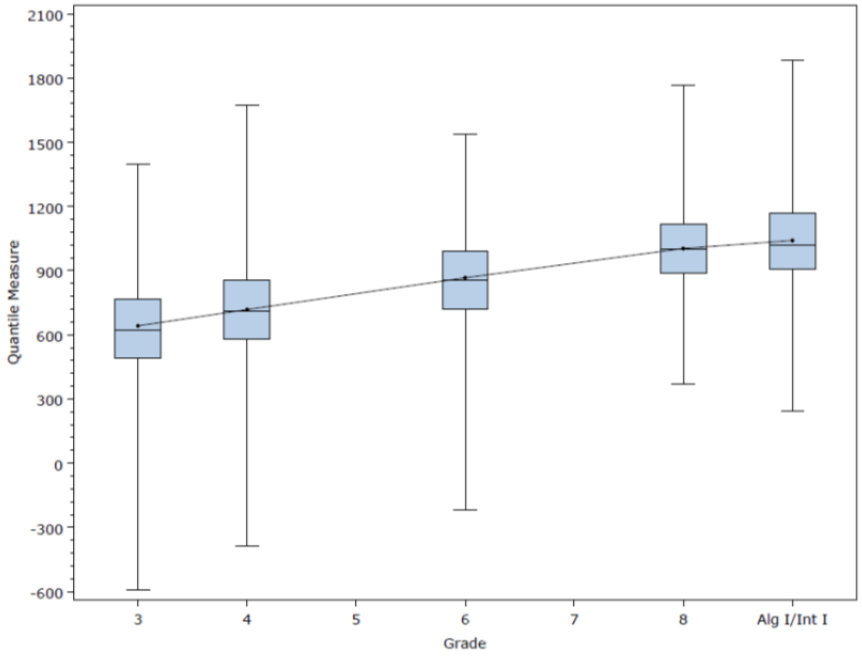


Figure 22. Box and whisker plot of the NC READY EOG Mathematics /EOC Algebra I/Integrated I Quantile measures by grade/course, final sample (*N* = 8,720).

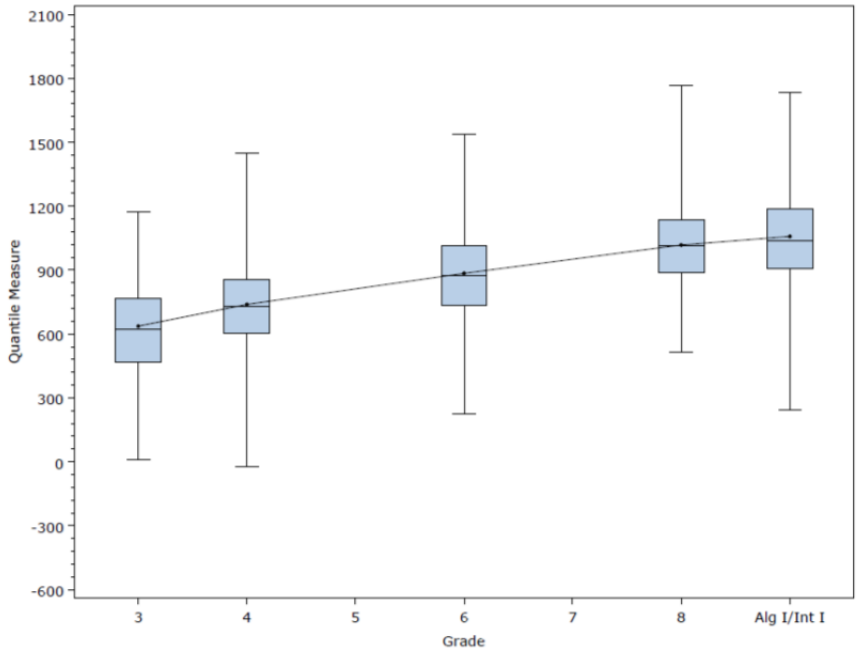
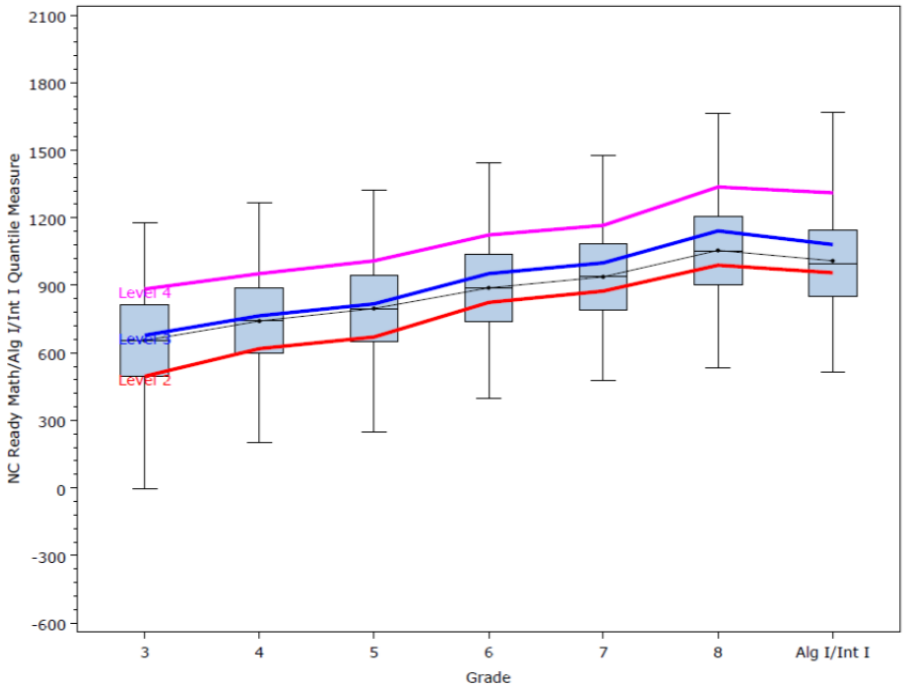


Figure 23. Box and whisker plot of the NC READY EOG Mathematics /EOC Algebra I/Integrated I Quantile measures with the performance standards by grade/course, state sample (*N* = 780,377).



