The North Carolina Testing Program

Technical Report

2012–2015

English Language Arts/Reading

Assessments (ELA)

End-of-Grade 3–8 and End-of-Course English II

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

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 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 (revise by SBE action in March 2014)

1.2 North Carolina ELA EOG and EOC Assessments

This technical manual addresses the End-of-Grade (EOG) assessments of ELA in grades 3 through 8 and the English II End-of-Course (EOC) assessment. End-of-grade and end-of-course assessments are only administered to students in English, and as explained above, are aligned to the Common Core State Standards.

Each operational base form of the EOG ELA/reading assessment has between 44 to 48 operational multiple-choice items constructed from six reading passages of which three are informational, two literature, and one poetry. The EOC English II assessment has 53 operational items: 46 multiple-choice, four technology enhanced (TE) items, and three constructed response items constructed from six reading passages of which three are informational, two literature, and one poetry. The EOG assessments were available in Paper format only in 2012–13.*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 8 field test items embedded within each form. EOC English II has 15 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.

Beginning in the 2014–15 school year, the EOG grade 7 was also available as a computer-based, fixed-form administration. EOC English II assessment was designed as a computer-based fixed form assessment with paper-based fixed forms available as accommodation for schools and individual students.

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 assessment must be administered the final ten instructional days of the school year. Students have up to four hours to complete each assessment.

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

Grade	MC Item		TE Item		CR Item	
	Number of Items	MSP per Item	Number of Items	MSP per Item	Number of Items	MSP per Item
Grade 3	44	1				
Grade 4	44	1				
Grade 5	44	1				
Grade 6	48	1				
Grade 7	48	1				
Grade 8	48	1				
English II	46	1	4	1	3	2

Note: MC=Multiple-Choice; TE=Technology-Enhanced; CR=Constructed 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 ELA EOG and EOC English II 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. The 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 tryout, 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 and ELL are able to take EOG and EOC assessments.

Chapter 6 describes the processes used for scoring items and procedure adopted to create final reportable scale scores. The first two sections of this chapter summarize the automated scoring procedures to transform students' responses into a number correct score for fixed response items and the human scoring process for assigning score category for constructed response items. Section three and four 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 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 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 ELA EOG and EOC 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 in 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 2 presents samples and a summary description of the various standardized reports created by NCDPI and available to LEA to provide and interpret assessments results to 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 these 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 EOG and EOC data. The final sections of the chapter document validity evidences: evidence based on content summarized from the alignment study, evidence based on relation to other variables summarized from the EOG/EOC Lexile 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 ELA

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."*

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 constructing, 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, p46) 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 ELA 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 ELA, EOG, and EOC Assessments

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

2. 2 Uses of NC ELA 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 English Language Arts 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- an –Career-Readiness standards

Scores from EOG and EOC are transformed, grouped, and reported into 1 of 5 achievement levels (in 2012–13 scores were reported using 4 achievement levels) corresponding to 1 of the 5

performance level descriptors adopted by the state to classify students based on their progress and readiness as defined by NCSCS College- and Career- Readiness standards.

- Assessment results are used for school and district accountability under the READY Accountability Model and for Federal reporting purposes.

EOG and EOC students' score data are part of the quantitative indicators used in two main components of the new state READY accountability model: educator effectiveness, and school performance grades. 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 English II has to come from their EOC assessment score. It is worth mentioning that the EOG in grades 4–8 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. …"* (p23). 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 NCSCS. To accommodate the shortened timeline, NCDPI made three modifications to the SBE 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 of item tryout for the newly developed innovative technology-enhanced item types.
II. The NCDPI also waived pilot testing (Step 18), because pilot tests are administered only for newly developed items not for assessments revised from a preceding test (GCS-A-013, Phase 4: Pilot/Operational Test Development, Step 18: Administer Test as Pilot, footnote 5).
III. The NCDPI used operational data (Step 21) instead of field test data for the Standard Setting process (Step 20).

Table 3.1 Flow Chart of Test Development of North Carolina Assessments

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

[a]Activities done only at implementation of new curriculum

[b] Activities involving NC teachers

3.1 Content Standards and Curriculum Connectors

As stated in Chapter 1 (see *Table 1.1*), the NCSBE adopted the revised NCSCS in June 2010. The revised NCSCS are aligned to the Common Core state standards (CCSS). Operational test forms aligned to the NCSCS for ELA and math were administered in 2012–13 testing administration (READY initiative). Testing of North Carolina students' skills relative to the standards and objectives in the NCSCS 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 NCSCS using Hess's matrix.

Table 3.2 Hess' Cognitive Rigor Matrix with Curricular Examples

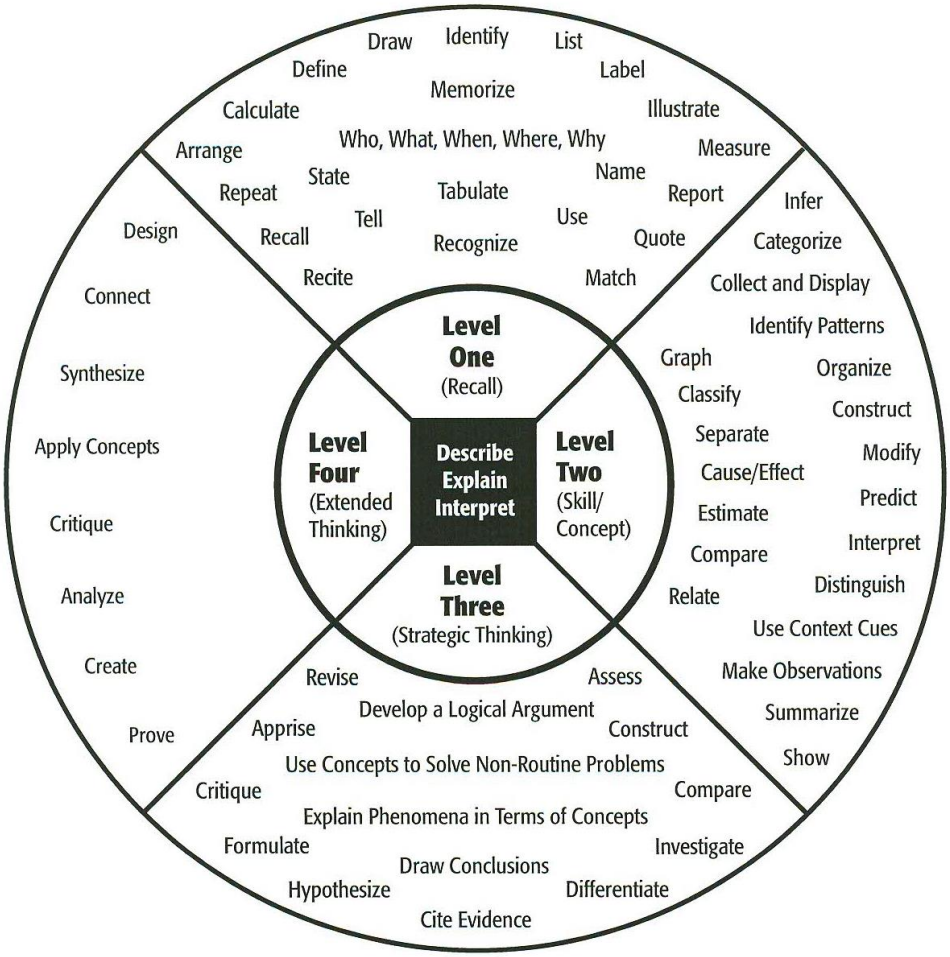
Bloom's Revised Taxonomy of Cognitive Process Dimensions	**Webb's Depth-of-Knowledge (DOK) Levels**			
	Level 1 **Recall & Reproduction**	**Level 2** **Skills & Concepts**	**Level 3** **Strategic Thinking/Reasoning**	**Level 4** **Extended Thinking**
Remember Retrieve knowledge from long-term memory, 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 signal words.	○ Use concepts to solve non-routine problems ○ Design investigation for a specific purpose or research question ○ Conduct a designed investigation ○ Apply concepts to solve non-routine problems ○ Use reasoning, planning, and evidence ○ Revise final draft for meaning or progression of ideas	○ Select or devise an approach among many alternatives to solve a novel problem ○ Conduct a project that specifies a problem, identifies solution paths, solves the problem, and reports results ○ Illustrate how multiple themes (historical, geographic, social) may be interrelated
Analyze Break into constituent parts, determine how parts relate, differentiate between relevant-irrelevant, distinguish, focus, select, organize, outline, find coherence, deconstruct (e.g., for bias or point of view)	○ Retrieve information from a table or graph to answer a question ○ Identify or locate specific information contained in maps, charts, tables, graphs, or diagrams	○ Categorize, classify materials ○ Compare/contrast figures or data ○ Select appropriate display data ○ Organize or interpret (simple) data ○ Extend a pattern ○ Identify use of literary devices ○ Identify text structure of paragraph ○ Distinguish: relevant-irrelevant information, fact/opinion	○ Compare information within or across data sets or texts ○ Analyze and draw conclusions from more complex data ○ Generalize a pattern ○ Organize/interpret data: complex graph ○ Analyze author's craft, viewpoint, or potential bias	○ Analyze multiple sources of evidence or multiple works by the same author, or across genres or time periods ○ Analyze complex/abstract themes ○ Gather, analyze, and organize information ○ Analyze discourse styles
Evaluate Make judgments based on criteria, check, detect inconsistencies or fallacies, judge, critique			○ Cite evidence and develop a logical argument for concepts ○ Describe, compare, and contrast solution methods ○ Verify reasonableness of results ○ Justify conclusions made	○ Gather, analyze, & evaluate relevancy & accuracy ○ Draw & justify conclusions ○ Apply understanding in a novel way, provide argument or justification for the application
Create Reorganize elements into new patterns/structures, generate, hypothesize, design, plan, construct, produce	○ Brainstorm ideas, concepts, or perspectives related to a topic or concept	○ Generate conjectures or hypotheses based on observations or prior knowledge	○ Synthesize information within one source or text ○ Formulate an original problem given a situation ○ Develop a complex model for a given situation	○ Synthesize information across multiple sources or texts ○ Design a model to inform and solve a real-world, complex, or abstract situation

In July, 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



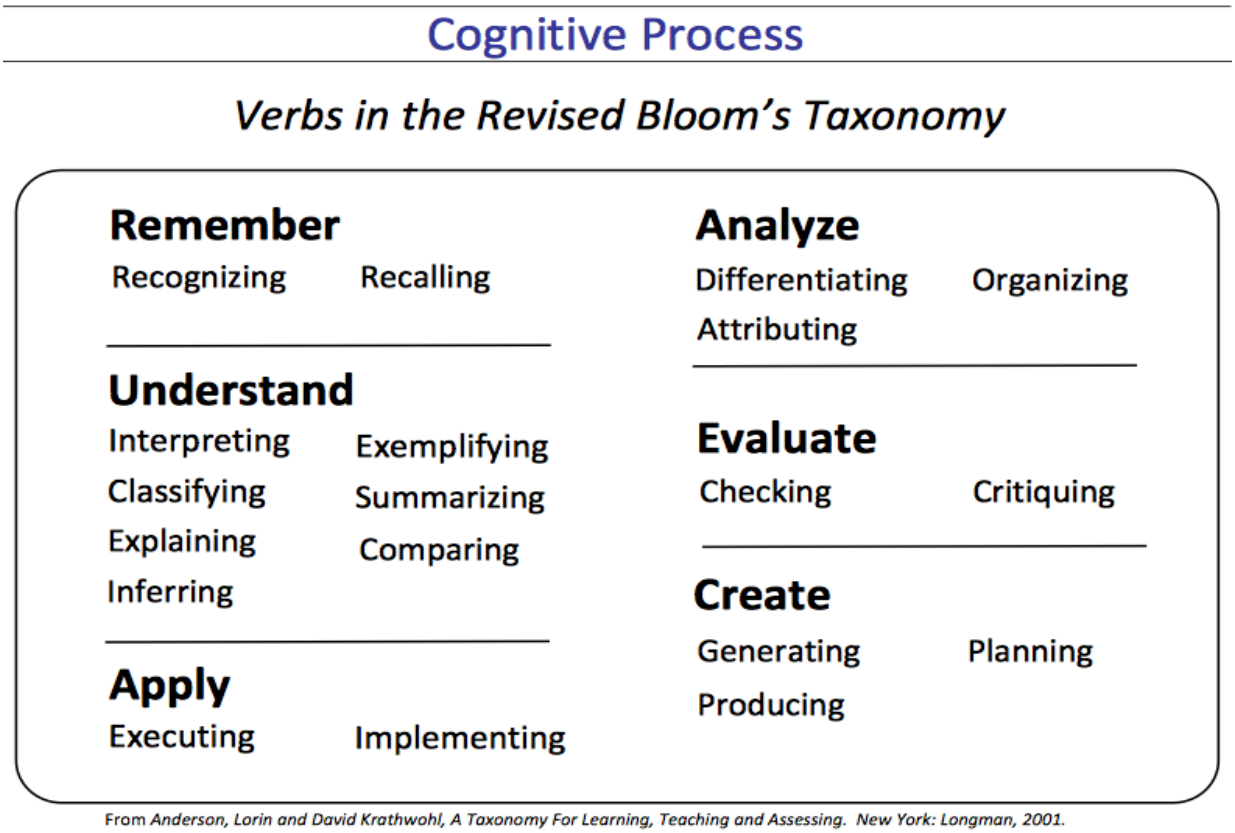
Level One Activities	Level Two Activities	Level Three Activities	Level Four Activities
Recall elements and details of story structure, such as sequence of events, character, plot and setting.	Identify and summarize the major events in a narrative.	Support ideas with details and examples.	Conduct a project 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[c] 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* shows the adopted content domain specification for ELA EOG grades 3-8 and EOC English II assessments.

[c] 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 Weights, Grades 3–8 ELA and English II

Domain/Standards	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	English II
Reading for Literature	32–37%	30–34%	36-40%	32–36%	34–38%	31–35%	30–34%
Reading for Information	41–45%	45–49%	37-41%	41–45%	41–45%	42–46%	32–38%
Reading Foundation Skills	NA	NA	NA	NA	NA	NA	NA
Writing	NA	NA	NA	NA	NA	NA	14–18%
Speaking and Listening	NA	NA	NA	NA	NA	NA	NA
Language	20–24%	19–21%	21-25%	21–25%	19–23%	20–24%	14–18%
Total	100%	100%	100%	100%	100%	100%	100%

The English Language Arts/Reading NCSCS consist 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 Content Standards and Weights, Grades 3–8 ELA and English II*

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 ELA Test Specifications & Blueprints. This iteration of EOG 3–8 assessments does not assess Reading foundation skills, writing, and speaking and listening. Writing is only assessed in English II in high school.

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 at 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 examine 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 ELL students. 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 NCSCS 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 NCSCS 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 develop test questions that closely agree with the NCSCS standards.

The "B" level course is designed as the next level course for potential item writers/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 Technology Enhanced 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 ELA was on computer-based technology-enhanced (TE) items. The evaluation criteria centered on aspects of accessibility, user-friendliness, and authenticity of construct measured. During the exploratory phase, the NCSTP looked at several varieties of TE items with their functionalities such as click-and-drag features, hot-spot features, special graphics, audio, or video. While these hold promise to improve student engagement and appeal of the assessment, they do require extra development safeguards to ensure that the items appear and function as intended while minimizing the introduction of construct irrelevant variance. Also, there needs to be evidence that the scoring protocol is accurate and all responses are scored properly, and that students with minimum computer skills are not disadvantaged. A usability study allowed test developers to observe students interacting with these new items and provided valuable feedback on the improvement, design and selection TE items.

Figure 3.3–Figure 3.6 shows a snap-shot of four types of TE items that were considered for English II. *Figure 3.3* shows an example of a Text Identified (TI) item with a stem and multiple options. Students are instructed to read the stem, then identify the correct text provided by clicking on all correct options. String Replace (SR) is shown in *Figure 3.4*. In this example, students are presented with a short text that has one word highlighted "**hot text**" and a list of four possible replacement words. The task is to select a response option by clicking or hovering using the mouse pointer over any choice from the list provided, with an appropriate replacement of the "hot-text." This action replaces the "hot-text" in the reading selection. These were the two TE item types test development staff evaluated that would be suited for operational administration. Other types of TE items considered are shown in *Figure 3.5* and *Figure 3.6*.

Figure 3.3 Text Identify TE Item Example

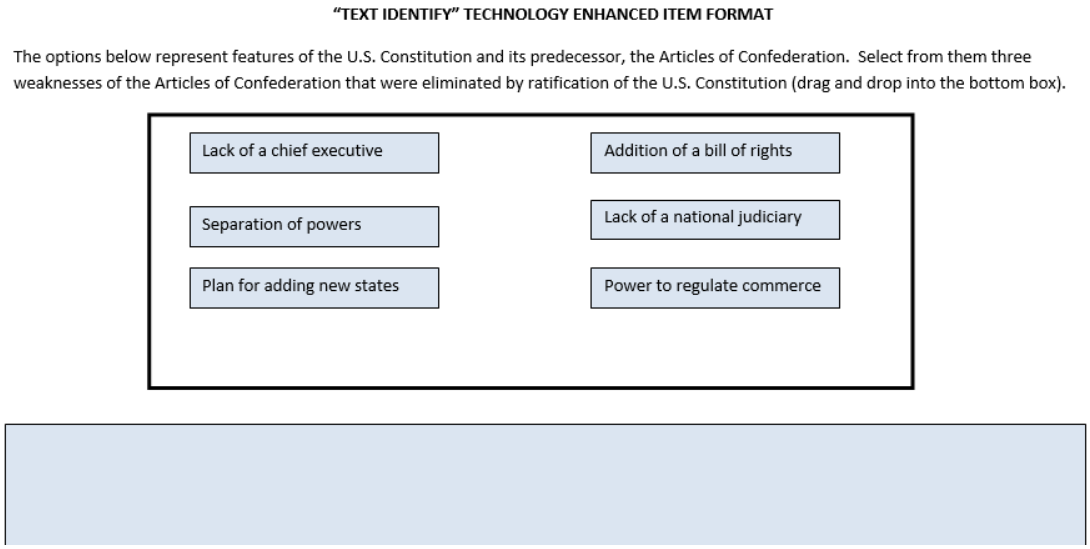


Figure 3.4 String Replace TE Item Example

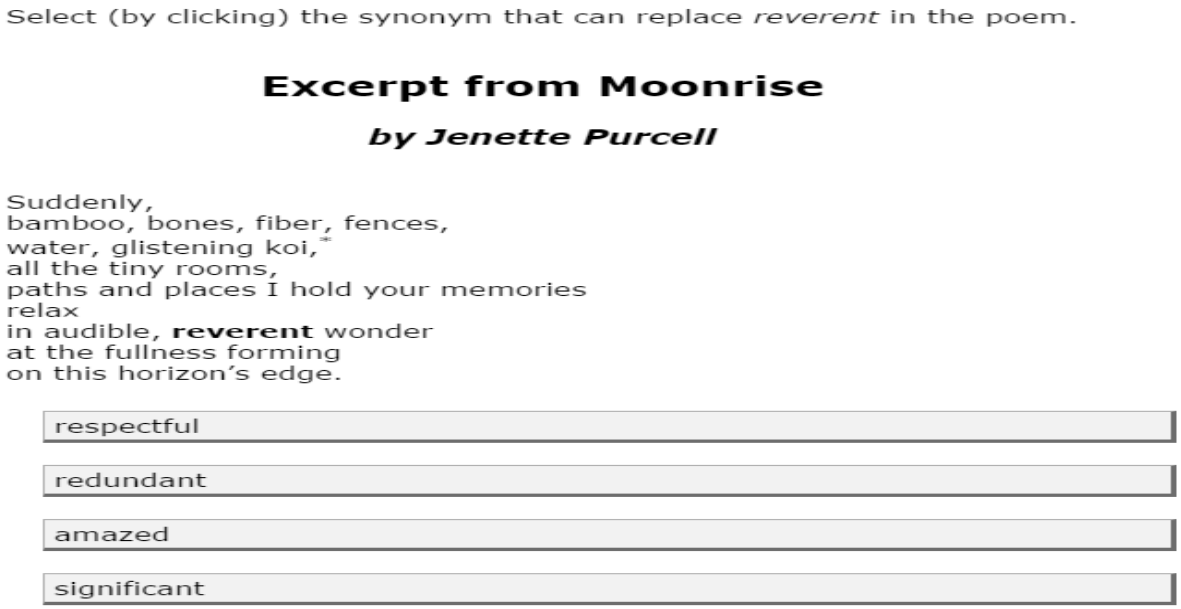


Figure 3.5 String Choice TE Item Example

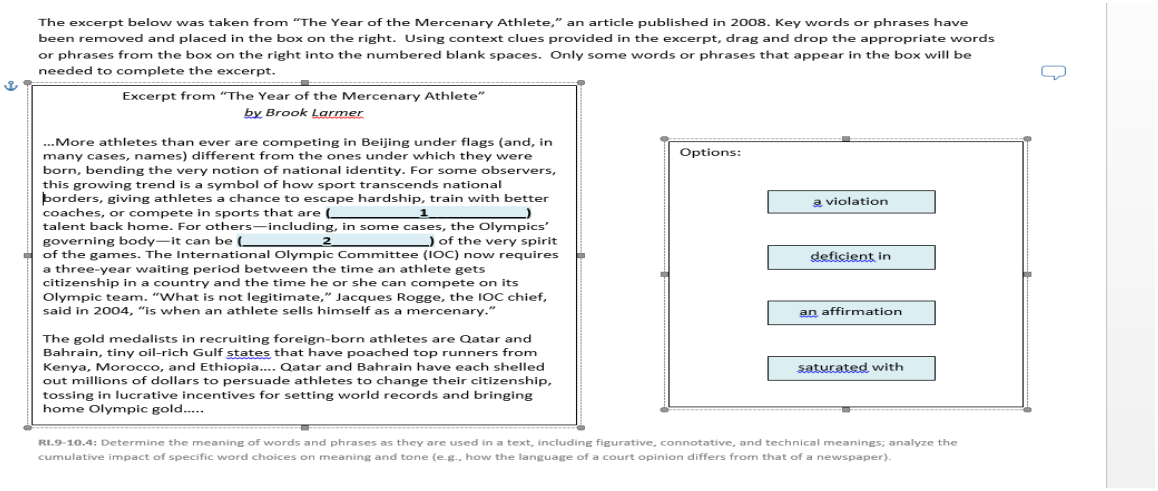
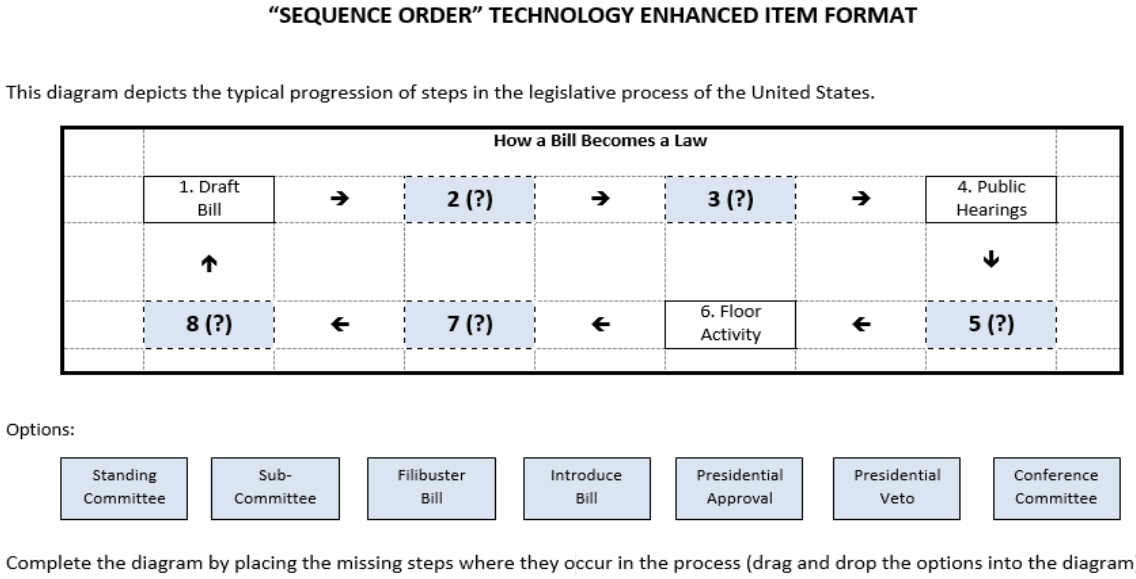


Figure 3.6 Sequence Order TE Item Example



The usability study for TE items in ELA was restricted to EOC English II because it was the only ELA assessment that was designed at that time to be administered on a computer. The goal was to design TE 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 English II, a total of 8 students from Fuquay-Varina High School in Wake County took part in the TE usability study. During the two day window, evaluators from the

NCDPI met with the selected students at their schools with laptops pre-loaded with assessment software.