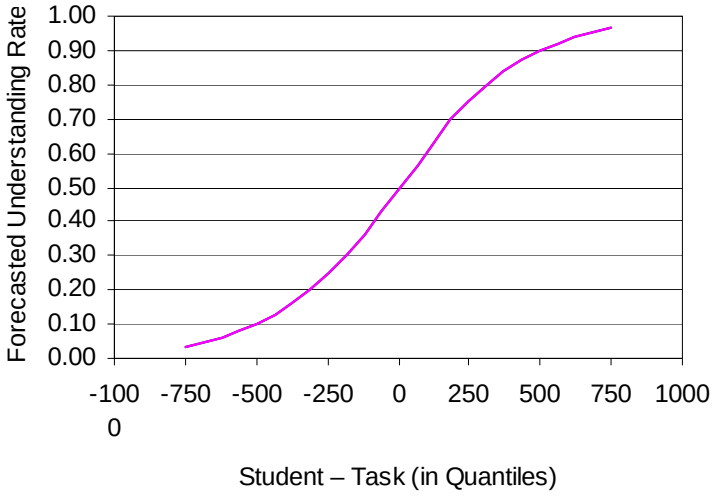
Quantile Framework and Instruction

Quantile measures are available from many norm-referenced and criterion-referenced assessments, in addition to state tests and instructional products. Students who take a mathematics achievement test that is linked with the Quantile Framework or one that reports directly in the Quantile metric will receive a Quantile measure. Educators can use these Quantile measures to match students, by *readiness level*, to level-appropriate instructional materials and forecast understanding. For example, a student with a Quantile measure of 500Q should be ready for instruction of mathematics problems at a demand level of 500Q.

Differentiated Instruction. A Quantile measure for materials is a number indicating the mathematical demand of the material in terms of the concept/application solvability. The Quantile measure for an individual student is the level at which he or she is ready for instruction (50% competency with the material) and has knowledge of the prerequisite mathematical concepts and skills necessary to succeed. The Quantile scale ranges from Emerging Mathematician (0Q and below) to above 1600Q. The Quantile measure does not relate to a specific grade, per se, so the score is developmental as it spans the mathematics continuum from kindergarten mathematics through the content typically taught in Algebra II, Geometry, Trigonometry, and Pre-calculus. The measure can be used by a teacher to determine what mathematical instruction the student is likely to be ready for next.

Figure 24 shows the general relationship between the student-task discrepancy and forecasted understanding. When the student measure and the task mathematical demand are the same (difference of 0Q), then the forecasted understanding, or success rate, is modeled as 50% and the student is likely ready for instruction on the skill or concept.

Figure 24. Relationship between student mathematical demand discrepancy and forecasted understanding (success rate).



An appropriate instructional range for the Quantile measure of a student is 50Q above and 50Q below the Quantile measure of the student (44% - 56% competency). This range identifies the "learning frontier" of mathematics skills in which a student has the prerequisite knowledge and skills needed to understand the instruction and will likely have success with tasks related to the skill/concept after this introductory instruction.

Quantile measures provide reliable, actionable results because instruction and assessment are described using the same metric. When instruction is measured at a unique mathematical level of understanding and any form of assessment can be reported using the same scale, equal levels of achievement are observed.

By understanding the interaction between student measures and resource measures (e.g., textbook lessons, instructional materials), any level of understanding can be used as a benchmark. An individual can modulate his or her own likely success rate by lowering the difficulty of the task (i.e., increase to 90% understanding) or increasing the difficulty of the task (i.e., lower to 40% understanding) depending on the situation (refer to *Figure 14*). This flexibility allows the teacher, parent, or student the ultimate control to modulate the fit between person and task.

The primary utility of the Quantile Framework is its ability to forecast what will likely happen when students confront resources and instruction on specific mathematical skills and concepts. With every application by teacher, student, or parent there is a test of the framework's accuracy. The framework makes a point prediction every time a resource or lesson is chosen for a student. Anecdotal evidence suggests that the Quantile Framework predicts as intended. That is not to say that there is an absence of error in forecasted understanding. There is error in resource measures based on QSC

(mathematical skills and concepts) measures, student measures, and their difference modeled as forecasted understanding. However, the error is sufficiently small that the judgments about students, resources, and understanding rates are useful.

The subjective experience of 25%, 50%, and 75% understanding/success as reported by students varies greatly. A 1000Q student being instructed on 1000Q QSCs (50% understanding) has a successful instructional experience—he has the background knowledge needed to learn and apply the new information. Teachers working with such a student report that the student can engage with the skills and concepts that are the focus of the instruction and, as a result of the instruction, are able to solve problems utilizing those skills. In short, such students appear to understand what they are learning. A 1000Q student being instructed on 1200Q QSCs (25% understanding) encounters so many unfamiliar skills and difficult concepts that the learning is frequently lost. Such students report frustration and seldom engage in instruction at this level of understanding. Finally, a 1000Q student being instructed on 800Q QSCs (75% understanding) reports that he is able to engage with the skills and concepts with minimal instruction, is able to solve complex problems related to the skills and concepts, is able to connect the skills and concepts with skills and concepts from other strands, and experiences fluency and automaticity of skills.