Each student worked on four TE items (2 TI and 2 SR) 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 TE questions was not challenging for the student, and no scores were reported. *Table 3.4* shows the usability study process in detail.

Table 3.4 Technology Enhanced Items Usability Process

Step	Purpose	Time (minutes)
1. Introductions	Introduce student to evaluator.	3–5
2. Ice breaker activity	Set the student at ease and establish a friendly atmosphere.	4–5
3. Overview of session	Preview the session. Provide directions.	3–5
4. Present item 1	Protocol 1. Evaluator begins recording 2. Present item and ask student to read directions and answer question 3. Student interacts with test question 4. Evaluator observes and takes notes 5. Evaluator stops recording when student is finished	7–10
5. Present 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 his/her 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**.**

Seven students completed the English II usability study. It took an average of about 13 minutes for students to complete the 5-item task (1 multiple-choice and 4 TE). Overall results were positive; all TE items worked as expected, and the scoring was applied properly. From the perspective of the students, below are summaries from the interviews.

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

  Survey results for English II participants confirmed six out of the seven students agreed the directions to the items were clear to follow. Five out of seven students spent one minute or less on the directions before starting working on the items, and only one spent more than two minutes. All of the students were able to locate the information they needed to answer the questions and knew how to indicate their response choice.

- Was anything confusing or unclear to you about these questions?

  All English II students reacted positively to taking the test on the computer. During the test, the students indicated that the items worked properly with no technical problem in storing answers, scoring items, or answering TE items. They did not report any issues about how the items displayed on the screen. None of the students stumbled or had problems during the test, so no intervention was provided.

  Most students (six out of seven) showed their preference to TE items over MC items, and the one who did not show preference treated the two item types similarly.

3.3.4 Item Tryout

In Spring 2011, the NCDPI also conducted an online item tryout for EOC English II with the purpose to evaluate new item types and assessments delivered via the new computer platform. As a part of the item tryout study, students were asked to respond to a short survey about their experience interacting with the test questions, their opinions about computer based testing, and their daily online experiences. The results summary from the student survey are present below.

More than 1,900 students who participated in the survey for EOC English II during the item tryout study are shown in *Table 3.5* for the complete demographic summary of participants. In general, 84% of students reported enjoying the experience of taking the assessment on the computer and easily navigated around the test platform. 82% of students also agreed that instructions for questions that required more than clicking on an answer choice were clear and

easy to understand. Regarding technology-enhanced item types, 75% of students agreed with the statement “items that required clicking on sentences within a paragraph to select choices worked as expected” while 9% of students did not agree with the statement. A complete summary of student responses regarding usability and accessibility of the testing platform and new item types are presented in *Table 3.6*

.

Table 3.5 Demographic characteristics of the students who took the survey.

Demographic Characteristics		**Frequency**	**Percent**
Ethnicity	White	1,095	57.1%
	Black	524	27.3%
	Hispanic	198	10.3%
	Asian	38	2.0%
	American Indian	4	0.2%
	Pacific Islander	3	0.2%
	Multiple	56	2.9%
Gender	Female	939	49.0%
	Male	979	51.0%
Grade	9	151	7.9%
	10	1,747	91.1%
	11	20	1.0%

Table 3.6 Usability / accessibility of the new item types on computer

	Agree		Neutral		Disagree		Did Not Respond	
	N	%	N	%	N	%	N	%
Instructions for questions that required more than clicking on an answer choice were clear and easy to understand.	1,576	82.2	200	10.43	108	5.6	34	1.8
I was able to easily navigate through questions with reading selections and easily read the text of each selection.	1,612	84.1	129	6.7	114	5.9	63	3.3
Items that required clicking on sentences within a paragraph to select choices worked as expected.	1,444	75.3	224	11.7	182	9.5	68	3.6
I was able to use my scrap paper effectively.	437	22.8	743	38.7	622	32.4	116	6.1

On questions regarding students' preference of mode, the responses were mixed with 42% of students affirming that they found questions that made them interact with a computer more interesting, and 58% of students agreed they like taking this kind of assessment on a computer (see *Table 3.7*). Also when asked if they felt online tests were better than paper-and-pencil tests for English II, 976 students (50.89%) responded "Yes," compared with 323 (16.84%) answered "No."

Table 3.7 Preference of item types / test modes

	Agree		Neutral		Disagree		Did Not Respond	
	N	%	N	%	N	%	N	%
Questions in which I had to interact with the computer were more interesting than multiple-choice questions.	812	42.3	586	30.6	460	24.0	60	3.1
I liked taking this kind of test on the computer.	1,129	58.9	400	20.9	313	16.3	76	4.0

Finally, students were asked to provide data on the overall computer use both in school and at home. The full results of students use as recorded from their survey responses is shown in *Table 3.8*. The majority (85%) of students reported having a social network page. In terms of using a computer in academics, 60% affirmed using emails or a web application to write and turn

in homework. About 79% of students reported to have used or handled computer devices at school, and 49% of participants plan to take an online course in the near future.

Table 3.8 Past experience with computer

	Yes		No		Did Not Respond	
	N	**%**	**N**	**%**	**N**	**%**
Have you used social network services (e.g., Facebook, MySpace, etc.)?	1,644	85.7	224	11.7	50	2.6
Have you turned in writing assignments using e-mail or web applications?	1,166	60.8	703	36.7	49	2.6
Have you used a handheld computing device at school (e.g., clickers, mp3 players, etc.)?	1,512	78.8	353	18.4	53	2.8
Have you taken an online course or do you plan to take one in the near future?	951	49.6	915	47.7	52	2.7

Among the student participants, 200 (10.43%) reported that they did not spend any time on a computer or a video game console each day, while 1,397 (72.8%) spent around 1 to 4 hours each day, 216 (11.3%) spent from 5 to 10 hours, and 52 (2.7%) students reported they spent more than 10 hours a day on a computer or on video games.

Regarding any issues students experienced during the test, 277 (14.4%) of students had problems navigating between pages/questions, 247 students (12.9%) reported they encountered issues when selecting answer choices, 239 (12.5%) had problems with highlighting text, 206 (10.7%) struggled with clicking on buttons or using tools, and 165 (8.6%) students claimed they had trouble selecting answer choices.

3.3.5 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.

3.3.6 Item Alignment

A critical aspect of item quality is alignment. Alignment refers to the extent to which an item agrees with and represents the content standard it is designed to measure. Assessments composed of items that are misaligned will generate scores that do not measure the breadth and depth of the intended construct. Scores from a misaligned assessment are characterized 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[d]. These groups independently reviewed proposed items through NC's online item writing system, the Test Development System (TDS), and classified them by the NCSCS and Depth of Knowledge (DOK) levels. Any items with discrepant classifications were prevented from continuing through item development until the discrepancy was resolved.

3.3.7 Item Format

The ELA grades 3–8 assessments consist of four-foil (distractor) multiple-choice items built around selected reading texts. For EOC English II, three main-items types were selected for the computer-based fixed forms. Traditional four-foil multiple-choice, two types of technology

[d]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 content standards.

enhanced items, and short constructed response. The two types of TE items referenced in the usability studies that were developed for EOC forms are: Text Identify (TI), and String Replace (SR). For examples and description of TI and SR items see *Figure 3.3* and *Figure 3.4*.

3.4 Step 9. Item Review for Field Testing

To ensure that items were developed to align to the NCSCS 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 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[e] based on the assessment content specification and blueprint. Field test forms were organized according to the blueprints to

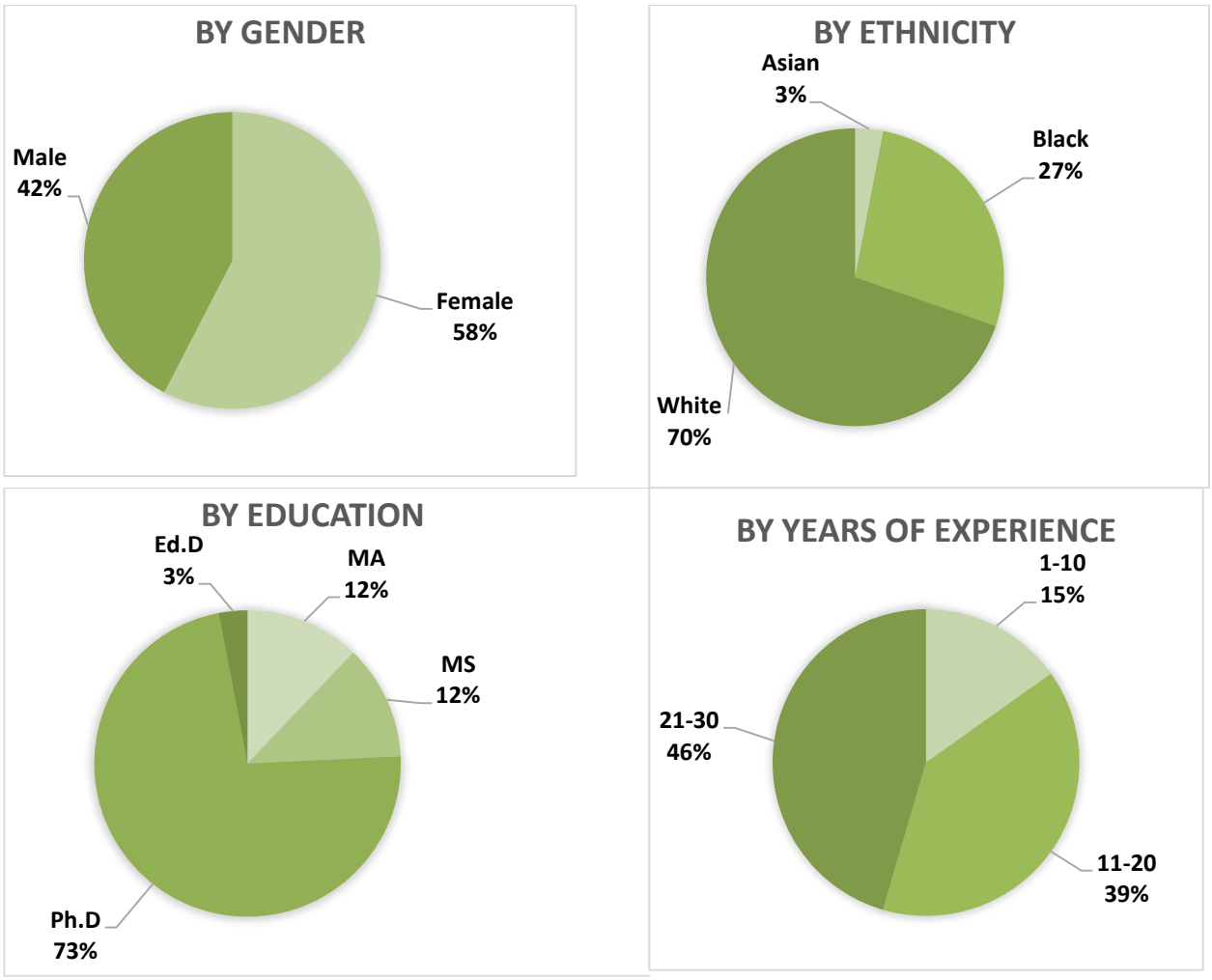
[e] See complete form assembly process described in chapter 5

be implemented for the operational assessment. *Table 3.9* 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. 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 EOG and EOC cycle. Descriptive summaries of their demographic and educational backgrounds are shown in the pie charts in Figure 3.7. 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.

Table 3.9 Number of items field tested for ELA, EOG, and EOC

Grade	Number of Forms	Items Per Form	Total Items Field Tested
ELA Grade 3	12	58	696
ELA Grade 4	12	58	696
ELA Grade 5	12	58	696
ELA Grade 6	12	62	744
ELA Grade 7	12	62	744
ELA Grade 8	12	62	744
English II	12	59	708

Figure 3.7 *Demographic Information for Outside Form Reviewers*



Chapter 4 Field-Test Administration and Operational Form Construction

The NC ELA 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[f]

Sampling for 2011–12 stand-alone field testing of the North Carolina ELA 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

[f] 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 comparison of the proportions of students selected for the stand-alone field test compared to the population. The desired sampling rate was set at 20% from each grade level. After attrition, the effective sampling rate across the grade levels ranged from 17% for English II to 22% for grade 4. Demographic proportions from the field test sample and population across the respective grades show a very similar distribution across the major demographic variables. 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 4% 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 with reasonable levels of sampling error.

Table 4.1 Demographic Summary for ELA Field Test 2012 Sample Participants

ELA		N	Gender		Ethnicity							Special Subgroup		
			% Female	% Male	% Asian	% Black	% Hispanic	% American Indian	% Multiracial	% Native Hawaiian/Pacific Islander	% White	% ELL[g]	% SWD[h]	% EDS[i]
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	26,095	49.57	50.43	2.44	23.94	14.93	1.25	3.87	0.06	53.51	10.73	9.63	57.88
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	27,709	49.91	50.09	2.50	25.66	15.38	1.35	3.67	0.08	51.35	8.87	10.07	58.73
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	23,467	49.27	50.73	2.68	25.95	14.75	2.12	3.85	0.09	50.56	5.87	8.87	56.12
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	26,335	49.55	50.45	2.84	25.99	13.15	1.01	3.61	0.07	53.33	4.97	8.49	54.12
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	25,624	49.73	50.27	2.10	24.67	11.37	2.14	3.37	0.11	56.25	4.55	8.14	52.36
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	22,983	50.16	49.84	1.98	25.77	12.17	1.22	3.97	0.11	54.78	4.24	8.09	53.24
English II	Population	118,544	48.85	51.15	2.54	28.10	10.94	1.58	3.53	0.08	53.23	3.35	11.08	48.10
	Sample	19,873	49.30	50.70	3.30	25.34	10.38	0.83	3.65	0.05	56.46	3.09	7.44	44.60

[g] English Language Learners
[h] Students with Disability
[i] Economically Disadvantaged Students based on free/reduced lunch

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 correlation

are not retained for use on the operational assessment. *Table 4.2* and *Table 4.3* show summary-descriptive classical statistics from a field test item pool.

Table 4.2 CTT Field Test 2012 Item Pool Descriptive Statistics for ELA, EOG 3–8

Grade	Number of Items	P-value Summary				Biserial Correlation Summary			
	MC	Average	SD	Min	Max	Average	SD	Min	Max
ELA 3	696	0.66	0.16	0.17	0.95	0.55	0.16	0.00	0.87
ELA 4	696	0.67	0.17	0.14	0.97	0.52	0.17	-0.09	0.91
ELA 5	696	0.67	0.17	0.09	0.98	0.51	0.17	-0.19	0.90
ELA 6	744	0.65	0.17	0.21	0.98	0.49	0.17	-0.06	0.89
ELA 7	744	0.64	0.18	0.12	0.98	0.49	0.19	-0.04	0.94
ELA 8	744	0.56	0.17	0.06	0.96	0.40	0.18	-0.29	0.84

Table 4.3 CTT Field Test 2012 Item Pool Descriptive Statistics for English II

Grade	Number of Items	P-value Summary				Biserial Correlation Summary			
		Average	SD	Min	Max	Average	SD	Min	Max
Multiple-Choice	602	0.48	0.15	0.12	0.93	0.38	0.17	-0.06	0.71
String Replace	23	0.53	0.20	0.19	0.87	0.43	0.14	0.23	0.69
Text Identify	47	0.32	0.16	0.04	0.60	0.37	0.18	0.04	0.82

Note: 36 CR items are not included

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 graded response model (GRM) for calibrating constructed response items. Equation 4-1 presents the mathematical representation for the 3PL, where:

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

(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.

Equation (4-2) shows the GRM, where

$$P_{i_g}^* = \frac{e^{a_i(\theta - b_{i_g})}}{1 + e^{a_i(\theta - b_{i_g})}}$$

(4-2)

where $P_{i_g}^*$ is the probability of responding in a particular category (0, 1, 2) to item *I*, a_i is the slope or the discrimination parameter, b_{ig} is the boundary location parameter.

The IRT parameter estimates were calibrated using IRTPRO software (Cai, Thissen, & du Toit, 2011) with the Bayesian prior distributions for the item parameter calibration set to a~lognormal(0, 1) and c~Beta(5, 15)]. For TE items, the Bayesian prior distribution of c~Beta (A, B) was set by dividing the number of possible response combinations for TI items. The use of the Bayesian prior distribution ensured appropriate parameter estimates of chance-scores were accounted for during calibration. Table 4.4 and Table 4.5 shows summary-descriptive IRT parameters statistics from a field test item pool.

Table 4.4 IRT Field Test 2012 Item Pool Descriptive Statistics for ELA EOG 3–8

Grade	Number of Items	Slope(a)				Threshold(b)				Asymptote(g)			
	MC	Average	SD	Min	Max	Average	SD	Min	Max	Average	SD	Min	Max
ELA 3	696	1.70	0.62	-1.90	4.57	-0.19	1.86	-3.30	41.48	0.21	0.06	0.09	0.45
ELA 4	696	1.55	0.62	0.01	4.56	-0.06	7.51	-3.63	193.6	0.21	0.06	0.10	0.56
ELA 5	696	1.55	0.57	0.06	4.16	-0.33	1.51	-4.12	27.60	0.22	0.06	0.07	0.49
ELA 6	744	1.53	0.66	-0.20	5.70	-0.10	2.51	-7.97	57.13	0.22	0.06	0.09	0.59
ELA 7	744	1.19	9.74	-261	5.08	-0.06	1.54	-5.69	17.81	0.21	0.06	0.09	0.46
ELA 8	744	1.40	0.83	0.06	14.50	0.42	1.54	-2.06	15.31	0.21	0.05	0.09	0.45

Table 4.5 IRT Field Test 2012 Item Pool Descriptive Statistics for ELA English II

Grade	Number of Items	Slope(a)				Threshold(b)				Asymptote(g)			
		Average	SD	Min	Max	Average	SD	Min	Max	Average	SD	Min	Max
MC	602	1.50	0.83	-2.11	5.69	0.90	1.57	-13.0	12.90	0.21	0.05	0.09	0.41
SR	23	1.84	0.74	0.52	3.49	0.33	1.01	-1.56	1.60	0.24	0.09	0.09	0.51
TI	47	1.24	0.69	0.29	3.26	1.56	1.86	-0.19	10.60	0.10	0.07	0.02	0.28

Note: 36 CR items are not included

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:

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

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

$$|D| = -2.35\beta_{MH} \qquad \textbf{(4-5)}$$

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). An item with a C- rating disfavors the focal group (favors males, whites, rurals, EDS).

Table 4.6 shows field test pool multiple-choice items by candidate DIF flag. During the initial construction of EOG and EOC assessments in 2011 the NCDPI investigated DIF for gender—male and female with male set as the reference group and female the focal group and two ethnicity categories—"White" versus "Black," and "White" versus "Hispanic." In both ethnic categories "White" was set as the reference group, and "Black" and Hispanic" were the respective focal groups. For example, for ELA EOG grade 3, females performed somewhat better on 327 items compared to males of similar ability, and males performed somewhat better on 338 items compared to females of similar ability. 15 items showed marginal DIF in favor of females, and 13 showed marginal DIF in favor of males. A total of 3 items showed significant DIF, 2 in favor of females, and 1 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 "Hispanic" category.

Based on recommendations from our 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 a category for Economically Disadvantage Students (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 4.6 Mantel-Haenszel Delta DIF Summary for ELA 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-
ELA 3	327	338	15	13	2	1	313	339	23	14	4	3	323	288	31	28	7	19
ELA 4	347	311	15	19	2	2	346	311	11	18		10	320	278	39	33	8	18
ELA 5	352	289	22	19	4	10	330	292	35	28	4	7	291	254	53	51	19	28
ELA 6	366	325	17	25	6	5	376	319	14	25	1	9	338	300	35	39	11	21
ELA 7	369	303	30	29	3	10	351	338	21	26	2	6	340	305	37	33	9	20
ELA 8	358	342	17	18	3	6	352	343	21	21		7	362	310	21	36	4	11
English II	329	314	10	19			328	318	8	17		1	295	309	31	25	3	9

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.

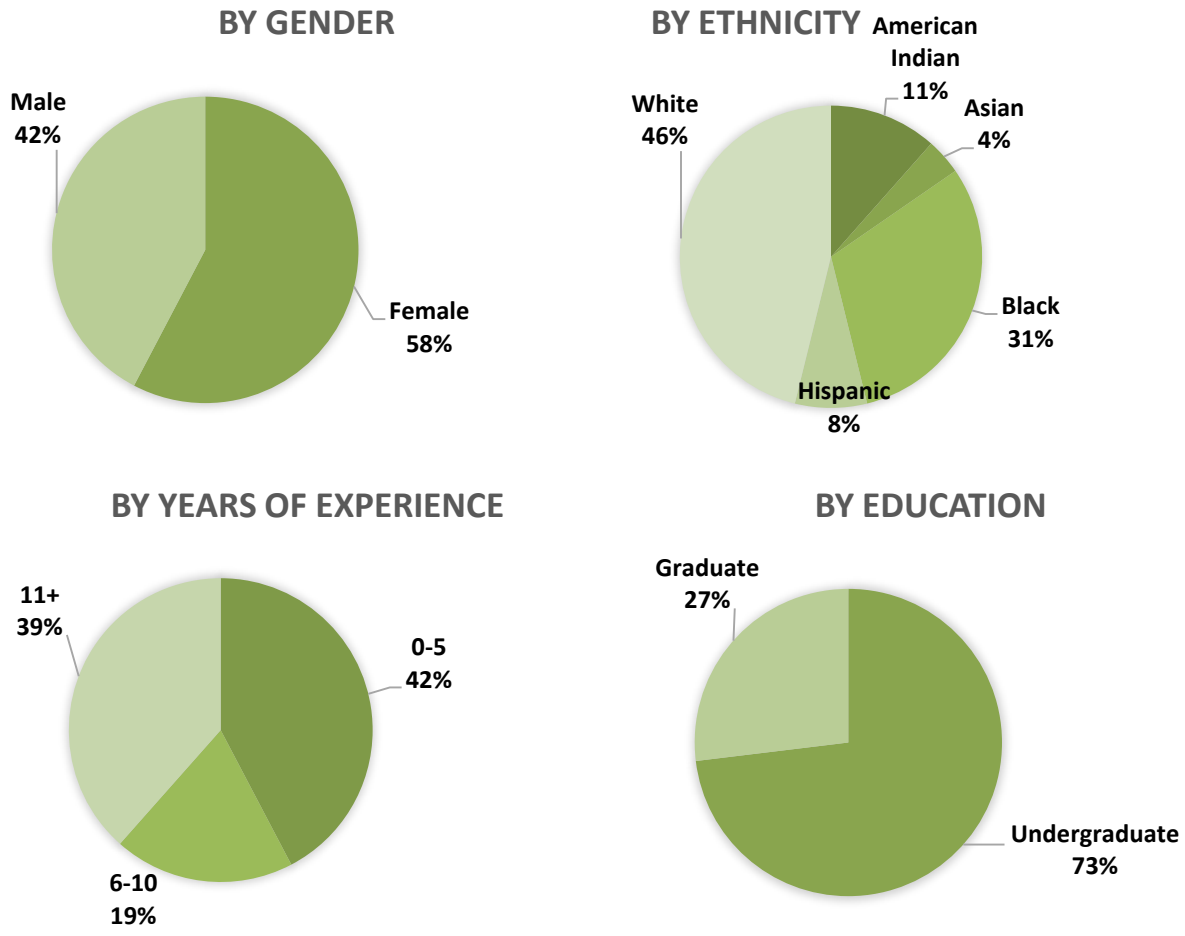
Standard 3.6 of the AERA/APA/NCME (2014) Standards states:

> "*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 with input from the NCTA convenes Bias Review panels.

The Bias 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 document 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 categories "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 4.7*. The table includes data for ELA EOG and EOC assessments administered under regular conditions—that is no accommodations of extended time and multiple test sessions. For all EOG, 75% of students completed the assessments within the 2-hours (120 Minutes) window. For EOG grades 3–7, it took about three hours and twenty minutes to three hours and thirty minutes for 99% of students in the sample to complete the assessment. In EOG grade 8, less than 1% of students spent over three hours (187 minutes) on the assessment. For EOC English II, only 1% or less of students spent more than two hours and thirty minutes on the assessment.

Table 4.7 ELA EOG and EOC Recorded Test Duration from Field Test 2012

EOG/EOC		Summary			Percentile				
	N	Number of Items	Avg.	SD	25th	Median	75th	95th	99th
Grade 3	22,415	58	89.52	34.02	65	84	107	150	200
Grade 4	23,088	58	97.42	35.22	71	90	117	162	207
Grade 5	20,392	58	102.4	35.13	77	96	122	168	210
Grade 6	22,839	62	99.17	32.02	75	95	118	155	202
Grade 7	22,331	62	96.67	31.00	75	92	115	150	200
Grade 8	19,756	62	95.59	30.02	75	91	113	148	187
English II	18,825	59	71.47	27.21	53	70	88	119	145

4.5 Step 15. Operational Test Construction

The field testing plan was designed to generate enough items to construct four equivalent forms for EOG ELA/Reading grades 4–8 and EOC English II. For ELA/Reading grade 3, the field test plan was designed to construct five equivalent forms with one of the forms to be administered as the Beginning-of-Grade 3 ELA/Reading assessment. 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 English II EOC, both computer-based and paper-based fixed forms were created. The paper-based fixed form is an exact replicate of the computer-based fixed form, with the exception of the TE items. For each grade level, one form was selected and published as a release form on the NCDPI website. The release forms were available to teachers, students, and all interested stakeholders so they could familiarize themselves with the new assessment prior to operational administration. 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)

To this end, NCDPI carefully considers all items prior to their inclusion in the operational pool and the operational test form.