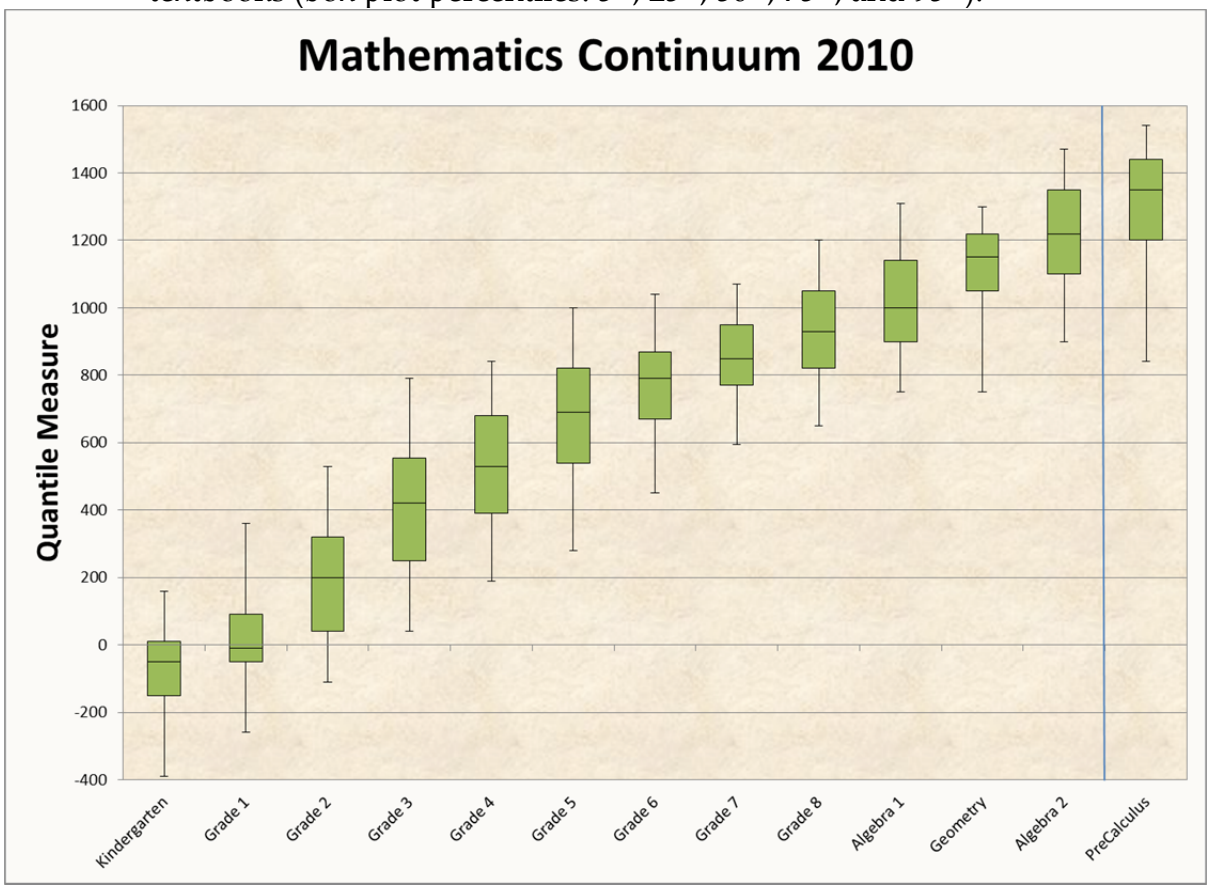
Quantile Framework and the CCSS. There is increasing recognition of the importance of bridging the gap that exists between K-12 and higher education and other postsecondary endeavors. Many state and policy leaders have formed task forces and policy committees such as P-20 councils. The Common Core State Standards (CCSS) for Mathematics were designed to enable all students to become college and career ready by the end of high school while acknowledging that students are on many different pathways to this goal: "One of the hallmarks of the Common Core State Standards for Mathematics is the specification of content that all students must study in order to be college and career ready. This 'college and career ready line' is a minimum for all students" (NGA Center & CCSSO, 2010b, p. 4). The CCSS for Mathematics suggest that "college and career ready" means completing a sequence that covers Algebra I, Geometry, and Algebra II (or equivalently, Integrated mathematics 1, 2 and 3) during the middle school and high school years; and, leads to a student's promotion into more advanced mathematics by their senior year. This has led some policy makers to generally equate the successful completion of Algebra II as a working definition of college and career ready. Exactly how and when this content must be covered is left to the states to designate in their implementations of the CCSS for Mathematics throughout K-12 (NGA Center & CCSSO, 2010a, p. 84).

The *mathematical demand* of a mathematical textbook (in Quantile measures) quantitatively defines the level of mathematical achievement that a student needs in order to be ready for instruction on the mathematical content of the textbook. Assigning QSC(s) and Quantile measures to a textbook is done through a calibration process.

Textbooks are analyzed at the lesson level and the calibrations are completed by subject matter experts (SMEs) experienced with the Quantile Framework and with the mathematics taught in mathematics classrooms. The intent of the calibration process is to determine the mathematical demand presented in the materials. Textbooks contain a variety of activities and lessons. In addition, some textbook lessons may include a variety of skills. Only one Quantile measure is calculated per lesson and is obtained through analyzing the Quantile measures of the QSCs that have been mapped to the lesson. This Quantile measure represents the composite task demand of the lesson.

MetaMetrics has calibrated more than 41,000 instructional materials (e.g., textbook lessons, instructional resources) across the K-12 mathematics curriculum (Smith and Turner, 2012). *Figure 25* shows the continuum of calibrated textbook lessons from Kindergarten through Pre-calculus where the median of the distribution for Pre-calculus is 1350Q. The range between the first quartile and the median of the first three chapters of Pre-calculus textbooks is from 1200Q to 1350Q. This range describes an initial estimate of the mathematical achievement level needed to be ready for mathematical instruction corresponding to the "college and career readiness" standard in the Common Core State Standards for Mathematics.

Figure 25. A continuum of mathematical demand for Kindergarten through Pre-calculus textbooks (box plot percentiles: 5th, 25th, 50th, 75th, and 95th).



This information describing college and career readiness in mathematics can be used to interpret the NC READY EOG Mathematics/EOC Algebra I/Integrated I performance standards. For each grade the "proficient" (Level 3) range of Quantile measures as defined by the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments is compared to the mathematical demands in the next grade/course. As can be seen in *Figure 26*, almost all students scoring at the "proficient" level should be prepared for the mathematical demands of the next grade/course. The Algebra I/Integrated I students at the proficient level are less ready for the next course work.

Figure 26. NC READY EOG Mathematics/EOC Algebra I/Integrated I "proficient" ranges (expressed as Quantile measures) compared with the mathematical demands of the next grade/course, by grade or course.