4.5.1 Criteria for Item Inclusion in Operational Pool

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 4.8*. The table shows that nearly 70% of the ELA items in all grades were retained or kept as "Reserve" for use on the operational test. This provided a sufficient item pool for the construction of four parallel forms in Grades 4 through 8 and English II, and five parallel form in Grade 3.

Table 4.8 Field Test 2012 Item Pool Summary for ELA

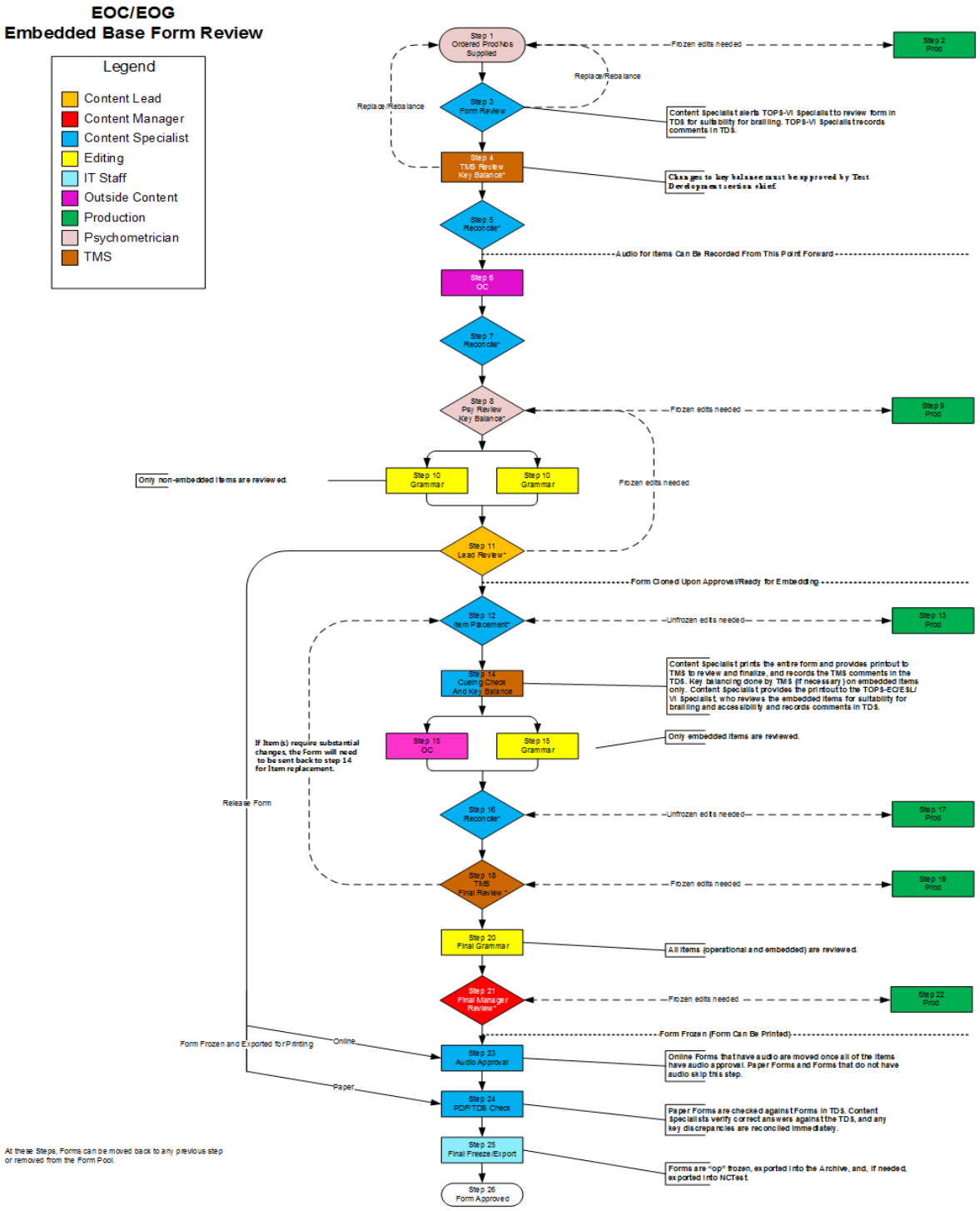
Grade Level	Psychometric Evaluation Summary					
	Keep		Reserve		Delete	
	N	Row %	N	Row %	N	Row %
ELA 3	467	67	152	22	77	11
ELA 4	363	52	186	27	147	21
ELA 5	382	55	204	29	110	16
ELA 6	398	53	188	25	158	21
ELA 7	390	52	197	26	157	21
ELA 8	358	48	184	25	202	27
English II	324	48	136	20	212	32
Total	2,682	54	1,247	25	1,063	21

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 form assembly 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 multiple items are replaced the entire form review process is reset.

At step 6 the form is sent to an outside content reviewer to offer general expert comments. Steps 8 through 11 involve grammar checks and key balance for multiple-choice items on the base form. Steps 12–18 are 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 step 18, grammar step 20 and content manager 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 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 ELA 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. Additionally, the ELA EOG was designed to have an underlying developmental scale. Therefore, a conscious decision was made to maintain a monotonically increasing difficulty (i.e., decreasing p-value) across the grade span. The rationale for this was that the material covered in each subsequent grade became more complex. The actual pool p-values generally followed the trend, and the resulting smoothing was relatively minor. *Table 7.3* and *Table 7.5* show 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 alternate 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 $-\infty$ to $+\infty$ 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 are selected from the item bank based on a predetermined blueprint to match a target or base TCC. According to Thissen et al (2001, p.158), 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 by $\sqrt{1-\rho}$. 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 that in IRT standard errors usually vary for different response patterns for the same test. Examinees with different response patterns or at different points on the theta scale will show variations in the amount of measurement precision. No single number characterizes the precision of the entire set for IRT scale score 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 document 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.

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

(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-ELA. The TCCs shows 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 ELA Forms A, B, and C

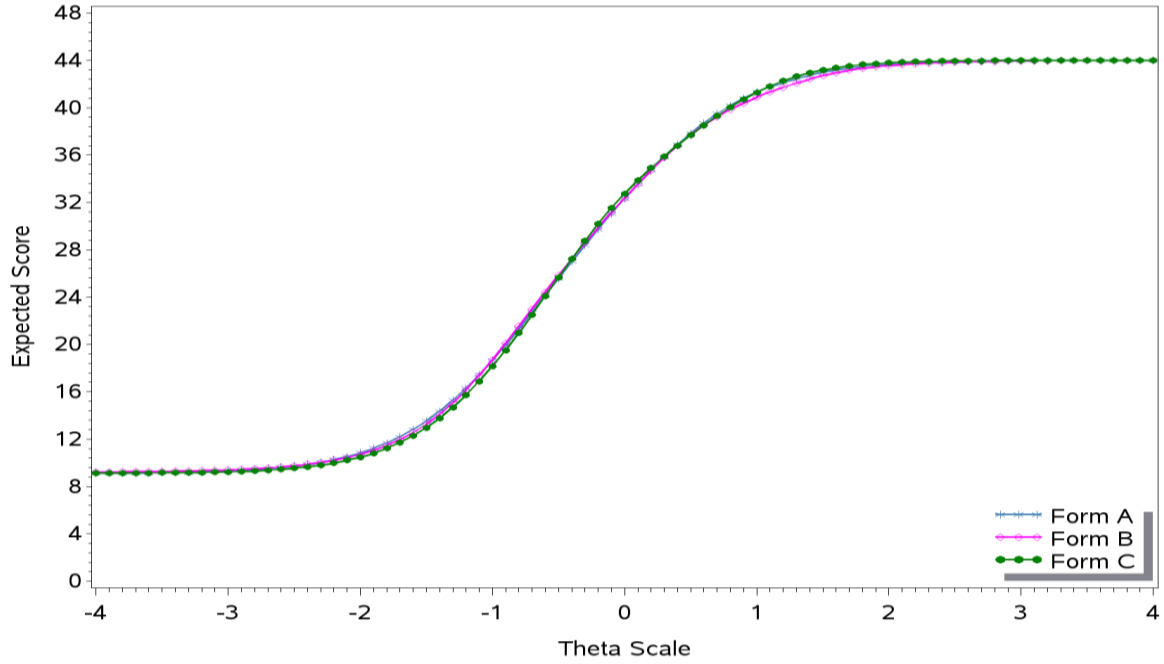


Figure 4.4 EOG Grade 4 TCC ELA Forms A, B, and C

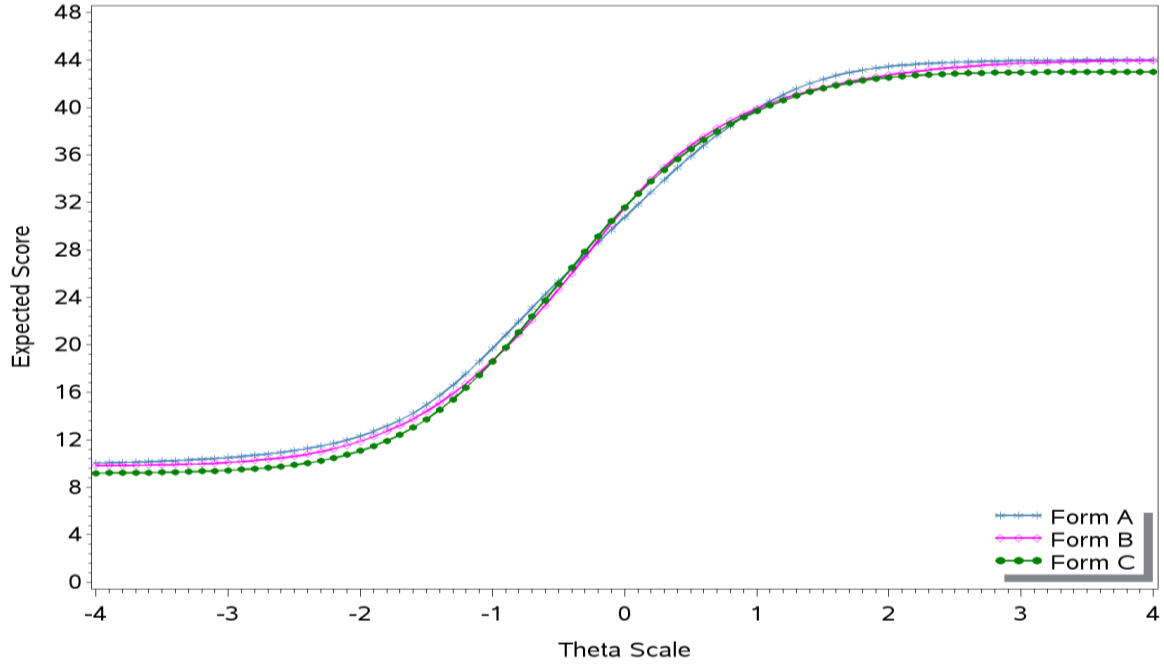


Figure 4.5 EOG Grade 5 TCC ELA Forms A, B, and C

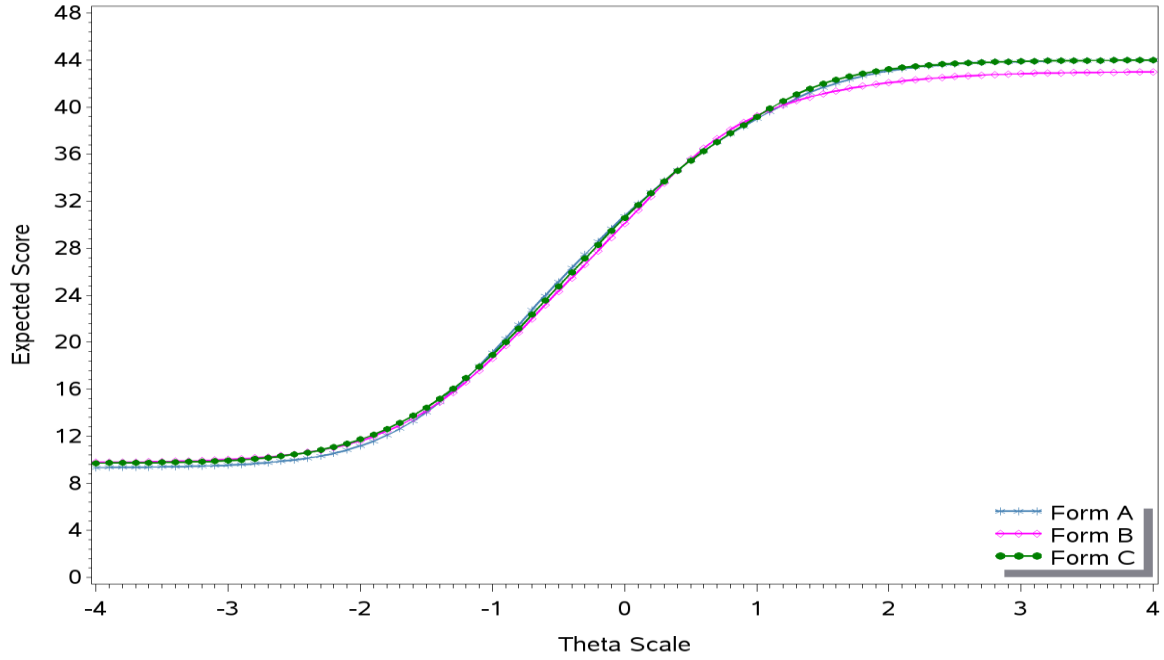


Figure 4.6 EOG Grade 6 TCC ELA Forms A, B, and C

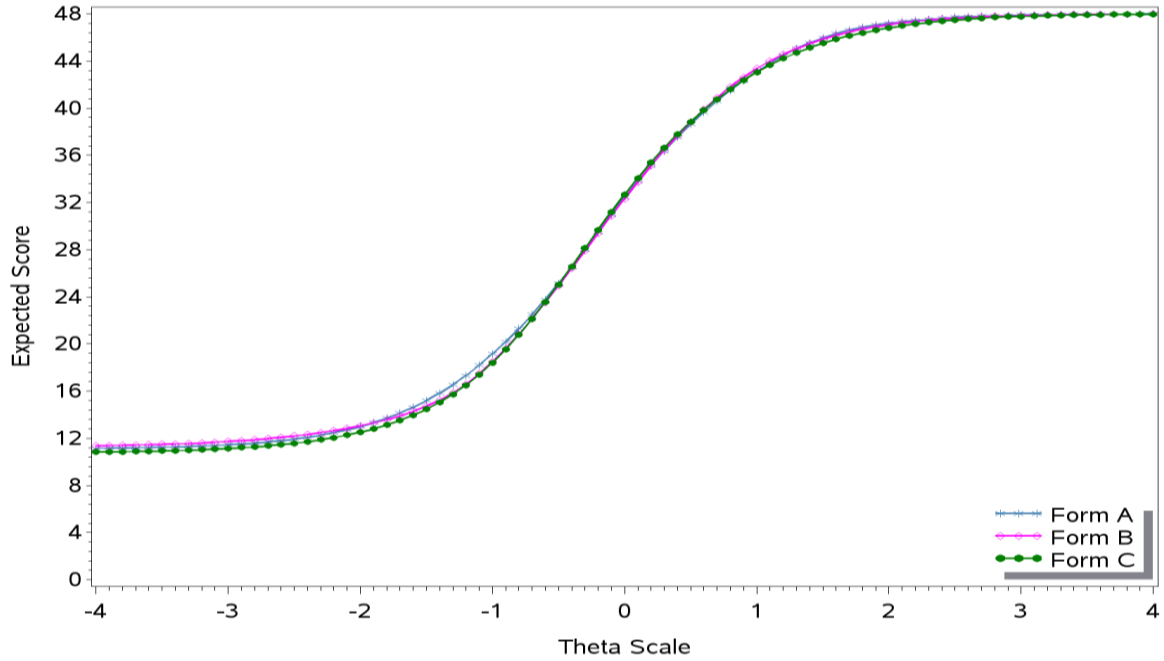


Figure 4.7 EOG Grade 7 TCC ELA Forms A, B, and C

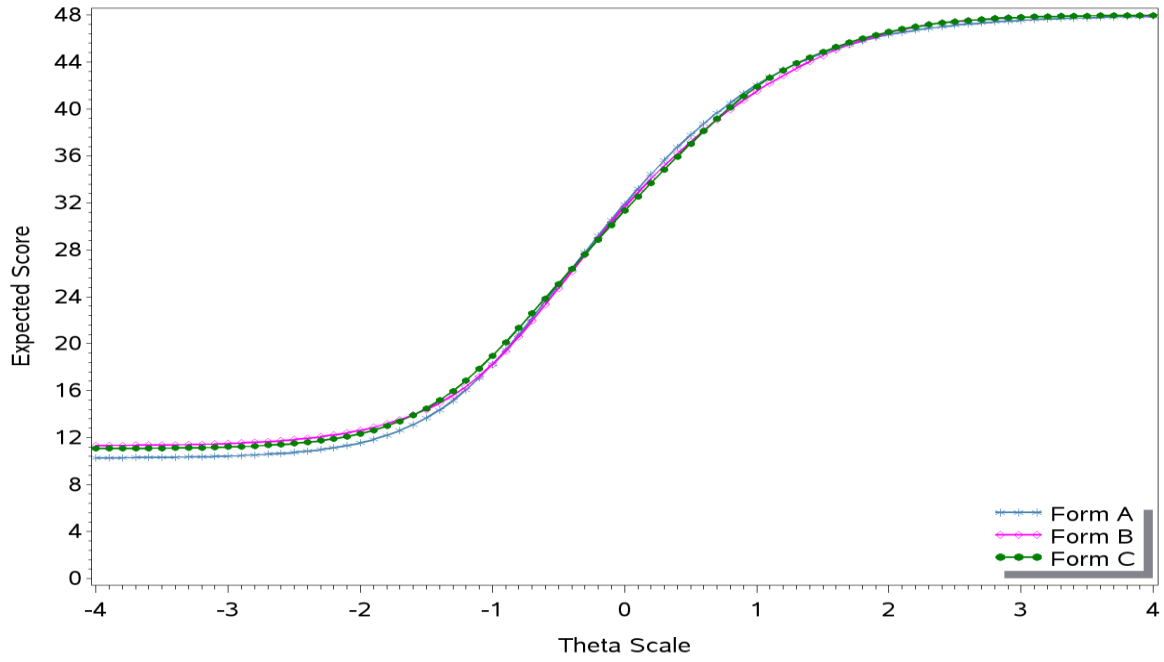


Figure 4.8 EOG Grade 8 TCC ELA Forms A, B, and C

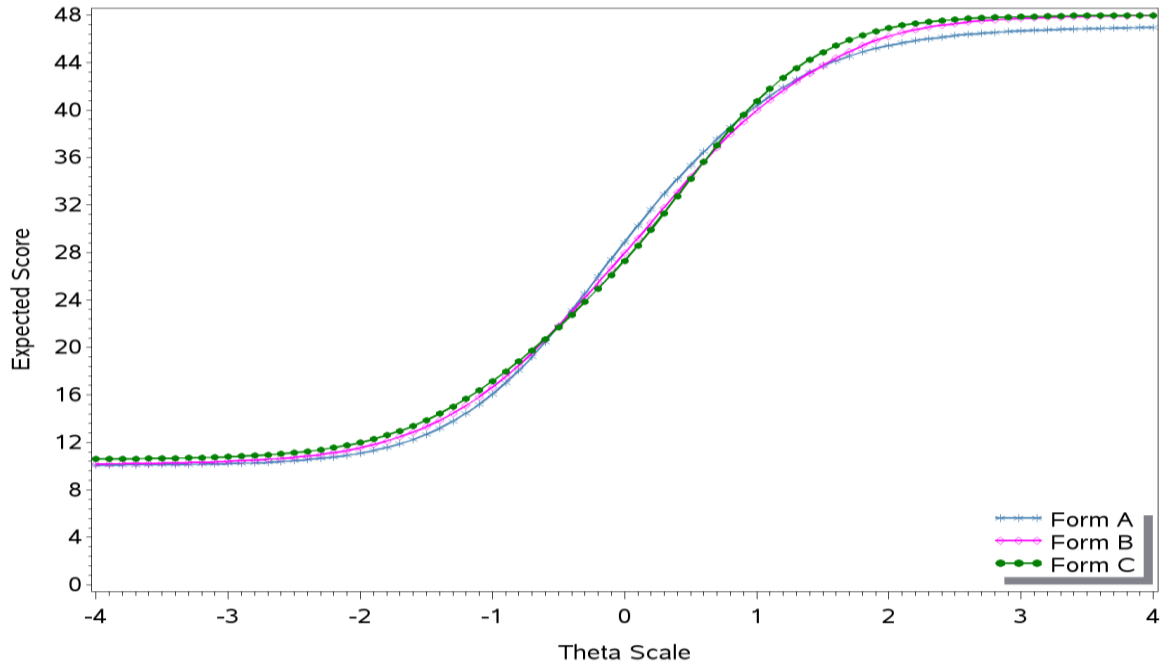
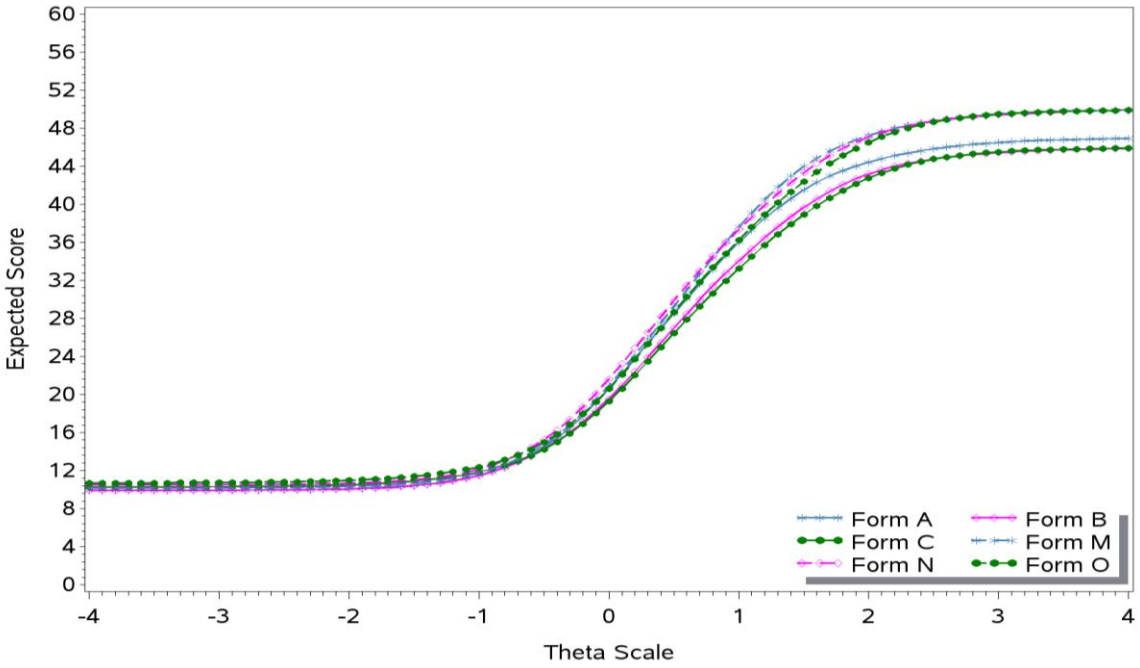


Figure 4.9 English II TCC forms A, B, C, M, N, and O



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 3.7) 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 is 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 and were reviewed by NCDPI and an 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 are exported from the Test Development System (TDS) application into the NCTest platform, a series of quality checks are performed to ensure all the specified interactions between items and the NCTest platform are fully functional across the different end users' approved devices. NSCU-TOPS and the NCDPI technology sections have instituted a five-phase quality check system 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 NC. 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 stated 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 include: 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* (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 described in the guides from NCDPI Testing Policy and Operations staff once a year for EOG assessments and twice a year for EOC assessments. Subsequently, the RACs provide training to Local Education Agency (LEA) test coordinators on the processes for proper test administration. LEA test coordinators then provide 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 ELL students), accommodated test administrations, test security (storing, inventorying, and returning test materials), and the *Testing Code of Ethics* (see Appendix 2-A).

5.3 Security Protocols Related to Test Administration

Test security is an ongoing concern in any testing program. When test security is compromised, it can undermine the validity of test scores. For this reason, 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* 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

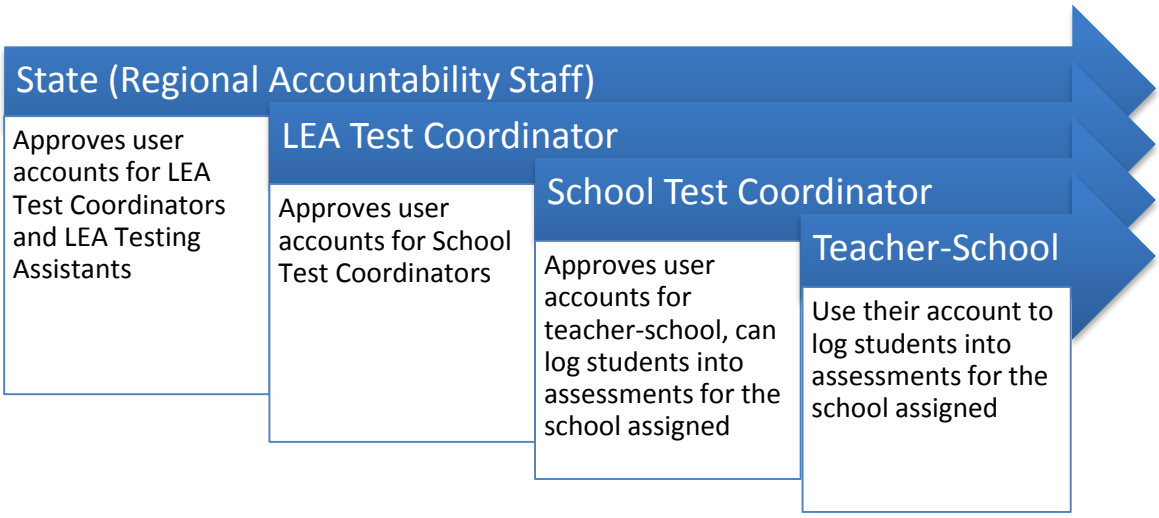
Test	Required Storage Time
All used answer sheets for operational tests (including scoring sheets for W-APT)	Six months after the return of students' test scores
Original responses recorded in a test book, including special print version test books (i.e., large print edition, one test item per page edition, 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: Algebra I/Integrated 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 English II 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 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 log in, the tests are sent securely from the NCTest server at NC State University to the local computer. Not all assessment content is sent at the time of login, only the text for all the test items 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 ELA 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 the EOG ELA is 180 minutes, with a maximum of 240 minutes. The estimated time allotted for EOC English II is 150 minutes, with a maximum of 240 minutes. For both the EOG and EOC, 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 English Language Learners (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 barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs." (the *Standards*, p. 67) Testing accommodations are defined as "changes in assessment materials or procedures that address aspects of students' disabilities that may interfere with the demonstration of their knowledge and skills on standardized tests." (Thurlow & Bolt, 2001, p. 3) Accommodations are provided to eligible students together with appropriate administrative procedures to assure that individual student needs are met and, at the same time, maintain sufficient uniformity of the test administration.

For any state-mandated test, the accommodation for an eligible student must (1) be documented in the student's current 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 in the *Assessment Guide*:

- 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
- 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 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)[j] 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 ELA include:

- Multiple testing session
- Testing in a separate room
- Student read aloud to self

Figure 5.2 ELL Proficiency Levels and Testing Accommodations

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

[j] 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®.

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 EOG and EOC test administration to 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 was available in both paper and computer mode.

The EOC English II assessment was developed as a computer-based fixed form. For the 2012–13 administration, districts could opt to use paper-based forms in place of the computer-based form. Beginning with the Fall 2014 administration, the state mandated all EOC English II assessments be administered as computer-based, fixed forms with the following exceptions:

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

Table 5.2 shows the total number of students who took ELA EOG and EOC test 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 from 2013 to 2015. In 2015, 87% of students took English II computer-based forms compared to 73% in 2013. For the EOG computer-based forms were administered for the first time in 2015 at grade 7, and approximately 20% of students took the computer-based form.

Table 5.2 EOG and EOC Test Administered by Mode

Type and Year		Administration Mode			
		Paper		Computer	
		Number of Assessments	Percent	Number of Assessments	Percent
EOG Grade 3	*2013*	106,518	100%	0	0
	2014	116,083	100%	0	0
	2015	118,510	100%	0	0
EOG Grade 4	*2013*	114,669	100%	0	0
	2014	107,388	100%	0	0
	2015	115,798	100%	0	0
EOG Grade 5	*2013*	114,435	100%	0	0
	2014	115,544	100%	0	0
	2015	108,385	100%	0	0
EOG Grade 6	*2013*	116,314	100%	0	0
	2014	115,280	100%	0	0
	2015	116,500	100%	0	0
EOG Grade 7	*2013*	115,381	100%	0	0
	2014	117,606	100%	0	0
	2015	92,935	79%	24,143	21%
EOG Grade 8	*2013*	112,944	100%	0	0
	2014	116,256	100%	0	0
	2015	118,869	100%	0	0
EOC English II	*2013*	29,988	27%	80,187	73%
	2014	22,050	19%	91,581	81%
	2015	15,529	13%	103,523	87%

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 English II 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 conditions 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 score scales. The first two sections of this chapter summarize the automated scoring procedures to transform students' responses into a number-correct score for fixed response items and the human scoring process for assigning score category for constructed-response items. Section three and four 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. 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

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 as described in the assessment guide. The regional accountability coordinator (RAC) has the responsibility of scanning and scoring tests for charter schools and for providing long-term storage for specific test materials such as used answer sheets and used test books (only available for the *Student Marks Answers in Test Book* accommodation).

Computer-mode forms are scored electronically via a centrally-hosted server at NCDPI using WinScan software. Once WinScan assigns scores for each item, data are 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.

6.2 Constructed Response Scoring

This section briefly describes the scoring process for constructed response (CR) items administered operationally in 2012–13 and beyond. Questar Assessment Inc. (QAI) is the scoring partner of NCDPI.

6.2.1 Transportation and Processing

There are three operational CR items in each EOC English II form. The forms are administered in both computer and paper modes. For scoring CR items in paper mode, Districts/Schools receive shipping labels from QAI to ship answer documents directly to QAI's facility. For CR items administered on computer, the student test records are transferred daily as Online Response Data File via NCDPI's secured File Transfer Protocol (FTP) site. The FTP serves two primary purposes: exchanging administrative documentation and exchanging student test material. The Student Test Data File Report with scored data are delivered by QAI to NCDPI within 14 business days after the administration has ended.

6.2.2 Rater Selection, Training and Qualification

AERA/APA/NCME (2014) Standard 4.20 specifies the following:

"The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers' responses that illustrate the levels on the rubric score scale, and

the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters' scoring." (p. 92)

1. Project Staffing

In general, QAI uses a hierarchy of Scoring Directors, Team Leaders, and Scorers. Scoring Directors are chosen for a project based on the following qualifications:

- 4-year degree
- Content expertise
- Previous project experience
- Experience with ScorePoint
- Ability to work under pressure to meet deadlines
- Ability to travel, facilitate, and interact with client
- Possesses good work ethic and integrity
- Good verbal and written communication skills
- Evaluations
- Schedule Flexibility

The Scoring Directors have the overall responsibility for the training of the project and content as well as the scoring expectations. They undergo extensive specialized training to prepare them for their roles as scoring experts and monitors by working with QAI or department content specialists.

Team Leaders report directly to the Scoring Directors and are typically in charge of a team of 10–12 scorers, depending on the item(s) and content area. They are specifically trained on the requirements and processes for scorer monitoring and intervention, including interpreting ScorePoint reports such as, Reader Reliability (RR) and Score Point Distribution (SPD) reports, conducting read behinds, holding one-on-one discussions, and scoring.

Team Leaders (TLs) are selected based on:

- 4-year degree
- Content knowledge
- Previous project experience
- Experience with ScorePoint (QAI proprietary system)
- Evaluations

Scorers must have fulfilled the following requirements:

- 4-year degree (in a related field in the content area for which they will be scoring as appropriate)
- Attend an open house for an introduction to Questar philosophy
- Complete an application process, complete with references
- Complete a sample of the content area for which they are applying
- Complete a one-on-one interview with Questar scoring staff

2. Training

Training Materials

AERA/APA/NCME (2014) Standard 6.8 states:

> *"Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented."* (p. 118)

Training materials for North Carolina include responses scored during rangefinding that represent the full range of score points as determined by the rangefinding committees, including responses that exemplify the nuances of the rubric (e.g., differentiation of a low "3" from a high "2").

Training materials consisted of the following:

- One **Passage**
- One **Prompt and Rubric**
- One **Scoring Guide (or Guide Set)** containing approximately 10 items with a minimum of 3 anchor responses (1 for each score point). During training, the Scoring Guide was discussed response by response within the group setting to identify any nuances of individual responses that have been selected as exemplary. This phase also includes a discussion of often seen acceptable and non-acceptable details for each item.
- A **Training Set** containing 10 responses, representing a variety of score points in random order. The training set was scored independently by each scorer, and each

response was discussed by the group. This set is used as a learning tool to assess whether the scorer understands the nuances as discussed in the Scoring Guide.

- A **Qualifying Set** containing 10 responses, representing a variety of score points in random order. The qualifying set is scored independently by each scorer, and each response is discussed by the group. This set was used to determine whether a scorer is eligible to continue on to scoring. Meeting the qualification standards on this set demonstrates that the scorer will be able to apply the necessary skills to score.

Team Leader Training

AERA/APA/NCME (2014) Standard 6.9 specifies:

> *"Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected."* (p. 118)

To meet this requirement, NCDPI's scoring vendor, QAI, had their Scoring Directors carefully selected and trained only the most qualified people to be Team Leaders. The Team Leaders were trained prior to scorers, so they were familiar with all of the training materials and the scoring procedures prior to scorer training.

Scorers were divided into teams, and each scorer was assigned a unique scorer identification number. That identification number allowed for the tracking of scorer performance via the scorer quality control reports throughout the online scoring.

Once the training staff was confident that the scorers understood and had an awareness of the need to be sensitive to the performances of students, nondisclosure forms were signed and training began.

Scorers, like Team Leaders, were required to meet the qualification standards before scoring student responses. Any scorer who was unable to meet the qualifying standards was dismissed–a stipulation understood by all scorers when they are hired. The qualification standard was 80% exact agreement on rubrics.

Prior to actual scoring, the scorers did the following:

- Signed a nondisclosure agreement
- Acknowledged the QAI harassment policy

- Reviewed NCDPI expectations and goals
- Set aside any biases they may have about students, student work, and the scoring criteria presented
- Trained to use the ScorePoint online scoring system

Once scorers were instructed on the above, individual training included the following process:

- Scorers were trained on the Scoring Guide, including discussion of the rubric, presenting the task or item (i.e., graphics and all related assets), reviewing the eligible score points, followed by group participation and discussion of each response using examples and annotations as appropriate. Questions by scorers were addressed as a group for consistent messaging and decisions.
- Scorers then completed a training set independently to assess their grasp of the scoring.
- Each response in the training set was reviewed with the group with an explanation and examples as needed to ensure scorer consistency on the nuances of each response and score point.
- Scorers completed a qualifying set independently. Results using the qualification criteria determined if they were allowed to score that particular task type.
- In addition, each nonscoreable code was explained and examples were provided as available. All nonscorable answers were assigned a code. Examples included blank (BL), illegible (IL), foreign language (FL), repeating prompt (RP), off topic (OT), incoherent (IC), and other reasons (OR).
- Protocol for "alerting" responses that require attention was discussed at this time.

Following the successful completion of training and qualifying, scoring center staff activated individual scorers in the system, allowing them to score student responses.

3. Qualification

In order to score an item, the scorer had to meet the qualifications standards for scoring. The qualification standard for all items was 80% each agreement. Successful completion of training also requires a minimum acceptable agreement rate of 80% on the task. A scorer can be

dismissed if retraining does not elicit satisfactory results or if it is determined that a scorer is not accurately scoring student responses.

6.2.3 Monitoring the Scoring Process

Scoring Directors and Team Leaders live monitor the scoring process in terms of valid responses, ongoing training, one-on-one discussion, and read-behinds. There are two kinds of read behinds used: random read behinds and prescribed read-behinds. The random read behinds are a part of the daily ongoing monitoring process, while prescribed read behinds are done in case something arises during the scoring. The read behinds may result in a change in a student's score. QAI also produces item reliability and score point distribution reports weekly as a part of monitoring reliability and validity of the scoring. The report includes the number of responses scored, agreement rates, and score distribution. For more details refer to Appendix 6-A NC Scoring Process – English II.

6.2.4 Inter-rater Agreement

NCDPI requires 10% of the random responses receive two readings as a part of the inter-rater agreement calculation. *Table 6.1* shows exact and adjacent agreement rates for the English II CR items from Fall 2012–Spring 2015 by administration. The results indicate that the exact agreement rates by item range from 82.7% to 98% with an average agreement rate over all items of 91.5%.

Table 6.1 Rater Agreement Rates by Administration and Mode Fall 2012–Spring 2015

Administration	Form A/M Agreement Rate (%)				Form B/N Agreement Rate (%)				Form C/O Agreement Rate (%)			
	Item	N	Exact	Adjacent	Item	N	Exact	Adjacent	Item	N	Exact	Adjacent
Fall 2012	#1	2,918	93	6	#1	2,767	94	6	#1	2,784	97	3
	#2	2,897	95	5	#2	2,858	96	4	#2	2,851	94	6
	#3	2,933	97	3	#3	2,726	95	4	#3	2,900	96	4
Spring 2013	#1	1,049	88	12	#1	1,027	88	12	#1	4,184	90	10
	#2	4,009	89.7	10.3	#2	4,200	93.5	6.5	#2	4,060	82.7	17.3
	#3	4,135	88.7	11.3	#3	4,090	91.7	8.3	#3	4,258	86.3	13.7
Fall 2013					#1	4448	86.2	13.8	#1	4404	85.3	14.7
					#2	4544	87.4	12.6	#2	4492	86.0	14.0
					#3	4382	83.3	16.7	#3	4552	86.7	13.3
Spring 2014					#1	4812	97.1	2.8	#1	4852	84.9	15.1
					#2	4766	95.9	4.1	#2	4928	93.2	5.9
					#3	4534	97.1	1.9	#3	4848	94.1	5.9
Fall 2014	#1	3,008	85.2	14.8	#1	3204	92.3	7.7	#1	2982	89.2	10.8
	#2	2932	83.5	15.5	#2	3298	90.7	9.3	#2	3040	94.0	6.0
	#3	3072	94.2	4.8	#3	3202	89.3	10.7	#3	3114	93.2	6.8
Spring 2015	#1	4162	90.5	9.5	#1	4144	93.3	6.7	#1	4472	89.2	10.8
	#2	3998	94.2	5.8	#2	4428	98.2	1.8	#2	4524	95.8	4.2
	#3	4350	96.5	2.7	#3	4302	97.4	2.6	#3	4406	94.2	5.8

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

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

(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. 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.

6.4 Developmental Scale for ELA EOG 3–8

The NDPI contracted with Pacific Metric Corporation to create a vertical developmental scale for ELA EOG 3–8 (see Appendix 6-B Developmental Scale for ELA.).
Data for the developmental scale was collected during the 2013 administration of ELA EOG following an embedding designed implemented by NCDPI.

Linking sections, which were administered to students in adjacent grades were embedded within operational forms. For example, some 5th-grade operational items were embedded into the 6th grade EOG form; the linking items did not count toward the 6th-grade students' scores. The linking plan only extended up, not down. For example, 5th grade items were embedded in 6th grade forms, but 6th-grade items were never embedded in 5th-grade forms. The developmental scale was derived by fixing the mean and standard deviation of grade 5 ELA at 450 and 10, then chain linking the other adjacent grades.

The difference in performance between grades on these linking items was used to estimate the difference in proficiency among grades. The flexMIRT™ version 1.88 (Cai, 2012)

described in Williams, Pommerich, and Thissen (1998) was used in creating the vertical developmental scale. The procedure was divided into four steps.

Step 1. flexMIRT™ was used to calibrate data from EOG assessments' item and population parameters for adjacent grades. It resulted in average mean difference and average standard deviation ratios (m_i and s_i) for each grade. Individual runs in flexMIRT™ were conducted for each of the grade-pair links. For ELA, each grade pair for grades 3 through 8 had twelve links (six below-grade and six above-grade). The linking sets varied between six and eight items, and each linking set was associated with a reading passage. Under the assumption of equivalent groups, the form results were averaged within grade pairs to produce one set of values per adjacent grade. Outlier values were dropped if they were greater than two standard deviations from the mean. Three sets of values were dropped as outliers—one each from the 3–4, 6–7, and 7–8 grade pairs. *Table 6.2* displays the average difference in adjacent-grade means and standard deviation ratios for the EOG ELA/Reading.

Table 6.2 Average Mean Difference in Standard Deviation Units Spring 2013 Item Calibrations

Grades	Average Mean Difference	Average Standard Deviation Ratio	Number of Grade-Pair Forms
3–4*	0.550	0.948	11
4–5	0.387	0.968	12
5–6	0.270	1.099	12
6–7*	0.298	1.011	11
7–8*	0.242	1.021	11

Note: An asterisk (*) denotes that one outlier was removed from the average for this grade pair

Step 2. In flexMIRT™, grade 3 was considered the reference group; its population mean and standard deviation were set to 0 and 1, respectively. The above-grade mean and standard deviation were estimated using the scored data and the IRT parameter estimates. These parameters were provided in the flexMIRT™ output and did not require independent calculation. Theoretically, a (0,1) growth scale anchored at grade 3 was constructed to yield the means ($M_i = M_{i-1} + m_i * S_{i-1}$) and standard deviations ($S_i = s_i * S_{i-1}$), for Grade i on (0,1) growth scale anchored at the lowest grade (with grade 3 indexed as i=3), where $M_2 \equiv 0$, and $S_2 \equiv 1$. This (0,1) growth scale was generated recursively upwards from grade 3 to grade 8.

Step 3. The scale was re-centered (re-anchored) at grade 5, yielding $M_i^* = \frac{(M_i - M_5)}{S_5}$ and $S_i^* = \frac{S_i}{S_5}$ as the means (M_i^*) and standard deviations (S_i^*).

Step 4. The final step in constructing the developmental scale was the application of a linear transformation in order to produce a developmental scale with the grade 5 mean and standard deviations equal to 450 and 10, respectively. For example, $\mu_i = 450 + M_i^*$ and $\sigma_i = 10S_i^*$, where μi is the mean of the final developmental scale in grade i and σi is the standard deviation for the developmental scale in grade i. The resulting Fourth Edition (2013) vertical developmental scales across grades are shown in ***Table 6.3***. For detail procedures please refer to this document:
http://www.ncpublicschools.org/docs/accountability/testing/technotes/devscaleeela1213.pdf

Table 6.3 Developmental Scale Means and Standard Deviations ELA EOG 2013

EOG	Mean	Standard Deviation
Grade 3	440.01	10.90
Grade 4	446.00	10.33
Grade 5	450.00	10.00
Grade 6	452.70	10.99
Grade 7	455.97	11.12
Grade 8	458.66	11.35

For the succeeding administrations of the EOG, the developmental scale was adjusted to population mean and standard deviation from the previous administration. For example, the mean and standard deviation for a given grade for 2012–13 population was used to scale for the 2013–14 administration and so on.

6.5 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, point-biserial, Cronbach alpha) and IRT-based analysis (item calibration and scoring, test characteristics curves, test information functions, and conditional standard errors).

7.1 Pre-Equated Parallel Forms 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 six alternate forms in EOC English II (A, B, C, M, N, O) were administered operationally for the first time in the 2012–3 school year. At every grade level, all parallel 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 on put on the same IRT scale and can be compared directly without need for equating.

Table 7.1 and *Table 7.2* show demographic descriptive summaries for students who were administered ELA EOG and EOC in 2012–13. The student counts listed in these tables are the number of valid tests administered, not the actual official enrollment records. The actual difference between the total student population and sample included in item analysis is trivial and given the very large sample sizes at every grade, such differences are not expected to impact final item and test statistics reported. On average, over 100,000 students per grade level at grades 3 through 8 and in high school were administered the EOG ELA or EOC English II assessments. For EOG grades 3–8 at least 35,000 were administered one of the three parallel forms. The differences across forms within grade are negligible, which is evident of the success of the random spiral process. In EOC English II, over 26,000 students were administered one of the three computer-based parallel forms, and about 10,000 students were administered one of the three parallel 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 ELA EOG Operational Test 2012–13

Grade and Form		N	Gender (%)		Ethnicity (%)						
			Female	Male	Asian	Black	Hispanic	American Indian	Multi-racial	Native Hawaiian/ Pacific Islander	White
ELA 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**
ELA 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**
ELA 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**
ELA 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**
ELA 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**
ELA 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**

Table 7.2 Student Demographic Summary for EOC English II Operational Test 2012–13

Grade and Form		N	Gender (%)		Ethnicity (%)						
			Female	Male	Asian	Black	Hispanic	American Indian	Multi-racial	Native Hawaiian/ Pacific Islander	White
English II	**A**	10,115	49.70	50.30	3.84	34.73	12.46	0.79	3.80	0.11	44.28
	B	9,801	49.12	50.88	3.54	34.15	12.08	0.56	3.30	0.15	46.22
	C	9,723	49.52	50.48	4.08	33.97	11.90	0.72	3.94	0.12	45.26
	M	26,569	49.51	50.49	1.95	24.35	11.14	1.81	3.29	0.12	57.34
	N	26,650	48.95	51.05	2.28	24.07	11.21	1.79	3.50	0.08	57.08
	O	26,382	49.31	50.69	2.19	24.14	11.03	1.68	3.25	0.09	57.62
	All	**109,240**	**49.31**	**50.69**	**2.60**	**26.93**	**11.41**	**1.47**	**3.44**	**0.10**	**54.06**

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—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. In all, three items were dropped from the operational set and scaled around: an item from EOG Grade 4 form C, an item from EOG Grade 5 form B, and an item from EOG Grade 8 form A. These forms are marked with asterisk in *Table 7.3*. Each EOG form was calibrated separately using the single-group design and 3PL IRT model to establish the final IRT parameters for scaling. In IRT, the need for equating is a non-issue if parameters from parallel forms are located on the same IRT scale either through the data collection design, as is the case with random spiraling of forms, or through concurrent calibration method. Once all items are calibrated onto the same IRT scale, then raw-to-scale tables are created for each parallel 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*."

7.3.2. EOC IRT Calibration Across Modes

For English II, all operational items in the three pairs of parallel forms (A and M, B and N, and C and O) created from field test data were reviewed using the psychometric criteria presented in section 4.5.1 Following these analyses, one item from EOC English II forms B and N was effectively removed from the final operational forms and scaled around.