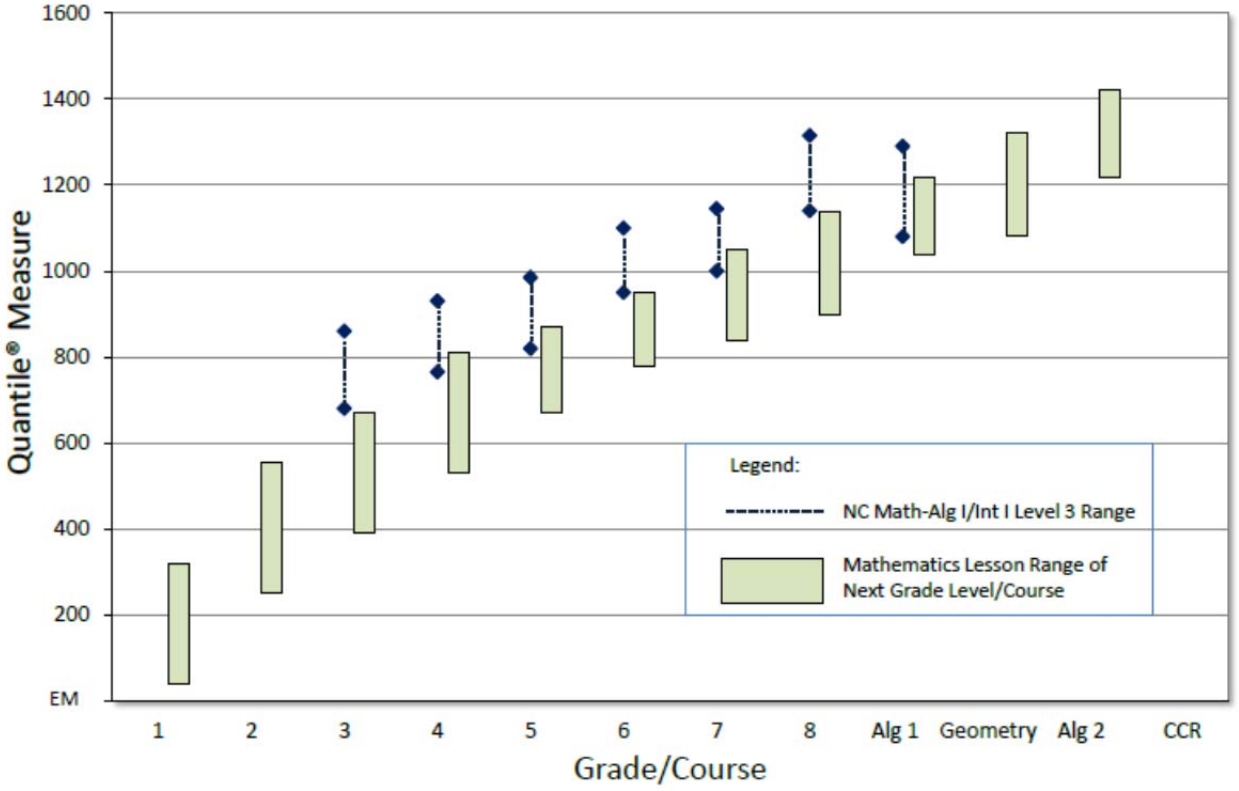
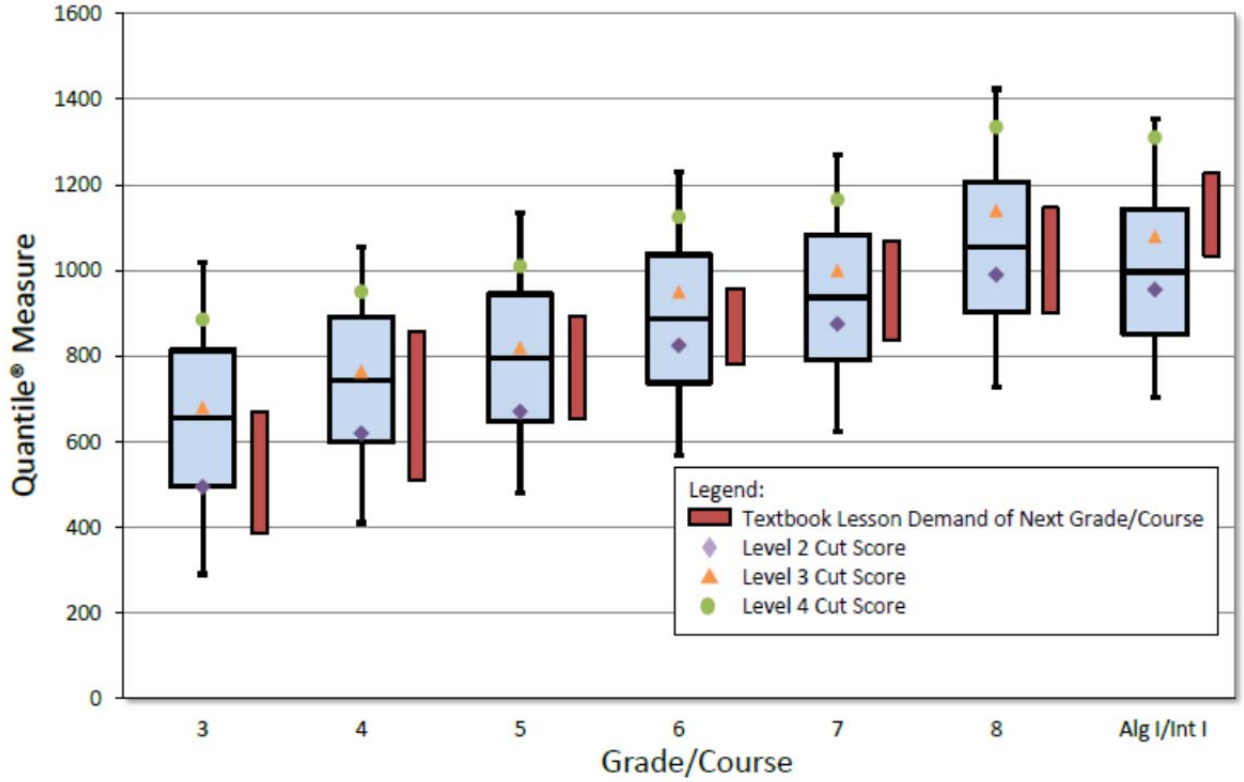


Figure 27 shows that the spring 2013 student performance on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments at each grade/course level is "on track" for college and career readiness in Grades 3 through 8. In comparing the performance of students in Algebra I/Integrated I, some students will need encouragement with supplemental materials at the next course. Students can be matched with mathematics materials that are at or above the recommendations in the Common Core State Standards for each grade/course.

Figure 27. NC READY EOG Mathematics/EOC Algebra I/Integrated I 2012-2013 student performance expressed as Quantile measures.



In 2009, MetaMetrics and the North Carolina Department of Public Instruction conducted a study to relink the NCEOG/EOC Mathematics Tests with the Quantile scale (MetaMetrics, 2010). The minimum score considered "proficient" (Level 3) at each grade level on the NCEOG/EOC Mathematics is presented in *Table 36*. In 2013, NCDPI transitioned their assessment program to the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment to align with the Common Core State Standards in Mathematics and to describe student mathematics performance in relation to college and career readiness. One outcome of this change was to set the performance standards for NC READY EOG Mathematics/EOC Algebra I/Integrated I at a higher level. For comparison purposes, the minimum "proficient" score for the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment is also repeated from *Table 35*. The Quantile scale can be used as an external "yardstick" to evaluate this change in the mathematical demand on the North Carolina Mathematics assessments. The information in *Table 36* shows that the NC READY EOG/EOC Mathematics standards are demanding more of students in terms of mathematical ability in 2013.