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) 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, and there no artificial restriction 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 difference between the TCCs from operational and field test administration are that the gradient of the operational TCCs is slightly lower, and the steepest section 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 ELA Operational Forms A, B, and C

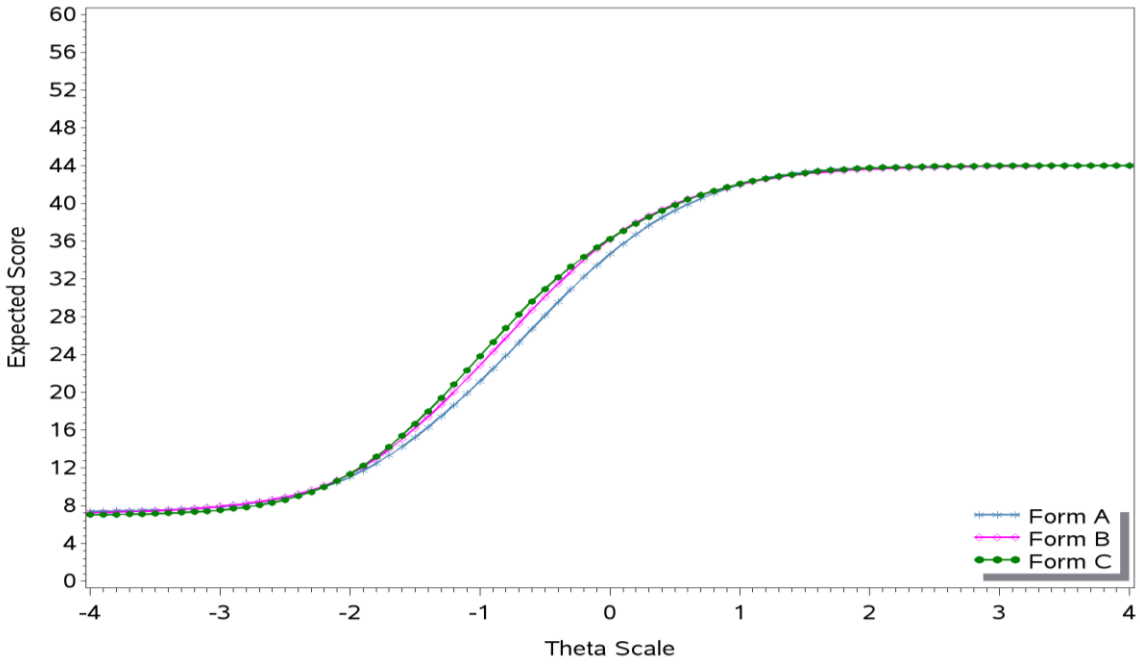


Figure 7.2 Grade 4 TCC ELA Operational Forms A, B, and C

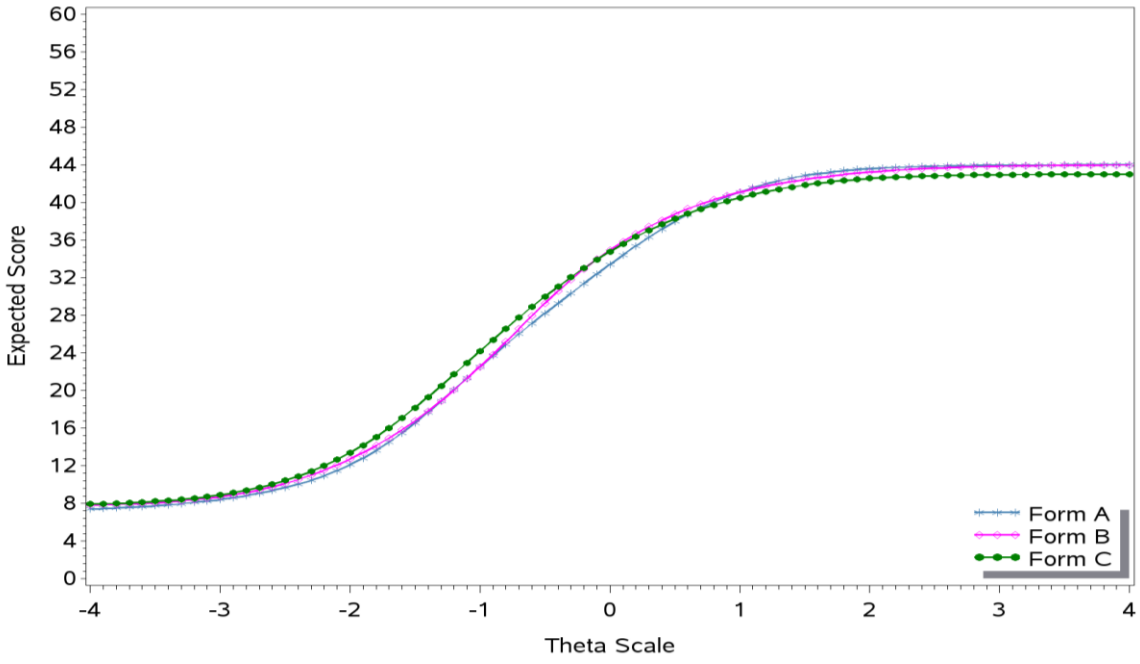


Figure 7.3 Grade 5 TCC ELA Operational Forms A, B, and C

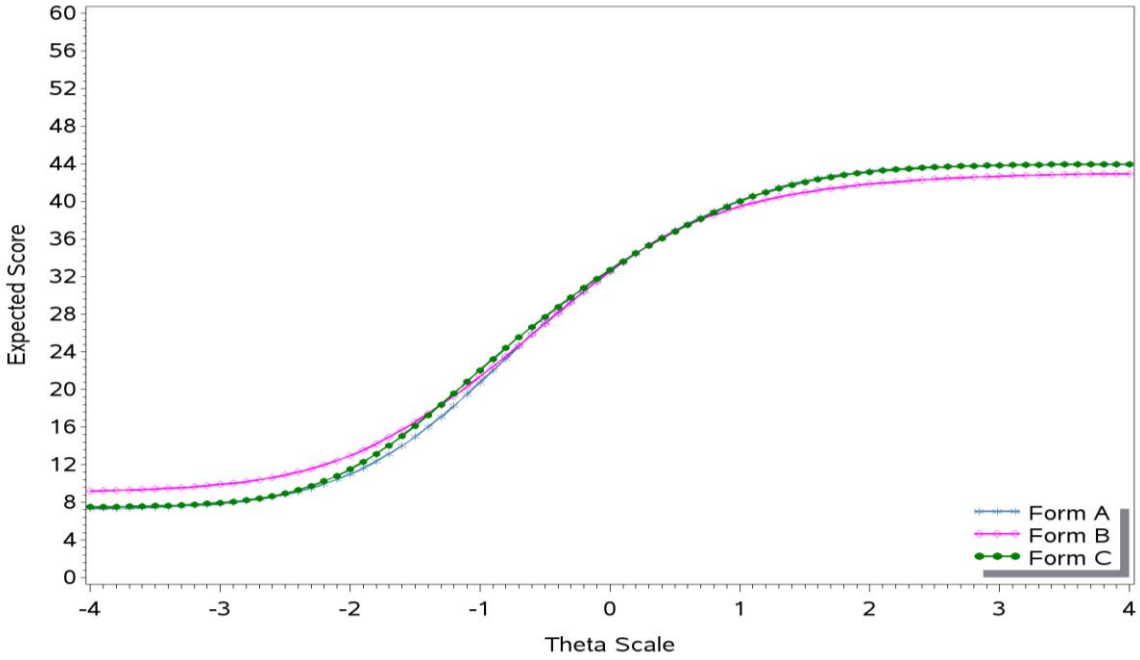


Figure 7.4 Grade 6 TCC ELA Operational Forms A, B, and C

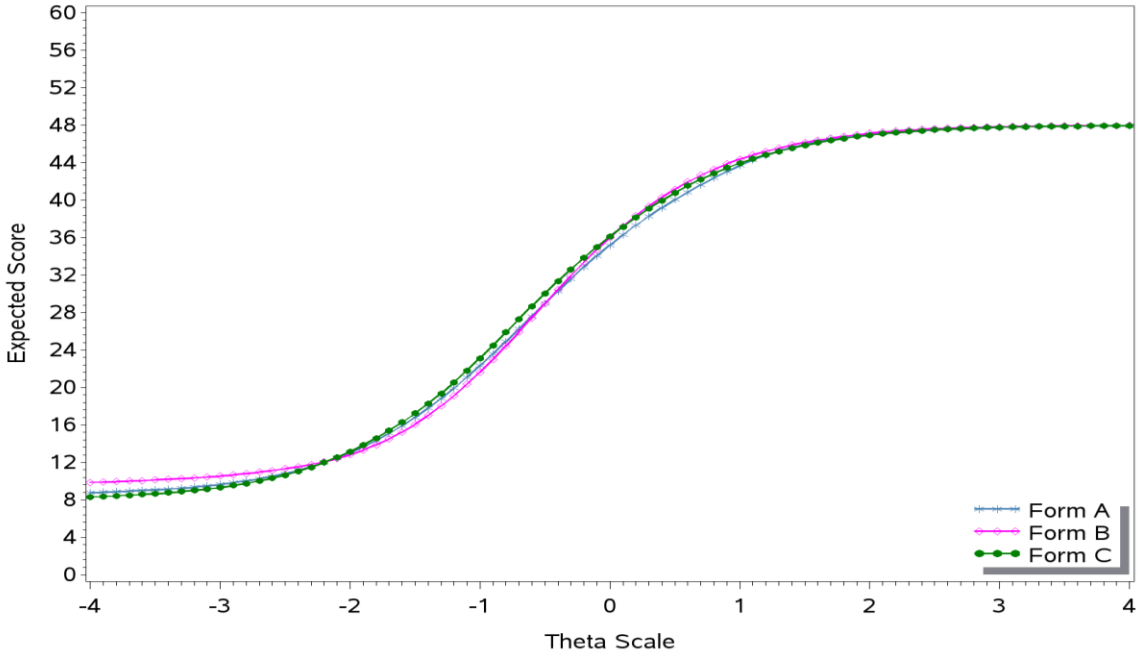


Figure 7.5 Grade 7 TCC ELA Operational Forms A, B, and C

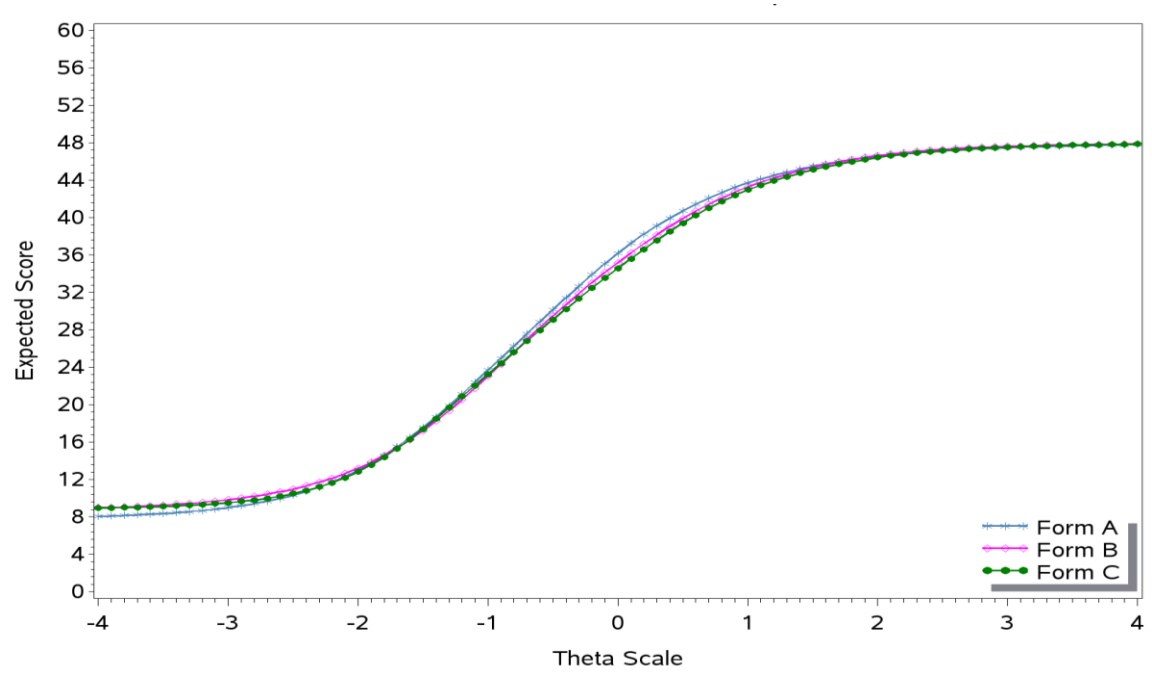


Figure 7.6 Grade 8 TCC ELA Operational Forms A, B, and C

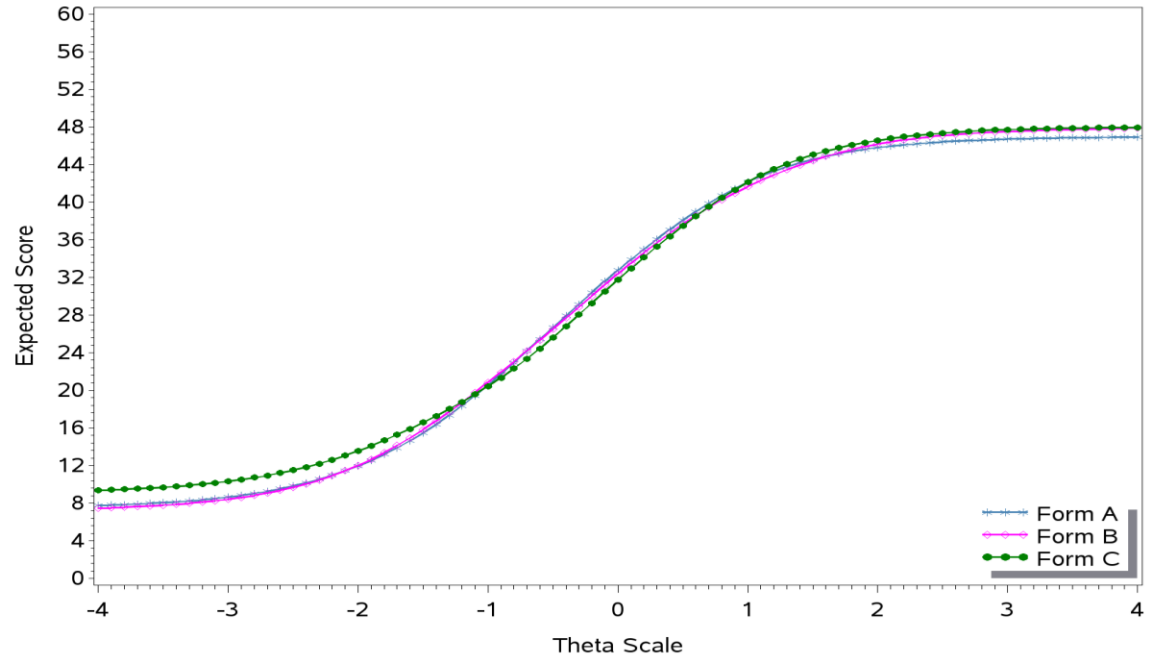
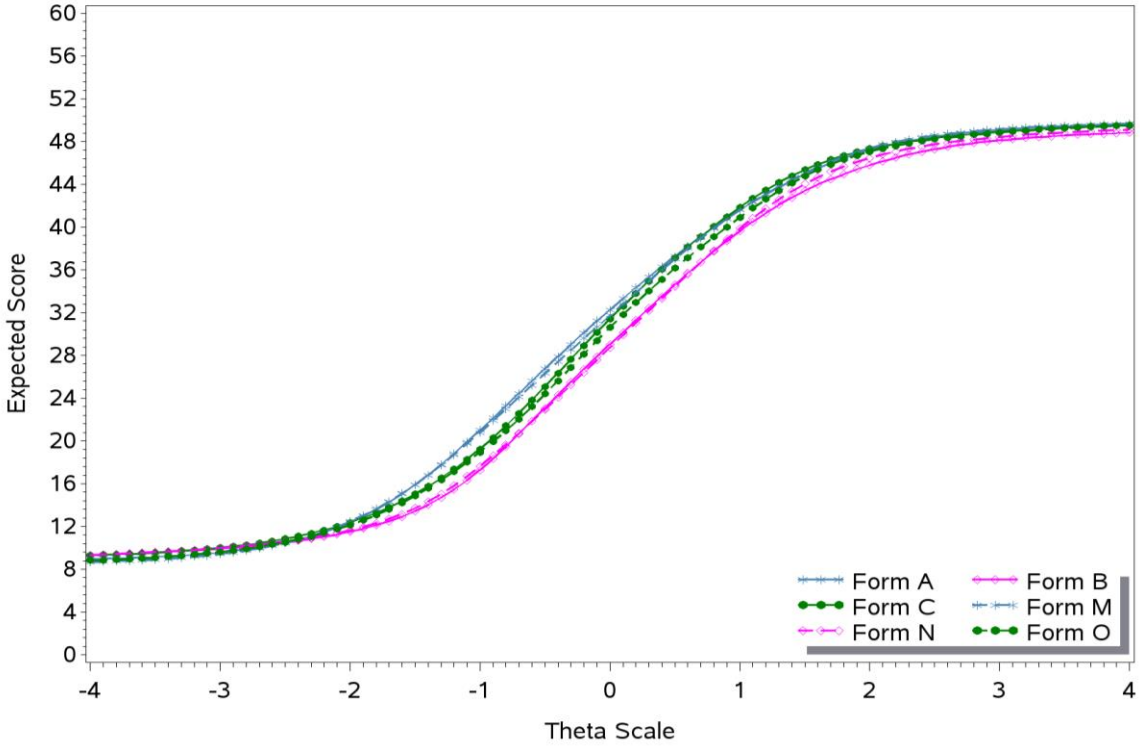


Figure 7.7 English II TCC ELA Operational Forms A and M, B and N and C and O



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 alternate forms, the goal was to select items that minimize the differences between TCCs of alternate forms. As a result, the displayed TIFs for alternate forms are not as closely uniform as the TCCs. The implication is that relative efficiency of alternate forms varies slightly. But overall, the forms are most efficient between theta range of -1 and 1.

In terms of standard errors, the figures show they are inversely related to TIF across all forms and are lowest between the theta range 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 ELA Grade 3 Test Information and Standard Errors for Operational Forms

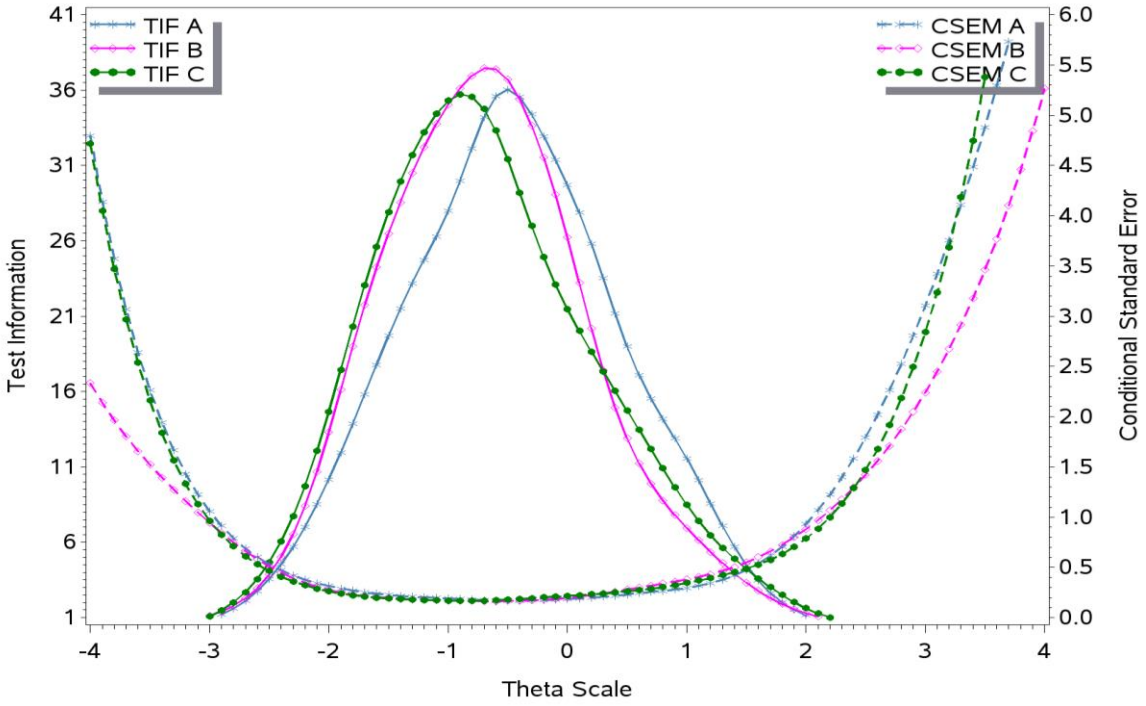


Figure 7.9 ELA Grade 4 Test Information and Standard Errors for Operational Forms

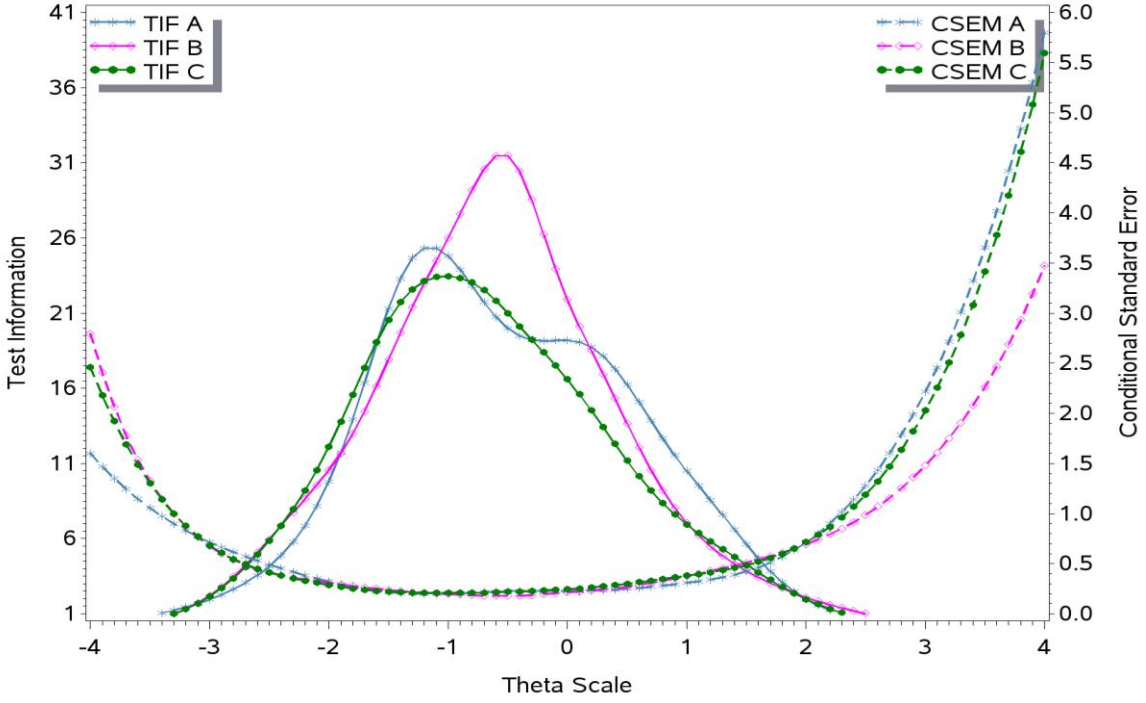


Figure 7.10 ELA Grade 5 Test Information and Standard Errors for Operational Forms

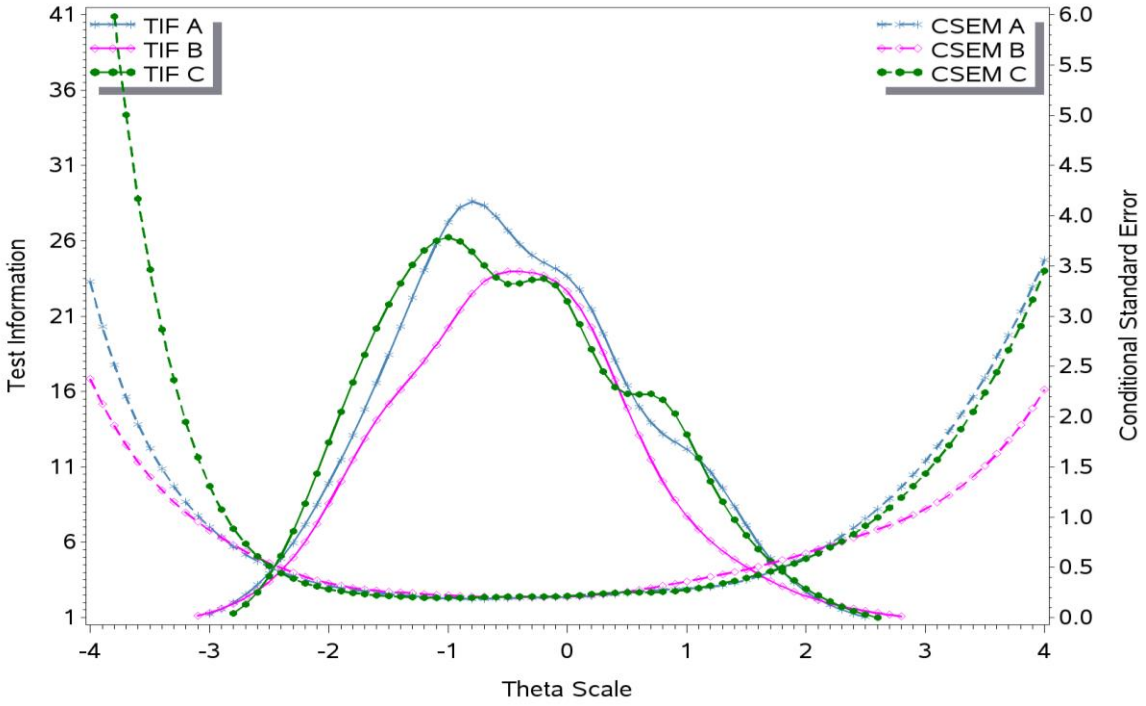


Figure 7.11 ELA Grade 6 Test Information and Standard Errors for Operational Forms