Table 36. Minimum "Level 3" Quantile measure on NCEOG/EOC Mathematics (2009) and NC READY EOG Mathematics/EOC Algebra I/Integrated I (2013).

Grade	Proficient Level 3 Cut Score (2009)	Proficient Level 3 Cut Score (2013)
3	515	680
4	645	765
5	775	820
6	795	950
7	860	1000
8	900	1140
Alg I/Int I	1020	1080

Conclusions, Caveats, and Recommendations

Forging a link between scales is a way to add value to one scale without having to administer an additional test. Value can be in the form of any or all of the following:

- increased *interpretability* (e.g., "Based on this test score, what mathematical skills and concepts does my child actually know?"),
- increased *diagnostic capability* (e.g., "Based on this test score, what are the student's weaknesses?"), or
- increased *instructional use* (e.g., "Based on these test scores, I need to modify my instruction to include these skills.").

The link that has been established between the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments and the Quantile Framework permits students to be matched with resources and materials that provide an appropriate level of challenge while avoiding frustration. The result of this purposeful match may be that students will be less fearful of mathematics, and, thereby become better mathematical thinkers. The real power of the Quantile Framework is in examining the growth in mathematical achievement of students—wherever the student may be in the development of his or her mathematical skills and concepts. Students can be matched with resources and materials for which they are forecasted to experience 50% understanding, therefore, they are ready for instruction on the topic. As a student's mathematical achievement grows, he or she can be matched with more demanding skills and concepts. And, as the skills and concepts become more demanding, then the student grows.

The development of the link between the scores on the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessments and the Quantile scale has been described and evaluated in this study. There are many factors that can affect the linking process. In this study two of the factors include:

- sample characteristics (e.g., gender, ethnicity), and
- relationship of sample distribution characteristics to the distribution characteristics of the state.

Conventions for Reporting. Quantile measures are reported as a number followed by a capital "Q" for "Quantile." There is no space between the measure and the "Q" and measures of 1,000 or greater are reported without a comma (e.g., 1050Q). All Quantile person measures should be rounded to the nearest 5Q to avoid over interpretation of the measures. As with any test score, uncertainty in the form of measurement error is present.

Next Steps. To utilize the results from this study, Quantile measures need to be incorporated into the NC READY EOG Mathematics/EOC Algebra I/Integrated I assessment results processing and interpretation frameworks. Suggested resources need to be developed for ranges of students. Care must be taken to ensure that the resources and materials on the lists are also developmentally appropriate for the students. The Quantile measure is one factor related to understanding and is a good starting point in the selection process of materials and resources for a specific student. Other factors such as student developmental level, motivation, and interest; amount of background knowledge possessed by the student; and characteristics of the resources and skills also need to be considered when matching resources and instruction with a student.

In this era of student-level accountability and high-stakes assessment, differentiated instruction—the attempt "on the part of classroom teachers to meet students where they are in the learning process and move them along as quickly and as far as possible in the context of a mixed-ability classroom" (Tomlinson, 1999)—is a means for all educators to help students succeed. Differentiated instruction promotes high-level and powerful curriculum for all students, but varies the level of teacher support, task complexity, pacing, and avenues to learning based on student readiness, interest, and learning profile. One strategy for managing a differentiated classroom suggested by Tomlinson is the use of multiple resources and supplementary materials that can be identified with the aid of the Quantile Framework. Equipped with a student's Quantile measure, teachers can connect him or her to textbook lessons, worksheets, games, websites, and trade books that have appropriate Quantile measures (Smith, no date; Smith and Turner, 2012). By incorporating Quantile measures into the planning of mathematics instruction, it becomes possible to forecast with greater probability how successfully students are likely to understand the material presented to them. Teachers can provide instruction on QSCs with Quantile measures below the targeted instruction when students are not ready for that instruction by focusing on prerequisite QSCs. On the other hand, teachers can focus enrichment activities on the impending QSCs.

Two resources are available on the Quantile Framework website – Quantile Teacher Assistant and Math@Home (Smith, no date; Smith and Turner, 2012). In order to support instruction with the many resources connected with the Quantile Framework, the Quantile Teacher Assistant (QTA) was developed to simplify and gather all relevant information. When using the QTA (http://qta.quantiles.com/), teachers can identify a specific state objective and determine the knowledge base. In addition, teachers can differentiate instruction by indicating the range of Quantile measures for their students in their classrooms. Math@Home (http://mah.quantiles.com/) activities reinforce mathematical skills covered in the previous school year and lay the groundwork for what will be taught when students return to class in the fall. By incorporating fun family games into everyday activities, students can practice mathematical skills year-round and parents can feel more confident about helping their children with mathematics.

MetaMetrics, in partnership with The Council of Chief State School Officers, has begun coordinating a national, state-led summer mathematics initiative to bolster student mathematics achievement during summer break. The Summer Math Challenge is designed to raise national awareness of the summer loss epidemic (Cooper, Nye, Charlton, Lindsay, and Greathouse, 1996), share compelling research on the importance of targeted mathematics activities, and provide access to a variety of free resources to support mathematics instruction and the initiative as a whole.

The 2013 "Summer Math Challenge" was a six-week, e-mail-based initiative designed to help students on summer vacation fight "summer slide" in mathematics skills. The initiative was designed to combat summer math slide by helping students retain mathematics skills acquired during the previous school year. The initiative targeted Grades 3 through 6 by reinforcing mathematics concepts presented from Grades 2 through 5 aligned with the Common Core State Standards. Participants received targeted instructional materials for a weekly concept along with personalized e-mail activity suggestions and resources that supported each concept. Twelve SEA chiefs requested assistance in launching a 2013 Summer Math initiative in conjunction with the CCSSO Chief's Summer Reading Challenge. North Carolina promoted the Summer Math Challenge through e-mail newsletters to educators. The "Chief's Summer Math Challenge" Flyer provides an overview of the CCSSO Chief's Math Challenge and MetaMetrics' 2013 Support to SEA leaders (URL: https://d1jt5u2s0h3gkt.cloudfront.net/m/cms_page_media/135/Chief's%20Summer%20Math%20Challenge%20Overview_2.pdf).

The following is a list of suggestions that can be used to leverage a student's Quantile measure in the classroom:

- Start class with warm-up problems and activities related to the prerequisite skills from a knowledge cluster.
- Enhance major themes of mathematics by building a bank of skills at varying levels that not only support a theme but also provide a way for all students to participate in the theme successfully. For example, consider how addition progresses from single numbers to multi-digit numbers, and then moves to decimals and fractions.
- Sequence mathematical skills according to their difficulty as much as possible.
- Develop a mathematics folder that goes home with students and returns weekly for review. The folder can contain examples of practice skills within a student's range, applications of topics outside the classroom, reports of recent assessments, and a parent form to record the amount of time spent working mathematics problems at home.