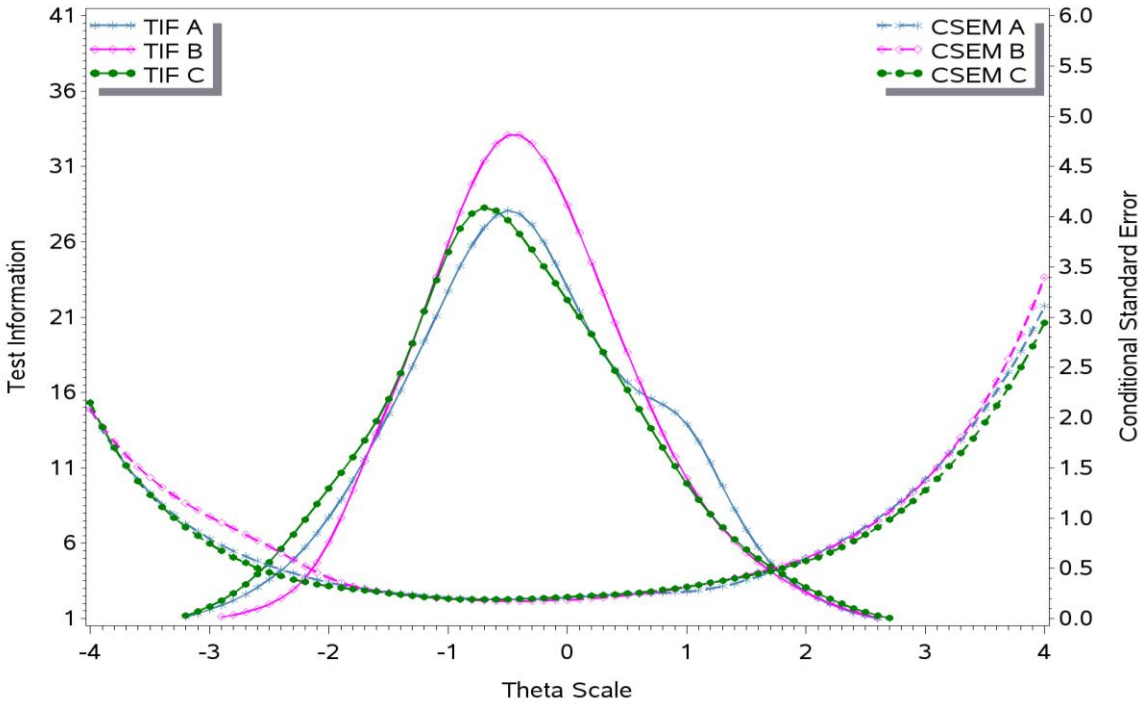


Figure 7.12 ELA Grade 7 Test Information and Standard Errors for Operational Forms

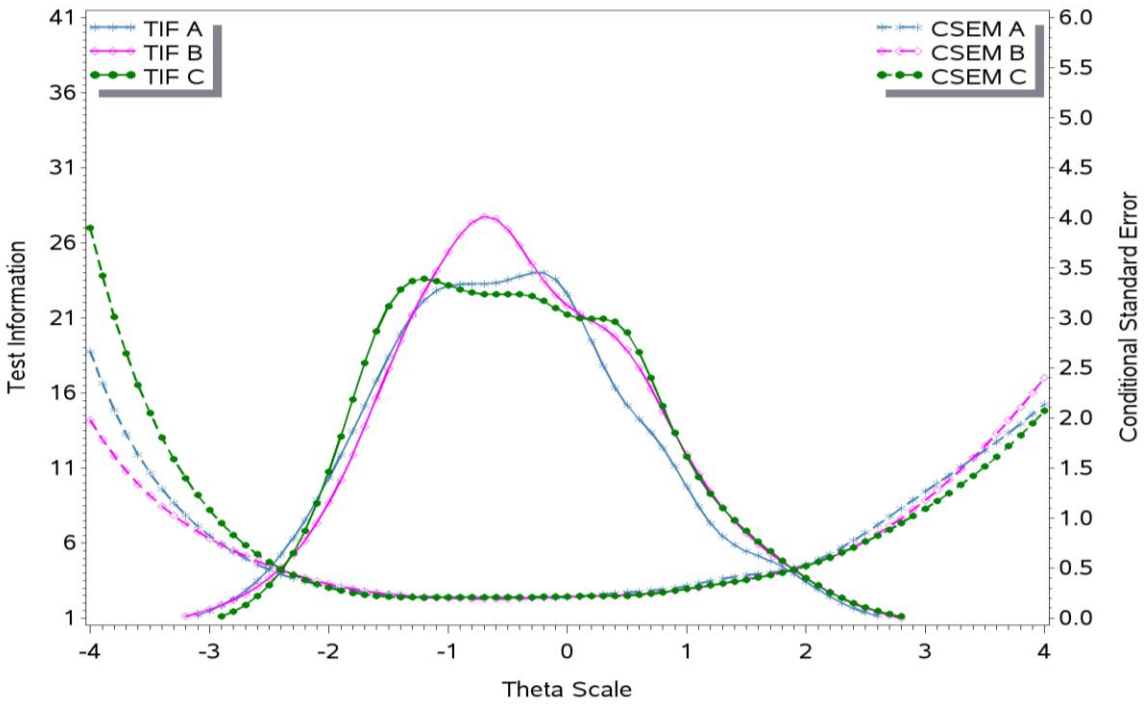


Figure 7.13 ELA Grade 8 Test Information and Standard Errors for Operational Forms

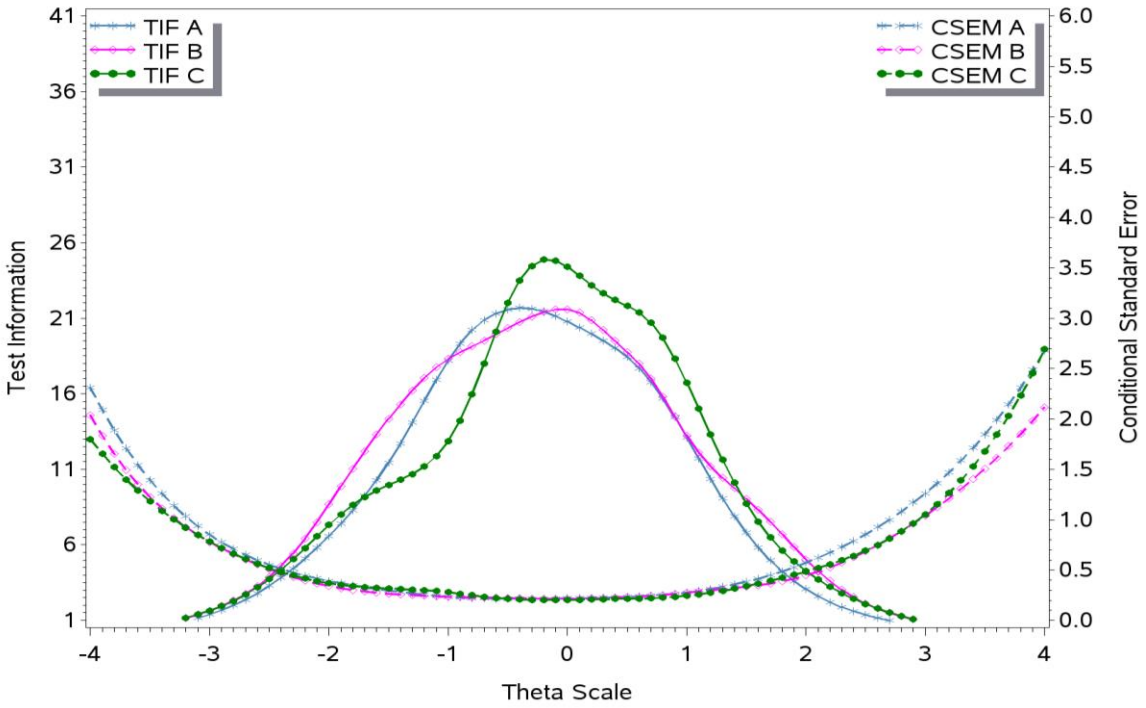
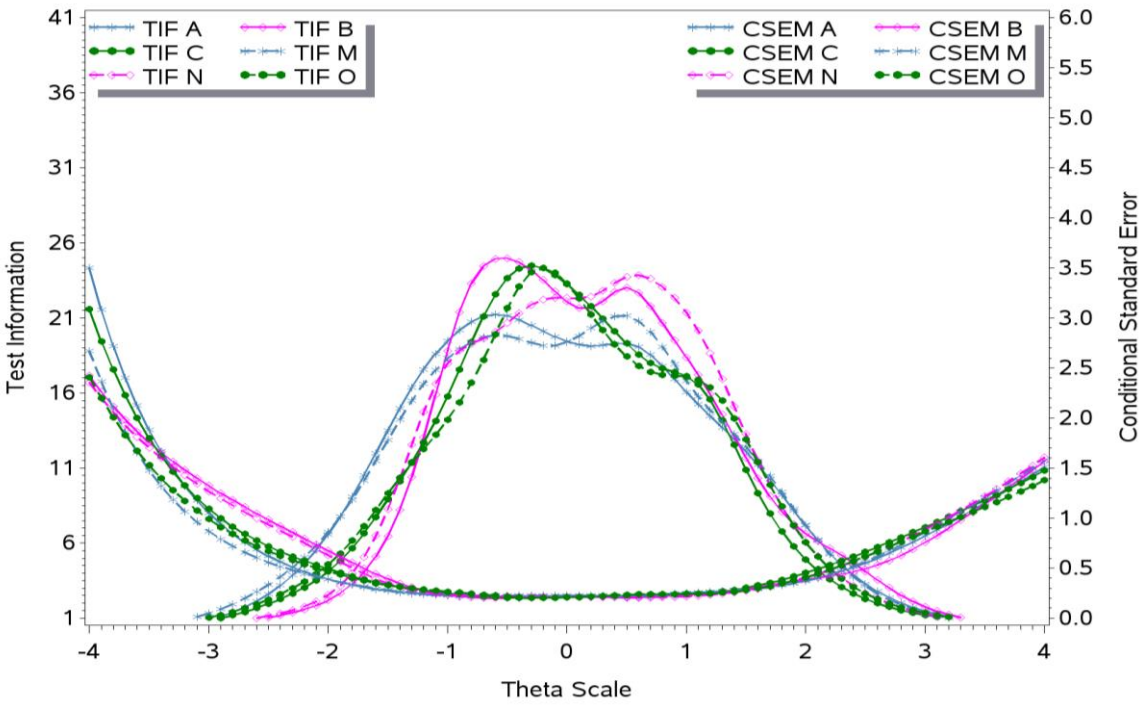


Figure 7.14 English II Test Information and Standard Errors for Operational Forms



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 7.3 and *Table 7.5* present comparison form-level average CTT summary statistics (p-values and 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 0.02 to 0.06 across all EOG grades 3–8. For English II, the average p-value difference ranged from 0.12 to 0.18 across 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 0.88 to 0.92, which is acceptable for tests of this length.

IRT parameters calibrated using field test data and again after the operational administration are presented in *Table 7.4* and *Table 7.6*. A similar trend as noted in the p-values was confirmed by the IRT b parameter. The ICC's 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 summary 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. The difference in b-parameter was most pronounced in English II, where the median absolute difference was between 0.5 and 1 across the forms.

Table 7.3 CTT Average Descriptive Statistics for ELA EOG 2012–2013

EOG		Number of Items	Field Test CTT Summary		Operational Test CTT Summary		
			Pvalue	Biserial Correlation	Pvalue	Biserial Correlation	Reliability (Cronbach Alpha)
ELA Grade 3	A	44	0.69	0.46	0.71	0.46	0.91
	B	44	0.69	0.46	0.74	0.47	0.92
	C	44	0.69	0.47	0.74	0.47	0.91
ELA Grade 4	A	44	0.67	0.42	0.71	0.42	0.89
	B	44	0.67	0.43	0.72	0.43	0.90
	C*	43	0.68	0.43	0.74	0.41	0.88
ELA Grade 5	A	44	0.66	0.43	0.69	0.44	0.90
	B*	43	0.67	0.43	0.70	0.42	0.88
	C	44	0.66	0.43	0.69	0.43	0.89
ELA Grade 6	A	48	0.65	0.42	0.68	0.41	0.89
	B	48	0.65	0.43	0.69	0.43	0.91
	C	48	0.65	0.43	0.69	0.41	0.89
ELA Grade 7	A	48	0.64	0.42	0.69	0.41	0.89
	B	48	0.64	0.43	0.69	0.42	0.90
	C	48	0.64	0.42	0.68	0.41	0.89
ELA Grade 8	A*	47	0.61	0.40	0.65	0.40	0.88
	B	48	0.59	0.40	0.64	0.39	0.88
	C	48	0.59	0.40	0.64	0.39	0.88

Table 7.4 IRT Average Descriptive Statistics for ELA EOG 2012–2013

EOG		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	1.82	-0.41	0.21	1.704	-0.659	0.17
	B	44	1.852	-0.372	0.21	1.733	-0.76	0.16
	C	44	1.838	-0.38	0.21	1.685	-0.776	0.16
Grade 4	A	44	1.534	-0.335	0.23	1.413	-0.683	0.16
	B	44	1.621	-0.293	0.22	1.531	-0.675	0.17
	C*	43	1.519	-0.373	0.21	1.39	-0.796	0.18
Grade 5	A	44	1.621	-0.277	0.21	1.572	-0.511	0.16
	B*	43	1.618	-0.302	0.23	1.482	-0.522	0.21
	C	44	1.723	-0.287	0.22	1.557	-0.566	0.17
Grade 6	A	48	1.64	-0.192	0.23	1.432	-0.489	0.18
	B	48	1.72	-0.173	0.24	1.532	-0.48	0.20
	C	48	1.624	-0.17	0.22	1.373	-0.573	0.17
Grade 7	A	48	1.56	-0.096	0.21	1.34	-0.573	0.16
	B	48	1.844	-0.052	0.24	1.496	-0.472	0.18
	C	48	1.739	-0.08	0.23	1.465	-0.428	0.18
Grade 8	A*	47	1.417	0.044	0.21	1.249	-0.411	0.16
	B	48	1.587	0.108	0.21	1.3	-0.345	0.15
	C	48	1.623	0.075	0.22	1.386	-0.266	0.19

Table 7.5 CTT Average Descriptive Statistics for EOC English II 2012–2013

EOC		Number of Items	Average CTT Field Test Administration		Average CTT Operational Administration		
			Pvalue	Biserial Correlation	Pvalue	Biserial Correlation	Reliability (Cronbach Alpha)
English II	A	53	0.48	0.41	0.65	0.39	0.89
	B*	52	0.47	0.40	0.63	0.39	0.89
	C	53	0.47	0.37	0.65	0.39	0.89
	M	53	0.47	0.42	0.64	0.39	0.89
	N*	52	0.48	0.41	0.60	0.39	0.89
	O	53	0.46	0.38	0.62	0.39	0.89

Table 7.6 IRT Average Descriptive Statistics for EOC English II 2012–2013

EOC		Number of Items	Average IRT Field Test Administration			Average IRT Operational Administration		
			Slope (a)	Threshold (b)	Asymptote (g)	Slope (a)	Threshold (b)	Asymptote (g)
English II	**A**	53	1.856	0.545	0.21	1.358	-0.207	0.18
	B*	52	1.704	0.609	0.22	1.364	-0.329	0.17
	C	53	1.829	0.671	0.22	1.299	-0.125	0.18
	M	53	1.898	0.576	0.20	1.355	-0.198	0.17
	N*	52	1.748	0.588	0.21	1.376	0.109	0.18
	O	53	1.797	0.663	0.21	1.3	-0.085	0.17

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

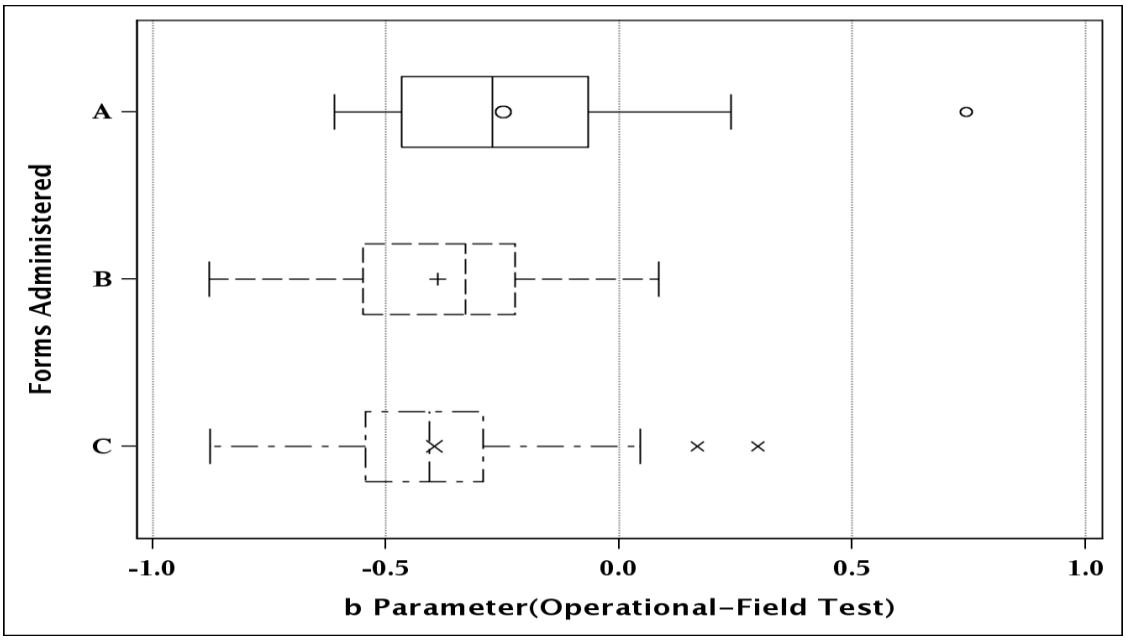


Figure 7.16 Grade 4 ELA b-parameter Difference Operational and Field Test

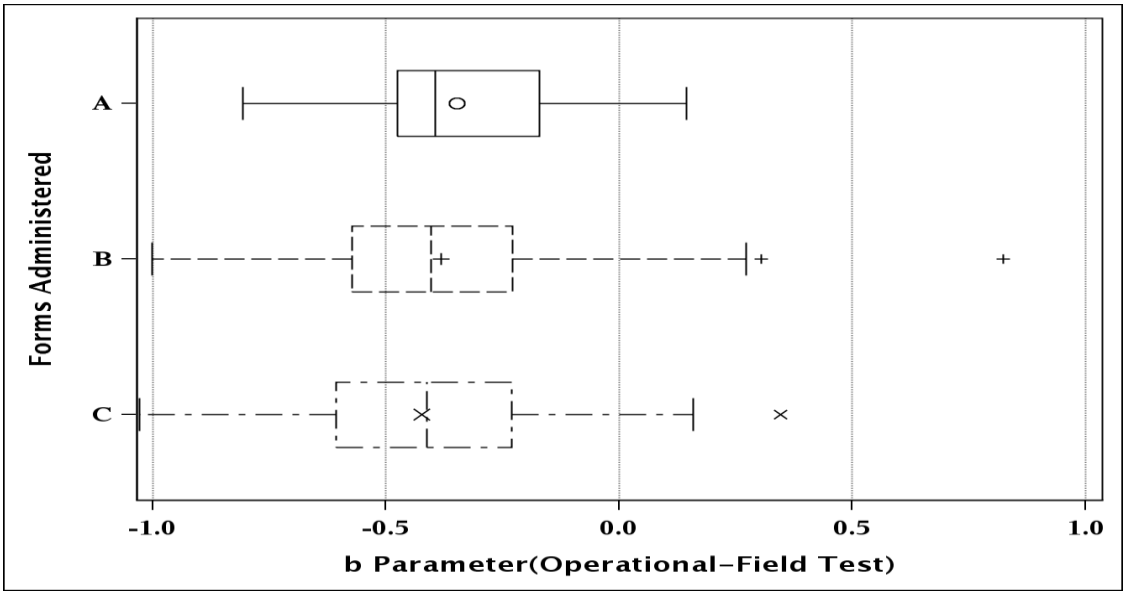


Figure 7.17 Grade 5 ELA b-parameter Difference Operational and Field Test

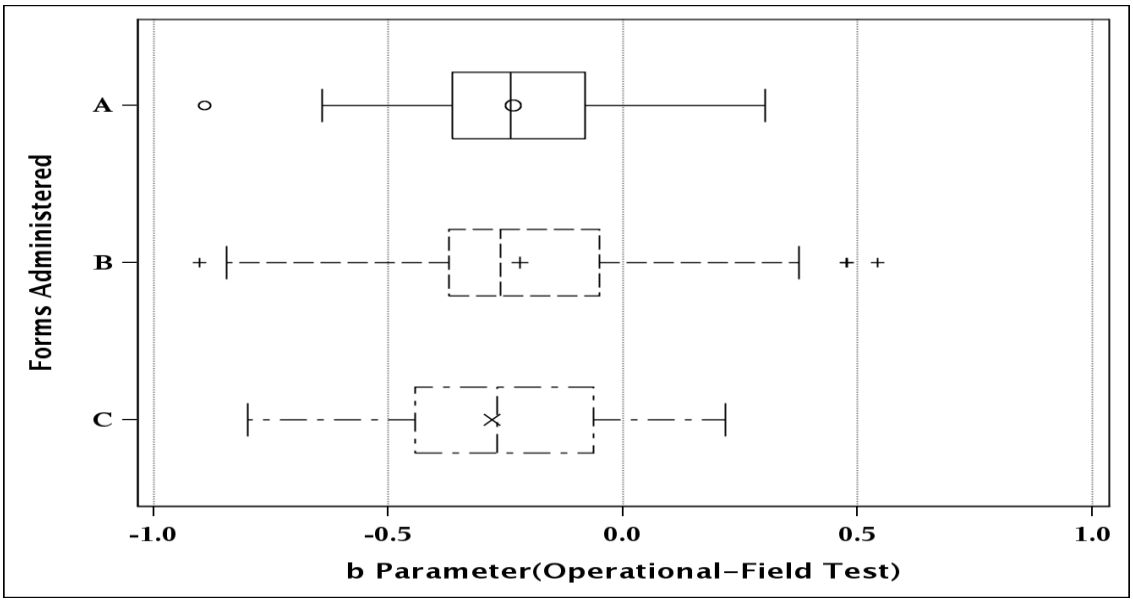


Figure 7.18 Grade 6 ELA b-parameter Difference Operational and Field Test

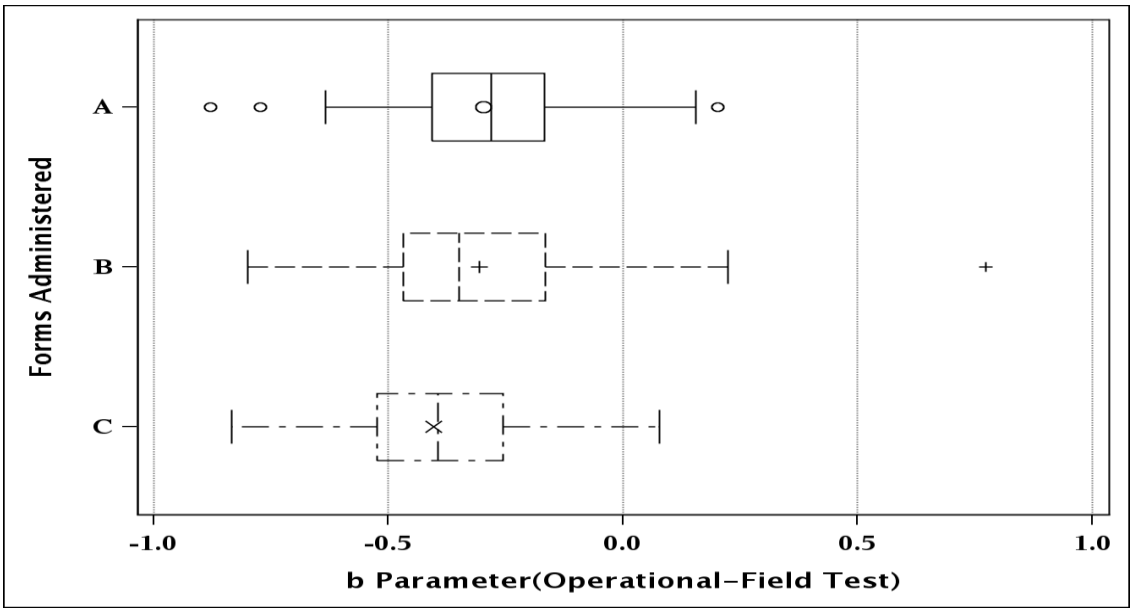


Figure 7.19 Grade 7 ELA b-parameter Difference Operational and Field Test

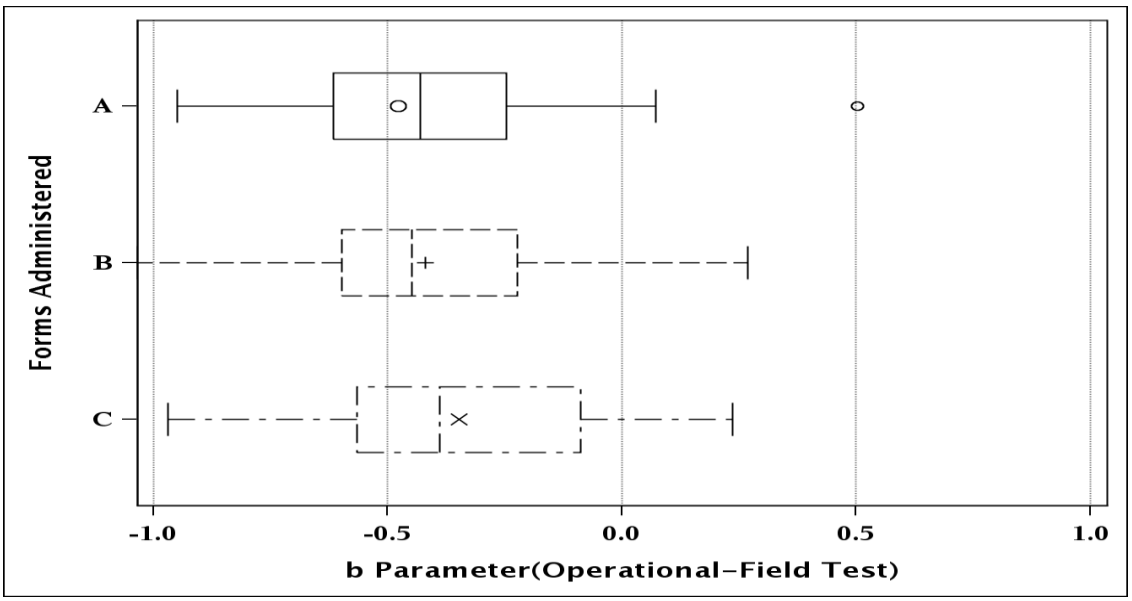


Figure 7.20 Grade 8 ELA b-parameter Difference Operational and Field Test

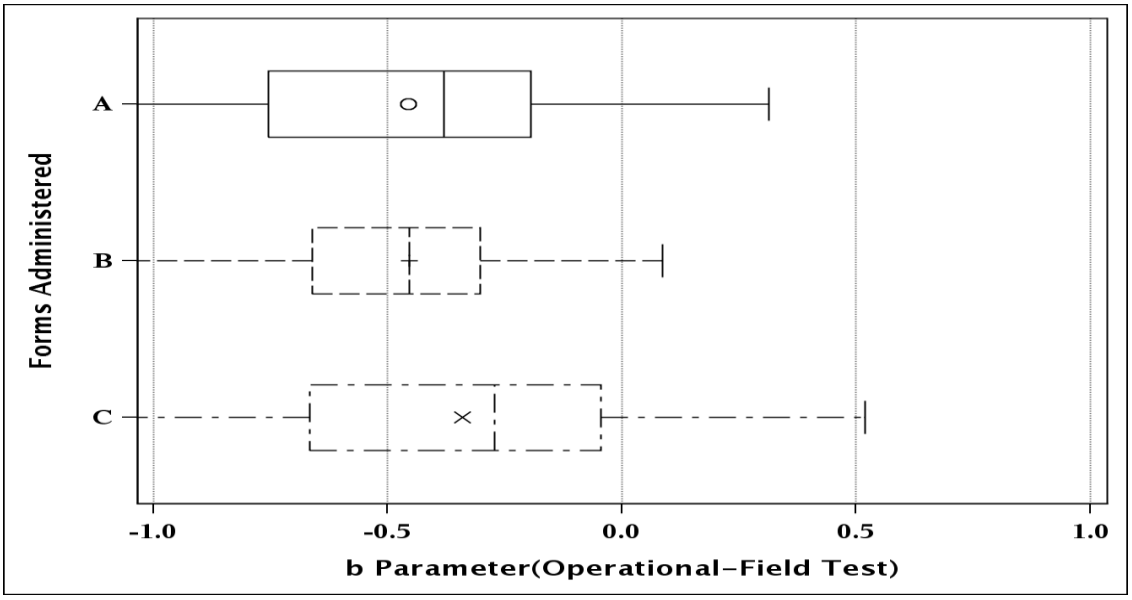
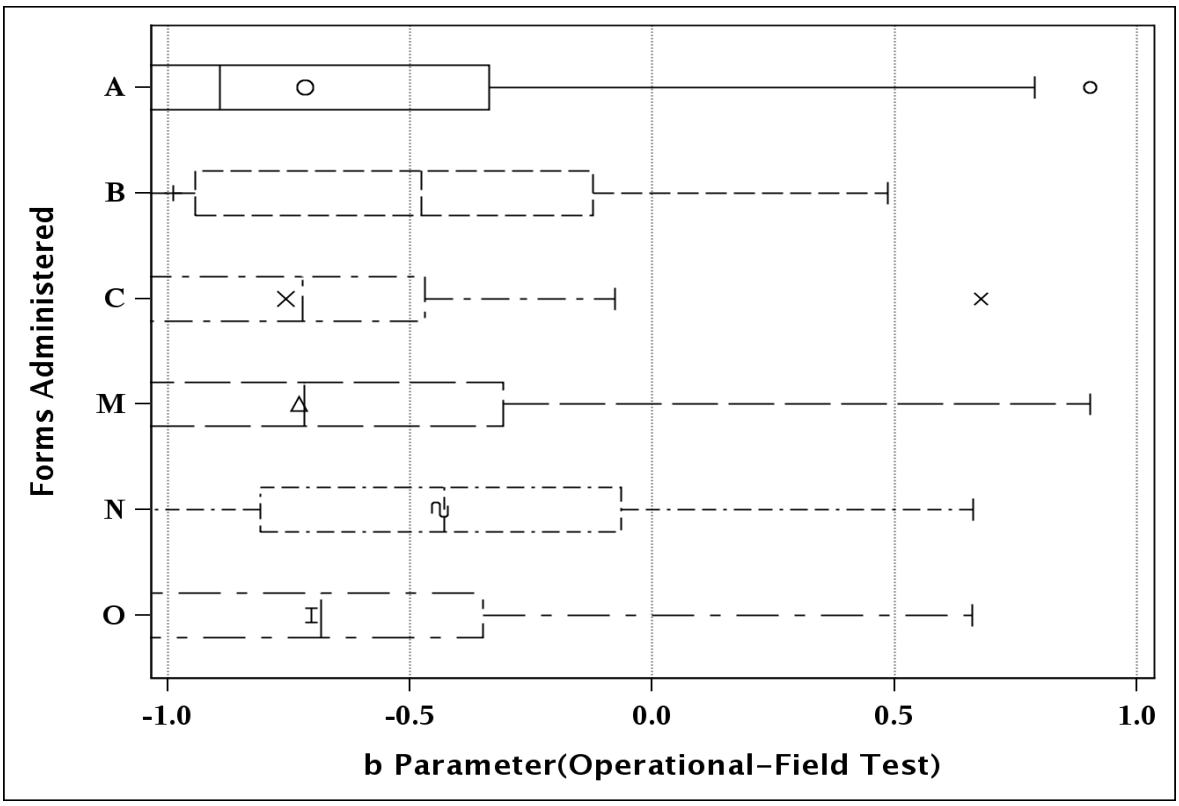


Figure 7.21 English II b-parameter Difference Operational and Field Test



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 7.7 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 0.17 to 1.12 and b-parameter range from -0.23 to as large as -0.94 indicating on average a medium-to-large effect from field test to operational parameters.

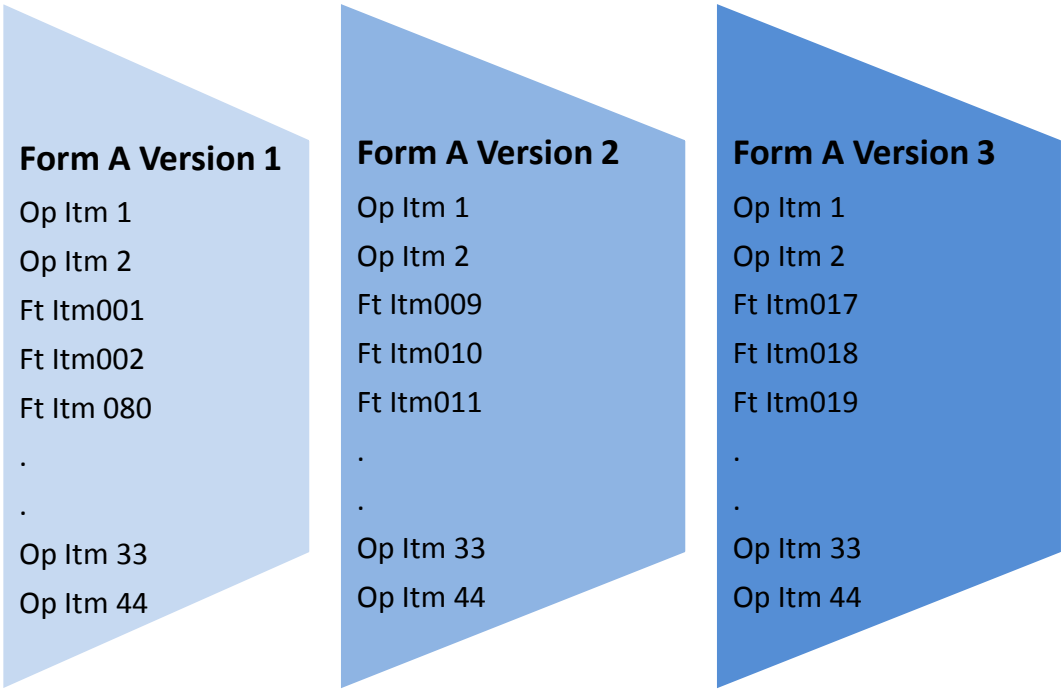
Table 7.7 ELA Effect Size Summary of Operational and Field Test Statistics

Grade		Operational Items	P-value Standardized Mean Difference	Threshold Standardized Mean Difference
ELA Grade 3	**A**	44	0.19	-0.33
	B	44	0.37	-0.49
	C	44	0.39	-0.52
ELA Grade 4	**A**	44	0.22	-0.39
	B	44	0.31	-0.42
	C*	43	0.40	-0.53
ELA Grade 5	**A**	44	0.14	-0.26
	B*	43	0.20	-0.23
	C	44	0.17	-0.30
ELA Grade 6	**A**	48	0.20	-0.35
	B	48	0.26	-0.36
	C	48	0.27	-0.45
ELA Grade 7	**A**	48	0.35	-0.53
	B	48	0.30	-0.43
	C	48	0.25	-0.36
ELA Grade 8	**A***	47	0.36	-0.57
	B	48	0.32	-0.48
	C	48	0.32	-0.37
English II	**A**	53	0.97	-0.84
	B*	52	0.89	-0.52
	C	53	1.12	-0.94
	M	53	0.92	-0.87
	N*	52	0.74	-0.57
	O	53	0.92	-0.87

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 ELA, a total of 8 field test items are embedded within each operational version of the EOG assessment. English II has 15 field test items embedded within the operational form. 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 forms 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.

Chapter 8 Standard Setting

Standard setting is a process used to set 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 NCSCS 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 ELA, 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." 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.

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 recommending cut scores and achievement levels for the newly developed ELA, EOG, and EOC assessments. Three panels (grades 3–5, grades 6–8, and English II) with a total of 54 North Carolina ELA/Reading educators (18 for grade 3-5, 19 for grades 6-8, and 17 for English II) convened in Chapel Hill, North Carolina, between July 22 and July 26, to make cut score recommendations for the ELA/Reading 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 4 achievement levels were present to the NCSBE for adoption. An abbreviated version of the final standard setting study prepared by Pearson[k] 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 54 panelist ranged from less than 5 years to more than 21 years of experience. The shows that a very diverse group educators participated in the standard setting.

Table 8.1 Panelist Experience as Educators

Panel	N	Years in Current Position					
		1-5	6-10	11-15	16-20	21+	NR
Reading 3–5	18	1	3	5	1	8	0
Reading 6–8	19	2	2	6	6	3	0
English II	17	3	5	5	2	1	1

Note: NR = no response.

The panelists' professional backgrounds are summarized in *Table 8.2* and *Table 8.3*. Panelist in Reading 3–5 and 6–8 groups made cut score recommendations for three grade levels of EOG and ELA, and the 17 panelists in the English II group made cut score recommendations for EOC English II. From these tables, teachers reported as teaching in lower, middle, or upper grades are reported in the context of their committee. For example, a lower-grade panelist in the Reading 3–5 panel teaches Grade 3 Reading, while a lower-grade panelist in the Reading 6–8 panel teaches Grade 6 Reading. Panelists who reported teaching more than one grade level within the subject area are listed under the multiple grades column, and panelists who primarily teach a grade level outside of the panel's range (e.g., a grade 2 teacher who participated in the

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

Reading 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, and a variety of professional backgrounds were 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
Reading 3–5	3	1	4	3	1	0	4	0	1	1
Reading 6–8	4	5	3	2	0	3	0	0	0	2

Note: LOW = lower grade, MID = middle grade, UP = upper grade, MUL = multiple grades, OFF = off-grade, SED = special education, SPE = specialist, COA = coach, GNS = grade level not specified, OTH = other.

Table 8.3 Panelist Professional Background: Single-Grade Panels

Panel	ON	OFF	SED	SPE	COA	HED	OTH	RET	NR
English II	11	2	1	0	0	2	1	0	0

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

In addition to reporting their own demographic characteristics (*Table 8.4*), 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 for 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
Reading 3–5	17	1	0	7	1	1	1	6	2	0
Reading 6–8	18	1	0	4	0	0	1	14	0	0
English II	14	3	0	1	0	2	0	14	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
Reading 3–5	2	1	1	0	4	3	4	2	0	1
Reading 6–8	0	1	1	4	2	5	5	0	1	0
English II	4	0	1	3	4	2	2	1	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	W
Reading 3–5	1	7	3	7	1	9	3	5
Reading 6–8	0	6	8	5	1	11	5	2
English II	1	6	5	5	4	1	11	2

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

8.1.2 Vertical Articulation Committee

Each standard setting breakout session room, which contained between 16 and 20 total panelists, was arranged to include three tables. At various points throughout the process, panelists within a committee broke up and worked together in groups of between 5 and 7 individuals at each table. Each of the three tables had at least one designated table leader, who was selected by 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 Reading (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 rooms.

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 provided 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 areas and grades. Panelists began by drafting and discussing "just barely" level descriptors: statements describing performance expectations for students who are *just barely* at the three cut points separating the four achievement levels. The "just barely" level descriptors are critical to standard setting for two reasons. First, discussing characteristics of students who are just barely at a particular cut point dividing two adjacent achievement levels aids panelists in developing a strong understanding of the differences in observed student performance across achievement levels. Second, in subsequent steps occurring during the standard setting process, panelists referred to the "just barely" level 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 tests with reading selections only), and linked content standard. After completing the OIB review, panelists were given an opportunity to

share their thoughts on and reactions to the test's content with their colleagues in the breakout session.

8.1.8 Standard Setting Training and Practice Round

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

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

which panelists affirm that they understand the process and feel prepared to begin (see Appendix F). Panelists were encouraged to seek clarification from the breakout session facilitator on any remaining questions or concerns, should they have any, prior to beginning the first round of judgment. Upon unanimous positive affirmation of readiness to proceed, committees began the standard setting process. The standard setting process consisted of three rounds of judgment. Panelists completed readiness surveys affirming their understanding of the process and willingness to proceed prior to beginning each of the three rounds. The committees were instructed to set their cuts in order starting at Level 2, then at Level 3, and finally at Level 4.

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

Behavioral Descriptors

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

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

Round 1 Feedback and Discussion and Round 2 Standard Setting

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

- How similar their cut scores were to that of the rest of the table (i.e., is a given panelist more lenient or stringent than the other panelists?)
- If a panelist had cut scores dissimilar to the table'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 these bookmarks. Using data provided in the feedback handouts, panelists discussed their judgments related to items in the range between the highest and lowest bookmarks for each achievement level. An example of the rating agreement feedback data provided to each table of panelists is provided in *Table 8.7*.

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**
	Minimum	**14**	**42**	**60**
	Maximum	**43**	**73**	**82**

In addition to the Round 1 cut score agreement data, panelists were shown external data to further inform their judgments in subsequent rounds of judgment. Panelists were provided with empirical item difficulty data showing the proportion of all test-takers from the spring 2013 administration who correctly answered each item (i.e., item *p*-values). The breakout session facilitator also shared with panelists the ACT Explore® cut score, which was linked to the North Carolina assessment by 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®
Reading 3	9	39	73	66
Reading 4	9	35	61	58
Reading 5	5	29	59	55
Reading 6	6	30	64	63
Reading 7	6	33	61	58
Reading 8	4	27	57	57
English II	3	25	61	*

*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 the table level as well as across the committee. Additionally, panelists were shown a graphical display of student impact data. The impact data displayed

the percentages of spring 2013 test takers who would be classified into the four achievement levels based on the panel's median cut score recommendation. Impact was shown for the overall North Carolina test-taking population, and impact was also broken down into gender and ethnicity subgroups. Panelists were given an opportunity to discuss the appropriateness of their cut scores given the current impact data. Following discussion of the Round 2 feedback data, panelists completed readiness surveys and proceeded to the third and final round of judgment.

Round 3 Feedback and Discussion

Following Round 3, panelists were shown their final recommended cut scores, which were based on the committee's median cut score judgments from this final round of judgment. Panelists were shown impact data, which again included overall impact as well as impact broken down 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 Reading) and also across subjects. Panelists were asked to evaluate and discuss, from a policy perspective, the reasonableness of the committees' content-oriented cut score recommendations and the impact of imposing these achievement expectations on student test scores. Panelists were guided through a process whereby they evaluated the reasonableness of impact for particular grades/subjects, both in isolation and in

contrast to other grades and subject areas. Table leaders from each committee were present in the vertical articulation meeting, which allowed them an opportunity to share with the entire group their reflections on the execution of the standard setting procedure as well as the discussions that occurred within their committees.

Following group discussions of the cuts and impact data, the lead facilitator asked the vertical articulation committee if they felt any cut score changes may be appropriate, given the observed patterns of impact data. The lead facilitator projected a spreadsheet with cut scores and impact data, and panelists were permitted to suggest potential revised cut scores to see real-time changes to impact data based on these potential revisions. Following NCDPI's instructions, the lead facilitator did not limit the range of potential cut score changes available to the vertical articulation committee. The lead facilitator 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:

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

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

Each participant on the vertical articulation panel considered the original recommended cut scores and their impact data as well as other potential cut scores and the changes in impact data associated with these potential cuts. Each member of the vertical articulation committee provided a unique, independent recommendation to either keep or change the cut scores. Consistent with the previous phase of the standard setting meeting, members of the vertical articulation committee completed readiness surveys and unanimously affirmed their understanding of the process and willingness to proceed prior to rendering their final

recommendations. The lead facilitator impressed upon the vertical articulation panel that their holistic, policy-oriented cut score recommendations would supplement, not overwrite, the content-oriented cut recommendations provided by the standard setting panels and would provide the North Carolina State Board of Education with additional information to consider when deciding which cut scores to adopt. Each member of the vertical articulation committee provided an independent recommendation to either keep or adjust the cut scores for every grade and subject. Panelists recorded their judgments on provided forms (see full report Appendix M) and returned them to the lead facilitator forprocessing. After completing the vertical articulation process for all grades and subjects, panelists completed an evaluation survey of the vertical articulation process.

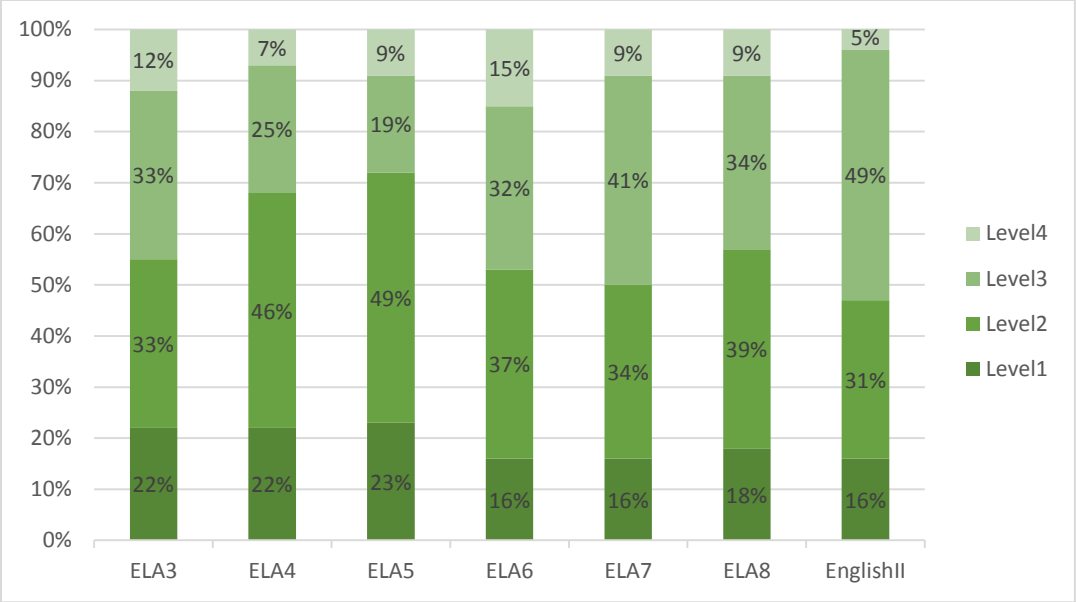
8.3 Results

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, which will be documented separately from this standard setting technical report. *Figure 8.1* displays impact data for the EOG Reading and End-of-Course English II assessments, respectively, based upon these cut score recommendations. Tables and figures showing individual panelists' page cuts across rounds are provided in the full report.

Table 8.10 Pre-Vertical Articulation Page Cuts

Assessment	Level 2	Level 3	Level 4
Reading 3	26	55	74
Reading 4	25	58	75
Reading 5	23	55	71
Reading 6	15	46	69
Reading 7	15	45	70
Reading 8	16	42	70
English II	9	34	79

Figure 8.1 Pre-Vertical Articulation Impact Data

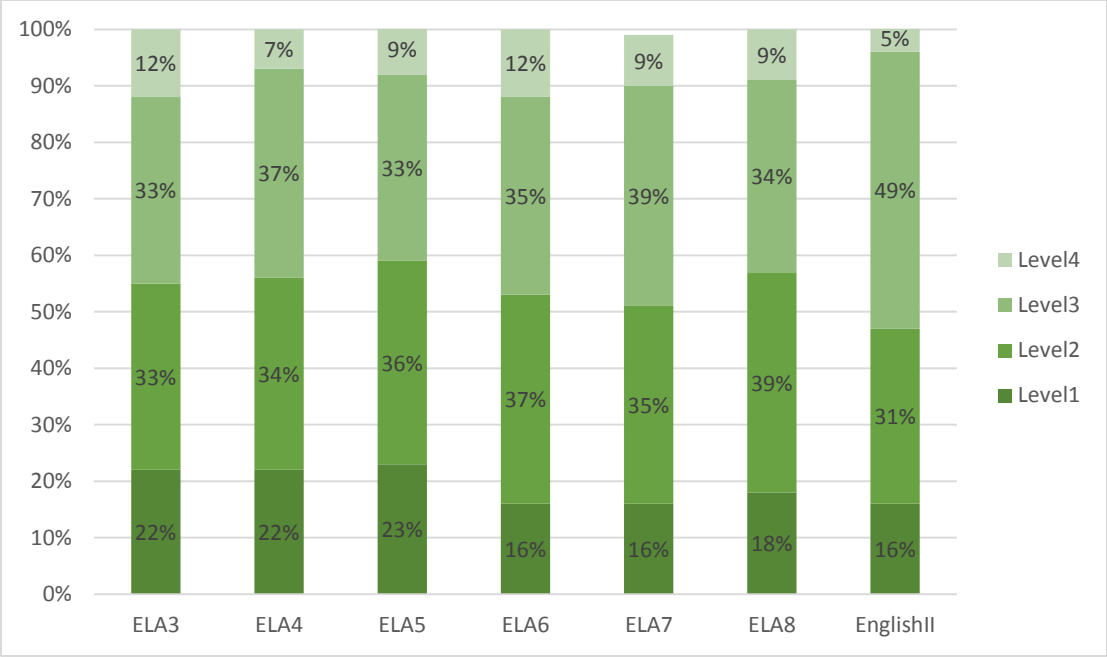


Cut scores obtained following the vertical articulation session are shown in *Table 8.11*, and impact data associated with these recommended cut scores are displayed in the subsequent figures.

Table 8.11 Post-Vertical Articulation Page Cuts

Assessment	**Level 2**	**Level 3**	**Level 4**
Reading 3	26	55	74
Reading 4	25	50	75
Reading 5	23	46	71
Reading 6	15	46	73
Reading 7	15	47	70
Reading 8	16	42	70
English II	9	36	79

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 Reading 3 "Level 2" cut of 432 is interpreted as students with a scale score of 431 or lower are placed in "Achievement Level 1," and student who score between 432 and 453 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**
Reading 3	432	442	452
Reading 4	439	448	460
Reading 5	443	453	464
Reading 6	442	454	465
Reading 7	445	457	469
Reading 8	449	462	473
English II	141	151	165

8.4 Validity of the Standard Setting

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

8.5 Standards Adoption and Revision

In October 2013, the North Carolina State Board of Education (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* and *Table 8.14*.