- Choose skills lower in a student's Quantile range when factors make the student view mathematics as more challenging, threatening, or unfamiliar. Select skills at or above a student's range to stimulate growth, when a topic holds high interest for a student, or when additional support such as background teaching or peer tutoring is provided.
- Develop individualized lists of skills that are tailored to provide appropriately challenging and curriculum suitable for all students.

Below are some suggestions related to leveraging a student's Quantile measure at home:

- Ensure that each child gets plenty of mathematical practice, concentrating on skills within his or her Quantile range. Parents can ask their child's teacher to print a list of appropriate skills or search the mathematics skill database on the Quantile website.
- Communicate with the child's teachers about the child's mathematical needs and accomplishments. They can use the Quantile scale to describe their assessment of the child's mathematical achievement.
- When a new topic proves too challenging for a child, use activities or other materials from the Web site to help. Review the prerequisite QSCs to ensure that gaps or misconceptions are not interfering with the current topic.
- Celebrate a child's mathematical accomplishments. The Quantile Framework provides an easy way for students to track their own growth. Parents and children can set goals for mathematics—spending so much time daily working on mathematical problems, discussing situational topics such as statistics from a newspaper or discounts at the store, reading a book about a mathematical topic, trying new kinds of Web sites and games, or working a certain number of mathematics problems per week. When children reach the goal, make it an occasion!

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1982). *Psychological Testing* (5th ed.). New York: MacMillan Publishing Company, Inc.

Bond, T.G. & Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Burg, S.S. (2007). *An investigation of dimensionality across grade levels and effects on vertical linking for elementary grade mathematics achievement tests*. University of North Carolina, Chapel Hill, NC.

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

Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs*. New York: John Wiley & Sons.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227-268.

Davis, F. (1944). Fundamental factors of comprehension in reading, *Psychometrika*, *9,* 185–197.

Dorans, N.J. & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Emenogu, B.C. & Childs, R.A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education*, *28*(1 & 2), 128-146.

Geary, D.C., & Hamson, C.O. (2002). Improving the mathematics and science of achievement of American children: Psychology's role. *American Psychological Association Online.* Retrieved May 23, 2002, from http://www.apa.org/ed/geary.html.

Goldstein, H. (2003). International comparisons of student attainment: Some issues arising from the PISA study. Retrieved December 14, 2004, from www.ioe.ac.uk/hgpersonal.

Gorsuch, R.L. (1983). *Factor analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Bulletin*. *9*, 139-164.

Hennings, S. S. & Simpson, M.A. (2012, October). *Quantile® Linking Test and a state assessment: How many dimensions?* Presentation at the 2012 semi-annual meeting of MetaMetrics' Technical Advisory Committee. Durham, NC: MetaMetrics, Inc.

Holland, P.W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Hudnutt, B. (2012). *Weaving mathematical connections from counting to calculus: Knowledge clusters and The Quantile® Framework for Mathematics*. Durham, NC: MetaMetrics, Inc.

Jaeger, R. M. (1973). The national test equating study in reading: The anchor test study. *Measurement in Education*, *4*, 1–8.

Kirsch, I. S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Kirsch, I. S., Jungeblut, A., & Mosenthal, P.B. (1994). *Moving toward the measurement of adult literacy*. Paper presented at the March NCES Meeting, Washington, DC.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. (2nd ed.) New York: Springer Science + Business Media, LLC.

Linacre, J.M. (2011). WINSTEPS (Version 3.73) [Computer Program]. Chicago: Author.

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

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. (1974). *Anchor test study final report: Project report* (Vols. 1–30). Berkeley, CA: Educational Testing Service. (ERIC Document Nos. ED 092 601-ED 092 631).

Messick, S. (1993). Validity. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan Publishing Company.

MetaMetrics, Inc. (October 29, 1998). *Linking*. Unpublished manuscript. Durham, NC: Author.

MetaMetrics, Inc. (2010). *Linking the North Carolina EOG/EOC Tests in Mathematics with the Quantile Framework*. Durham, NC: Author.

MetaMetrics, Inc. (2005). *PASeries* mathematics technical manual. Durham, NC: Author.

MetaMetrics, Inc. (2008). *The Quantile® Framework for Mathematics norms* [Unpublished normative data]. Durham, NC: Author.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Governors Association Center for Best Practices (NGA Center) & the Council of Chief State School Officers (CCSSO). (2010a). *Common core state standards for mathematics.* Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.

National Governors Association Center for Best Practices (NGA Center) & the Council of Chief State School Officers (CCSSO). (2010b). *Common core state standards for mathematics: Appendix A.* Retrieved from http://www.corestandards.org/assets/CCSSI_Mathematics_Appendix_A.pdf.

National Research Council. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, D.C.: National Academy Press.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Leaning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2002). *Helping children learn mathematics*. Mathematics Learning Study Committee, J. Kilpatrick and J. Swafford, Editors. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

North Carolina Department of Public Education. (2013a). *Achievement level descriptors for North Carolina end-of-course tests.* Retrieved on November 6, 2013 from http://sbepolicy.dpi.state.nc.us/Policies/GCS-C-036.asp?Acr=GCS&Cat=C&Pol=036

North Carolina Department of Public Education. (2013b). *Achievement level descriptors for North Carolina end-of-grade tests.* Retrieved on November 6, 2013 from http://sbepolicy.dpi.state.nc.us/Policies/GCS-C-033.asp?Acr=GCS&Cat=C&Pol=033

North Carolina Department of Public Instruction. (2013c). *Common core state standards (CCSS) for mathematics: North Carolina assessment specifications summary, READY EOG assessments, Grades 3-8, READY EOC algebra I/integrated I assessments.* Retrieved on October 31, 2013 from http://www.dpi.state.nc.us/docs/acre/assessment/math.pdf

North Carolina Department of Public Instruction. (2013d). *End-of-course assessment: Number of items and total test time (minutes)*. Retrieved on December 6, 2013 from http://www.ncpublicschools.org/docs/accountability/testing/eoc/eocadmininfo12.pdf

North Carolina Department of Public Instruction. (2013e). *End-of-grade assessment: Number of items and total test time (minutes)*. Retrieved on December 6, 2013 from http://www.ncpublicschools.org/docs/accountability/testing/eog/eogadmininfo13.pdf

North Carolina Department of Public Instruction. (2013f). *North Carolina READY end-of-course assessments*. Retrieved on October 24, 2013 from http://www.ncpublicschools.org/docs/accountability/policyoperations/assessbriefs/assessbriefeoc13.pdf

North Carolina Department of Public Instruction. (2013g). *North Carolina READY end-of-grade assessments*. Retrieved on October 24, 2013 from http://www.ncpublicschools.org/docs/accountability/policyoperations/assessbriefs/assessbriefeog13.pdf

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, Norming, and Equating. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed. pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

Poznanski, J.B. (1990). A meta-analytic approach to the estimation of item difficulties. Unpublished doctoral dissertation, Duke University, Durham, NC.

Price, L. R., Lurie, A., & Wilkins, C. (2001). EQUIPERCENT: A SAS program for calculating equivalent scores using the equipercentile method. *Applied Psychological Measurement, 25*, 332-341.

Rasch, G. (1980). *Probabilistic models for some intelligence and attachment tests*. Chicago: The University of Chicago Press. (First published in 1960).

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reder, S. (1998). Dimensionality and construct validity of the NALS assessment. In M.C. Smith (Ed.), *Literacy for the twenty-first century: Research, policy, practices and the National Adult Literacy Survey*. Westport, CT: Preager Publishing.

Rentz, R. R., & Bashaw, W. L. (1975). *Equating reading tests with the Rasch model* (Vol. 1-2). Athens, GA: University of Georgia, Educational Research Laboratory.

Rentz, R. R., & Bashaw, W. L. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement*, *14*, 161–179.

Roussos, L., Schnipke, D., & Pashley, P. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Behavioral and Educational Statistics*, 24, 293-322.

Salganik, L. H., & Tal, J. (1989). A Review and Reanalysis of the ETS/NAEP Young *Adult Literacy Survey*. Washington, DC: Pelavin Associates.

Salvia, J. & Ysseldyke, J.E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin Company.

SAS Institute, Inc. (1985). The FREQ procedure. In *SAS Users Guide: Statistics, Version 5 Edition*. Cary, NC: Author.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, *20*(2), 135-154.

Smith, M. (2010). *The need for differentiating mathematics instruction* (MetaMetrics Policy Paper). Durham: MetaMetrics, Inc.

Smith, M., & Turner, J. (no date). *Supporting differentiated math instruction in a Common Core world*. Durham: MetaMetrics, Inc.

Smith, M., & Turner, J. (2012, February). *A mathematics problem: How to help students achieve success in mathematics through college and beyond* (MetaMetrics Policy Brief). Durham, NC: MetaMetrics, Inc.

Starr, L. (2002). Math wars! *Education World*. Retrieved January 27, 2003, from http://www.eduationworld.com/a_curr/curr071.shmtl.

Stenner, A.J. (1990). Objectivity: Specific and general. *Rasch Measurement Transactions*, *4*, 111.

Thurstone, L. L. (1946). Note on a reanalysis of Davis' Reading Tests. *Psychometrika, 11*(2), 185.

Tomlinson, C.A. (1999). *The differentiated classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wright, B.D. & Linacre, J.M. (1994, August). *The Rasch model as a foundation for the Quantile Framework*. Unpublished manuscript.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Zwick, R. (1987). Assessing the Dimensionality of the NAEP Reading Data. *Journal of Educational Measurement, 24,* 293–308.

Appendix A

THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

Imagine empowering and accelerating students' learning in mathematics by better differentiating instruction and monitoring growth in student ability. With the Quantile Framework, educators can help achieve this goal by identifying level-appropriate mathematical tasks for students and track their progress!

HOW IT WORKS

The Quantile Framework for Mathematics is a unique measurement system that uses a common scale and metric to assess a student's mathematical achievement level and the difficulty of specific skills and concepts. The Quantile Framework describes a student's ability to solve mathematical problems and the demand of the skills and concepts typically taught in kindergarten mathematics through Algebra II, Geometry, Trigonometry and Precalculus. The Quantile Map provides educators with a sampling of primary mathematical skills and concepts from over 500 Quantile Skills and Concepts (QSCs) throughout the Quantile scale. This sampling of QSCs ranges from EM (Emerging Mathematician) for early, foundational mathematical skills and concepts to 1500Q for more advanced skills and concepts. As the difficulty, or demand of the skill increases, so does the Quantile measure.

HOW TO USE IT

With the Quantile Framework, educators can explore the interconnectedness of mathematical skills and concepts and identify those elements that are critical for progressing student learning. Educators are better able to inform their instruction on how to best teach a skill or concept by pinpointing which skills build upon each other. The skill mapping of mathematical concepts enables educators to build an instructional path that best fits their students' unique abilites. Both students and QSCs receive a Quantile measure. Numerous tests report Quantile student measures including many state end-of-year assessments, national norm-referenced assessments and math programs. On the QSC side, more than 580 textbooks, 64,000 lessons and 3,100 downloadable resources have received Quantile measures.

Quantile measures provide educators with the information they need to identify gaps in mathematical knowledge, as well as serve as a guide for progressing to more advanced topics. Every QSC is part of a knowledge cluster that shows relationships and connections between mathematical skills and offers their relative difficulty among different skills. Both the *prerequisite* and *impending* skills are elements of knowledge clusters and serve as building blocks that support students' success. Educators can advance student learning by using prerequisite and impending skills to build mathematical knowledge and understanding. Prerequisite skills help educators see the pieces of the puzzle that make up a skill or concept, showing what needs to be understood first. Impending skills are skills and concepts that build upon a focus skill and allow educators to see a trajectory of knowledge across grades and content strands.

15000+
14000Q
13000Q
12000Q
11000Q
10000Q
9000Q
8000Q
7000Q
6000Q
5000Q
4000Q
3000Q
2000Q
1000Q
0Q
EM1000Q

MetaMetrics®
LINKING ASSESSMENT WITH INSTRUCTION

For more information, free resources, and to search the Math Skills Database, visit Quantiles.com.

THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

High School Example
James

Heritage High School | Geometry Course

Quantile Measure: 1190Q

James is exploring theorems about lines and angles in his Geometry class. In his current learning path, the focus skill being taught is *use properties, definitions, and theorems of angles and lines to solve problems related to adjacent, vertical, complementary, supplementary, and linear pairs of angles.* This focus skill is part of a knowledge cluster that contains prerequisite and impending skills. Working with prerequisite skills can help students struggling to learn and impending skills can help students progress to the next level of learning.

Since James' Quantile measure is within the range of the focus skill being taught (his Quantile measure +/- 50Q), James will be ready for this type of instruction. With his mathematical ability being at the same level as the focus skill, learning will be optimal. Once James is performing well with the focus skill, he will be better prepared to learn the impending skills connected with this focus skill.

1250Q
IMPENDING SKILL
Use definitions and theorems of angles formed when a transversal intersects parallel lines.

1220Q
IMPENDING SKILL
Use properties, definitions, and theorems of polygons to solve problems related to the interior and exterior angles of a convex polygon.

1160Q
FOCUS SKILL
Use properties, definitions, and theorems of angles and lines to solve problems related to adjacent, vertical, complementary, supplementary, and linear pairs of angles.
CCSS G.CO.9

1010Q
PREREQUISITE SKILL
Define and identify alternate interior, alternate exterior, corresponding, adjacent and vertical angles.

1020Q
PREREQUISITE SKILL
Define and identify complementary and supplementary angles.

800Q
PREREQUISITE SKILL
Write a linear equation or inequality to represent a given number or word problem; solve.

◆ ALGEBRA & ALGEBRAIC THINKING ★ NUMBER SENSE ■ NUMERICAL OPERATIONS ● MEASUREMENT ▲ GEOMETRY ▲ DATA ANALYSIS , STATISTICS & PROBABILITY

MetaMetrics.

For more information, visit Quantiles.com.

THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

Middle School Example
Sophia

Heritage Middle School | Grade 6

Quantile Measure: 770Q

Sophia is using variables to represent mathematical expressions in her math class. In her current learning path, the focus skill being taught is *translate between models or verbal phrases and algebraic expressions.* This focus skill is part of a knowledge cluster that contains prerequisite and impending skills. Working with prerequisite skills can help students struggling to learn and impending skills can help students progress to the next level of learning.

Since Sophia's Quantile measure is within the range of the focus skill being taught (her Quantile measure +/- 50Q), Sophia will be ready for this type of instruction. With her mathematical ability being at the same level as the focus skill, learning will be optimal. Once Sophia is performing well with the focus skill, she will be better prepared to learn the impending skills connected with this focus skill.

810Q
IMPENDING SKILL
Write an equation to describe the algebraic relationship between two defined variables in number and word problems, including recognizing which variable is dependent.

800Q
IMPENDING SKILL
Identify parts of a numerical or algebraic expression.

800Q
IMPENDING SKILL
Write a linear equation or inequality to represent a given number or word problem; solve.

750Q
FOCUS SKILL
Translate between models or verbal phrases and algebraic expressions.
CCSS 6.EE.6

620Q
PREREQUISITE SKILL
Translate between models or verbal phrases and numerical expressions.

430Q
PREREQUISITE SKILL
Describe the meaning of an unknown in the context of a word problem.

ALGEBRA & ALGEBRAIC THINKING | NUMBER SENSE | NUMERICAL OPERATIONS | MEASUREMENT | GEOMETRY | DATA ANALYSIS , STATISTICS & PROBABILITY

MetaMetrics.

For more information, visit Quantiles.com.

THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

Late Elementary Example
Donald

Heritage Elementary School | Grade 4

Student Quantile Measure: 450Q

Donald is learning about line graphs with very large data values. In his current learning path, the focus skill being taught is *organize, display, and interpret information in graphs containing scales that represent multiple units.* This focus skill is part of a knowledge cluster that contains prerequisite and impending skills. Working with prerequisite skills can help students struggling to learn and impending skills can help students progress to the next level of learning.

Since Donald's Quantile measure is within the range of the focus skill being taught (his Quantile measure +/- 50Q), Donald will be ready for this type of instruction. With his mathematical ability being at the same level as the focus skill, learning will be optimal. Once Donald is performing well with the focus skill, he will be better prepared to learn the impending skills connected with this focus skill.

800Q

IMPENDING SKILL

Identify and use appropriate scales and intervals in graphs and data displays.

480Q

FOCUS SKILL

Organize, display, and interpret information in graphs containing scales that represent multiple units.

CCSS 3.MD.3

480Q

IMPENDING SKILL

Organize, display, and interpret information in bar graphs.

470Q

IMPENDING SKILL

Organize, display, and interpret information in line graphs.

200Q

PREREQUISITE SKILL

Organize, display, and interpret information in line plots and tally charts.

110Q

PREREQUISITE SKILL

Skip count by 2s, 5s and 10s beginning at any number.

90Q

PREREQUISITE SKILL

Skip count by 3s, 4s, 6s, 7s, 8s, and 9s.

EM10Q

PREREQUISITE SKILL

Organize, display, and interpret information in picture graphs and bar graphs using grids.

ALGEBRA & ALGEBRAIC THINKING | NUMBER SENSE | NUMERICAL OPERATIONS | MEASUREMENT | GEOMETRY | DATA ANALYSIS , STATISTICS & PROBABILITY

MetaMetrics.

For more information, visit Quantiles.com.

THE QUANTILE® FRAMEWORK FOR MATHEMATICS MAP

Linking assessment with mathematics instruction

Early Elementary Example
Aliyah

Heritage Elementary School | Kindergarten

Quantile Measure: EM100Q

Aliyah is exploring unknown-addend problems in her class. In her current learning path, the focus skill being taught is *know and use related addition and subtraction facts*. This focus skill is part of a knowledge cluster that contains prerequisite and impending skills. Working with prerequisite skills can help students struggling to learn and impending skills can help students progress to the next level of learning.

Since Aliyah's Quantile measure is within the range of the focus skill being taught (her Quantile measure +/- 50Q), Aliyah will be ready for this type of instruction. With her mathematical ability being at the same level as the focus skill, learning will be optimal. Once Aliyah is performing well with the focus skill, she will be better prepared to learn the impending skills connected with this focus skill.

EM25Q

IMPENDING SKILL

Model the concept of subtraction using numbers less than or equal to 10.

EM80Q

FOCUS SKILL

Know and use related addition and subtraction facts.

CCSS 1.OA.4

EM110Q

PREREQUISITE SKILL

Identify missing addends for addition facts.

EM260Q

PREREQUISITE SKILL

Model the concept of addition for sums to 10.

MetaMetrics.

For more information, visit Quantiles.com.

ALGEBRA & ALGEBRAIC THINKING | NUMBER SENSE | NUMERICAL OPERATIONS | MEASUREMENT | GEOMETRY | DATA ANALYSIS , STATISTICS & PROBABILITY