Table 8.13: Revised 5 Achievement Levels Descriptors

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

Table 8.14 Scale Scores Cuts Based on Five Achievement Levels 2014 and Beyond

Achievement Levels Cuts	Level 2	Level 3	Level 4	Level 5
ELA	Partial Command	Sufficient Command	Solid Command	Superior Command
EOG 3	432	439	442	452
EOG 4	439	445	448	460
EOG 5	443	450	453	464
EOG 6	442	451	454	465
EOG 7	445	454	457	469
EOG 8	449	458	462	473
English II	141	148	151	165

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. The new Level 3 minimum scale score was created by subtracting one conditional standard error of measurement (CSEM) from the original Level 3 scale score. Level 1 "Limited Command" and Level 2 "Partial Command" remained unchanged and describes students who have neither met on-grade-level proficiency standard nor college- and career-readiness standards.

Chapter 9 Test Results and Reports

This chapter is divided into two main sections and presents test-level summary statistics for ELA EOG and EOC based on reported scale scores and achievement levels from 2012–13 and 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 grade levels are slightly negatively skewed. The score distribution also shows a slight rightward shift for EOG grade 3 through EOG 8 as a result of the developmental scale that was implemented during scaling. Overall variability across all grades is consistent around 10.

Figure 9.1 English Grade 3 Scale Score Distribution 2012–13

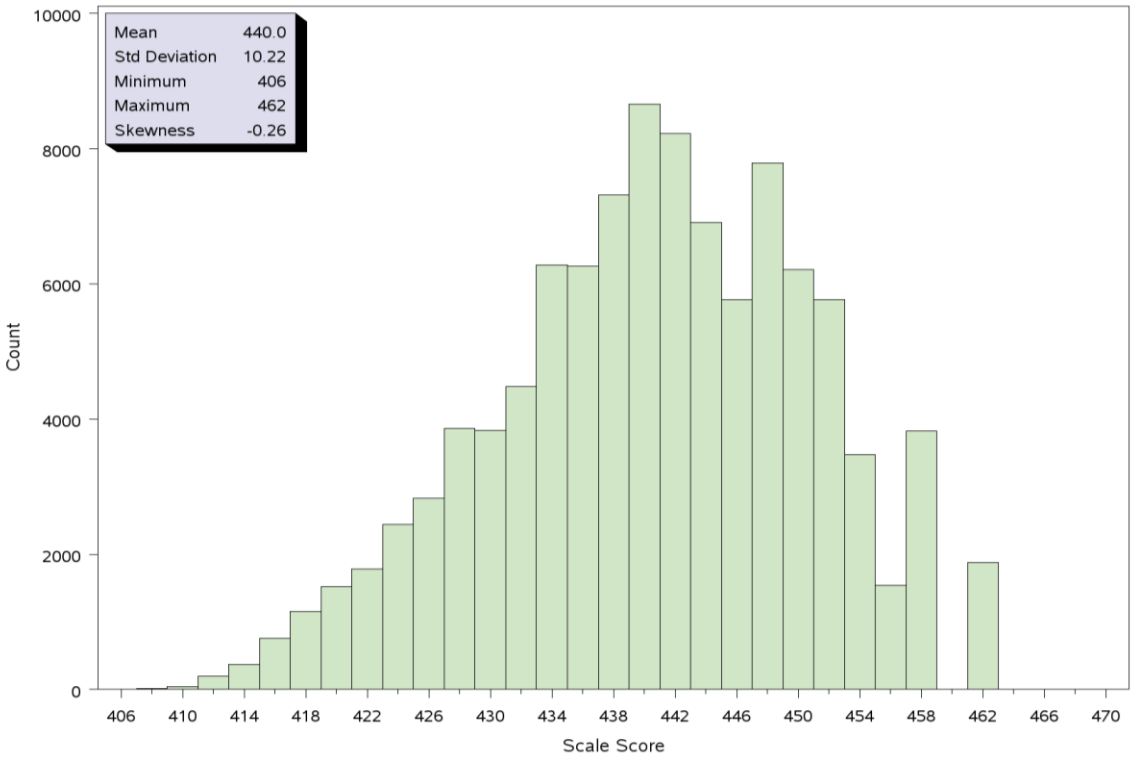


Figure 9.2 English Grade 4 Scale Score Distribution 2012–13

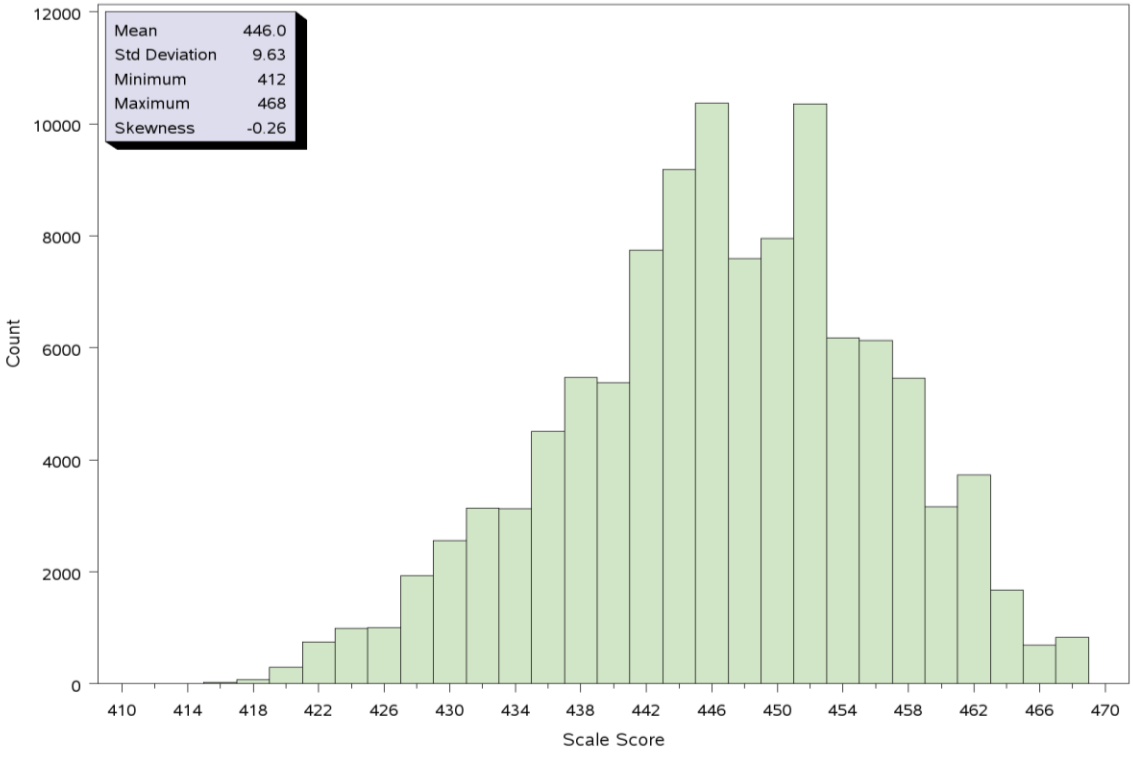


Figure 9.3 English Grade 5 Scale Score Distribution 2012–13

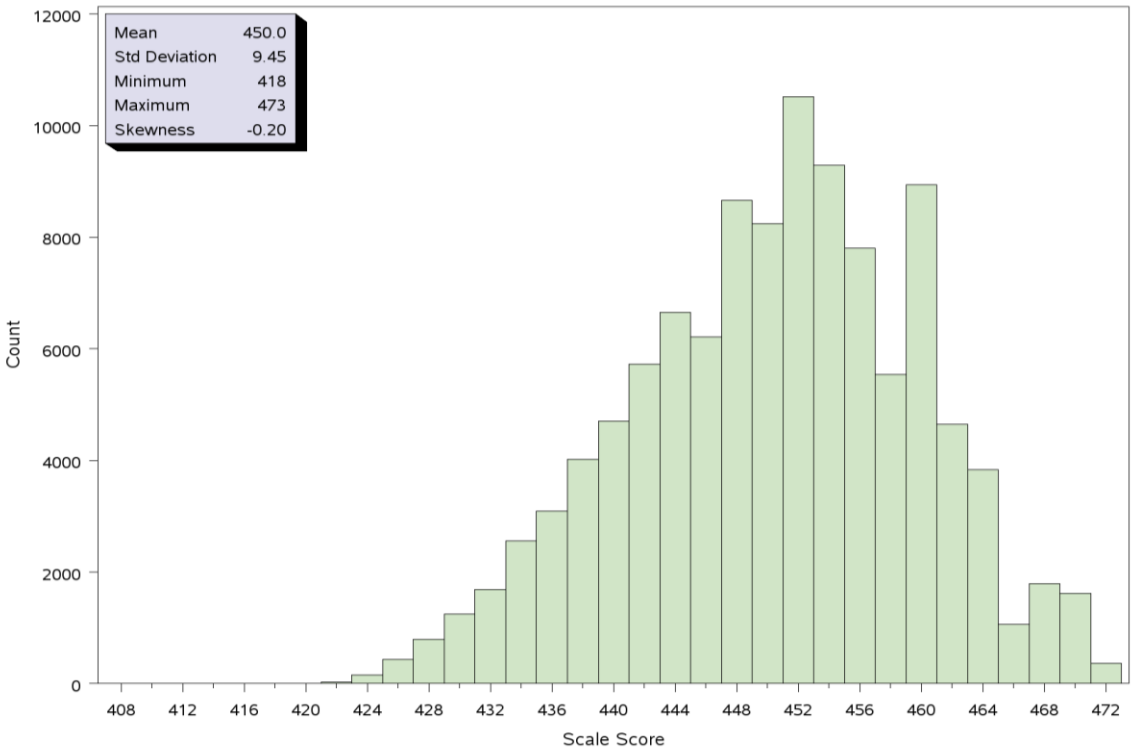


Figure 9.4 English Grade 6 Scale Score Distribution 2012–13

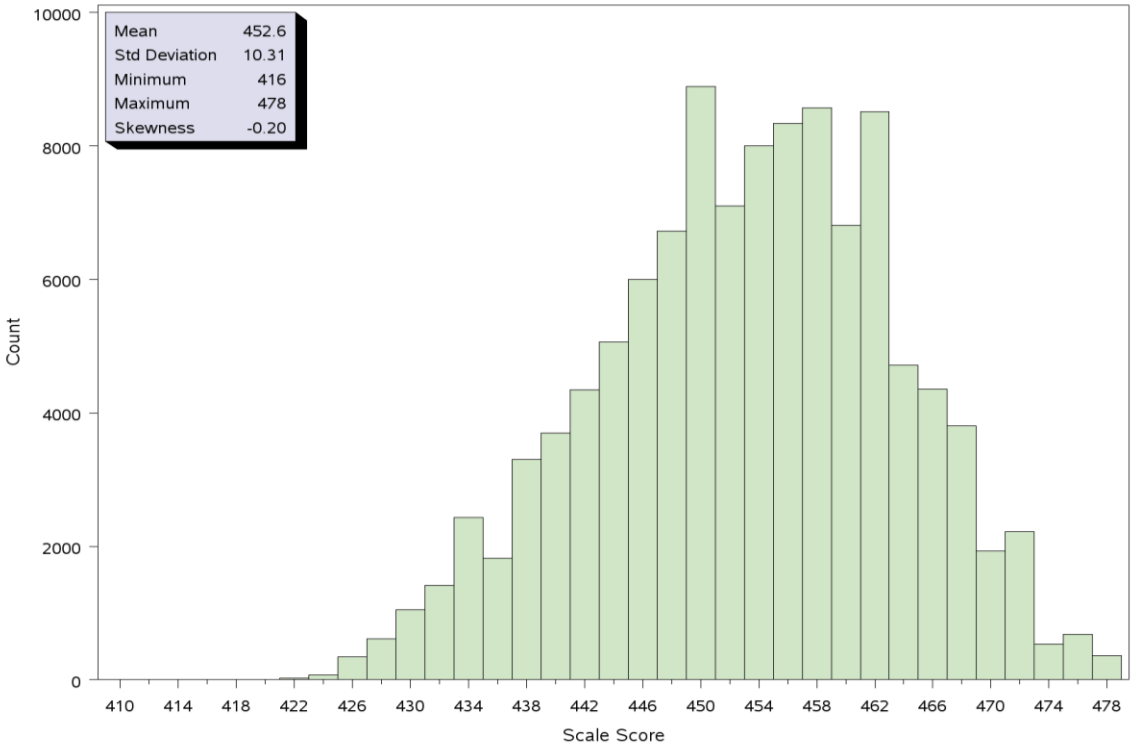


Figure 9.5 English Grade 7 Scale Score Distribution 2012–13

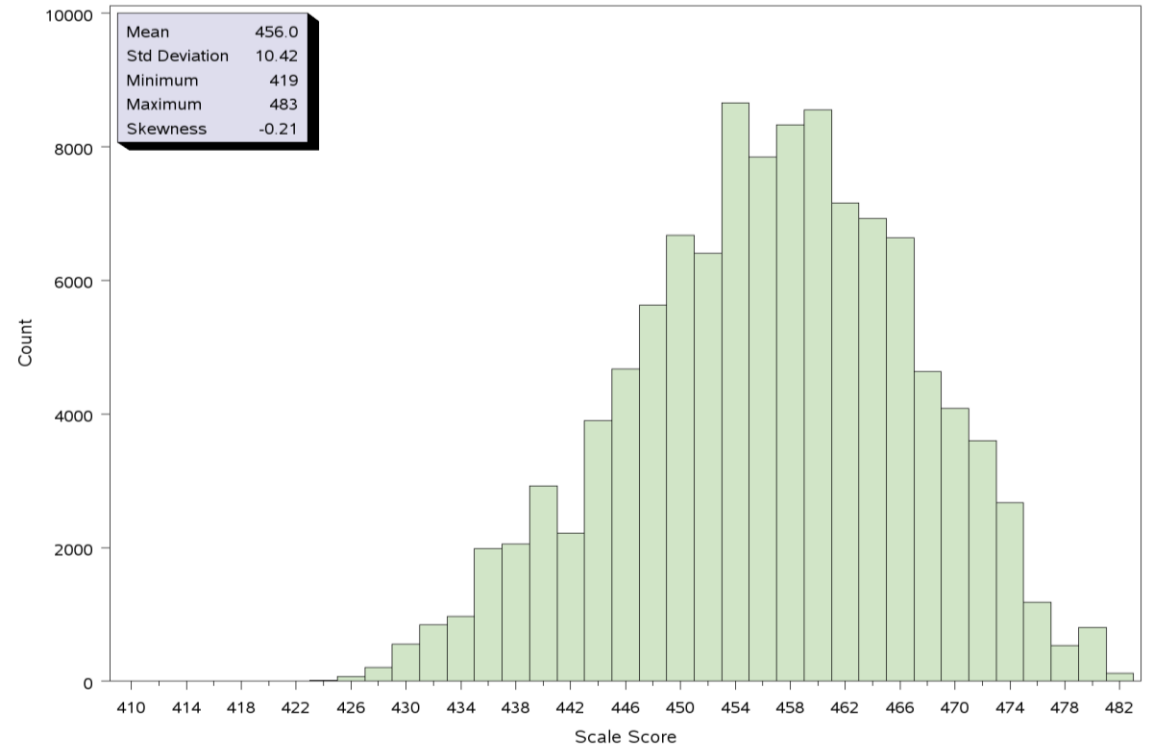


Figure 9.6 English Grade 8 Scale Score Distribution 2012–13

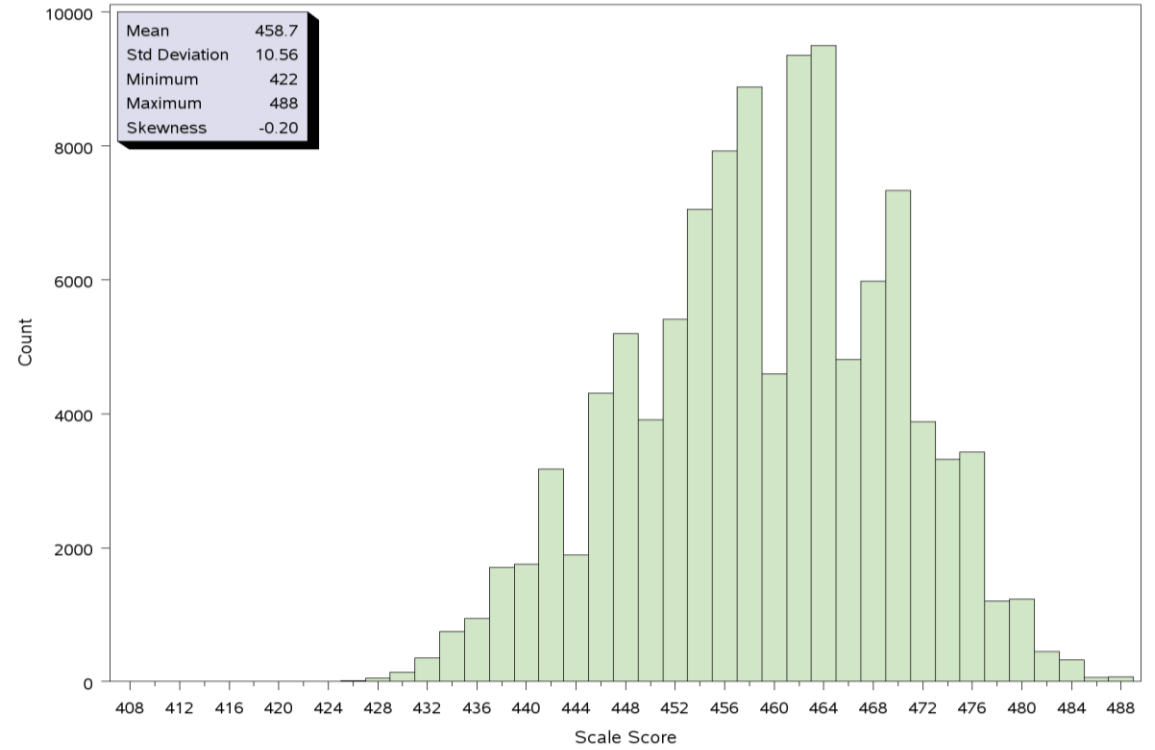
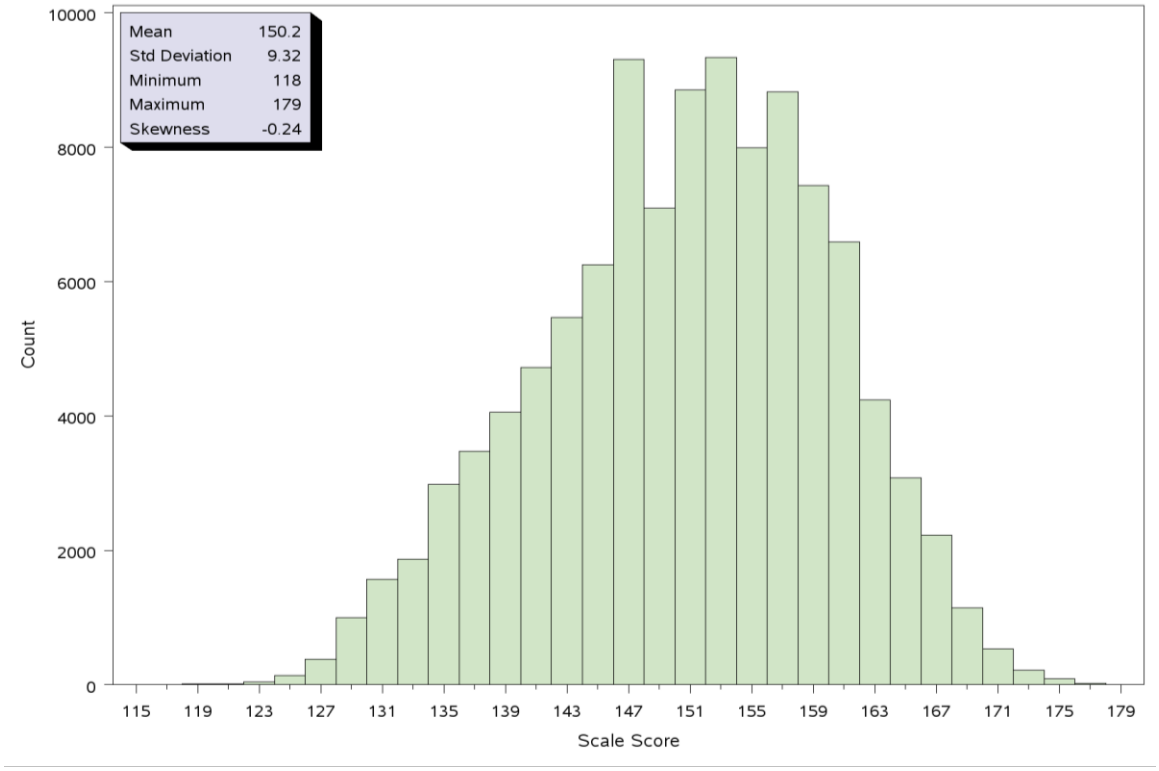


Figure 9.7 English II Scale Score Distribution 2012–13



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 with the exception of EOG grade 5. Descriptive summary evidence from *Table 9.1* indicates average scale scores have been consistent across the past three years. In general, average scales scores across all assessments for the past three years have either stayed flat or are slightly trending downwards. But the effect of the difference across years is very small and can be mostly explained 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 slight upward trend in overall variability across years 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	103,048	440.0	10.2	111,182	440.5	10.3	116,376	439.6	10.9
EOG 4	110,147	446.0	9.6	103,553	445.7	10.1	113,959	445.8	10.2
EOG 5	109,702	450.0	9.4	111,175	450.0	9.6	106,589	449.5	10.3
EOG 6	111,575	452.7	10.3	110,955	452.6	10.6	114,459	452.0	11.2
EOG 7	110,784	456.0	10.4	113,012	455.8	10.7	114,661	454.8	11.4
EOG 8	108,855	458.7	10.6	111,946	458.9	10.7	116,751	458.1	11.2
EOC English II	105,779	150.5	9.2	109,569	150.5	9.5	114,680	149.8	9.9

9.1.2 Scale Score by Gender

Scale score summaries by gender for EOG and EOC across three administration cycles show similar trends 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 about 1 to 3 scale points higher than males. The average difference between females and males seems to be larger: about .33 standard deviation in grades 7, 8 and high school see *Table 9.2*. Scale score variance was very similar in both gender groups and followed a similar pattern, with a slightly increasing trend of 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	50,888	440.8	10.0	55,082	441.3	10.0	56,926	440.6	10.4
	Male	52,160	439.3	10.4	56,100	439.7	10.5	59,450	438.6	11.2
EOG 4	Female	54,630	446.5	9.5	50,912	446.5	9.8	55,848	446.8	9.9
	Male	55,517	445.6	9.8	52,641	444.9	10.2	58,111	444.9	10.5
EOG 5	Female	54,482	450.5	9.3	54,950	450.6	9.4	51,929	450.3	10.0
	Male	55,220	449.5	9.6	56,225	449.5	9.8	54,660	448.7	10.5
EOG 6	Female	55,292	453.4	10.0	54,630	453.3	10.3	55,825	452.9	10.8
	Male	56,283	451.9	10.5	56,325	451.9	10.8	58,634	451.1	11.5
EOG 7	Female	55,006	456.7	10.0	55,820	456.5	10.4	55,939	455.8	10.9
	Male	55,778	455.3	10.7	57,192	455.1	11.0	58,722	453.9	11.7
EOG 8	Female	54,279	459.8	10.1	55,395	460.2	10.2	57,159	459.7	10.7
	Male	54,576	457.5	10.8	56,551	457.6	10.9	59,592	456.6	11.5
English II	Female	52,422	151.7	8.9	53,936	151.9	9.1	56,272	151.4	9.5
	Male	53,357	149.2	9.4	55,633	149.2	9.6	58,408	148.3	10.1

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 the 2013–14 administration and beyond; and 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 level 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 short-term trend between 2013–14 and 2014–15 on average, shows a 2% decline in the proportion of students classified as college- and career-ready (Levels 4 and 5) for EOG grades 3, 6, 7, 8 and EOC English II. For EOG grades 4 and 5, the proportion has actually increased by 1.4% and 0.7% respectively.

The achievement-level classifications by gender across grades and administrations are shown in *Table 9.4* and *Table 9.5*. These tables should be interpreted with similar precaution as the previous table with regards to achievement levels for 2012–13. A similar trend as the total population can be observed between genders. The results across all administrations and grades further indicated that there are higher proportions of female students over male students who scored at level 4 or above (college- and career-readiness). Overall about 5% more female students are classified as college-and-career ready compared to their male counterpart. The range of the difference is 2.8% to as high as 9.6% in high School. The differences were more pronounced in EOC English II, ranging from 10.7% in the 2012–13 administration to 12.2% in the 2014–15 administration.

Table 9.3 Achievement level classifications by Grade and Year

			% Achievement Level				
	Year	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*	103,048	20.3	33.1		34.6	12.1
	2013–14	111,182	19.1	19.1	12.8	36.7	12.3
	2014–15	116,376	22.2	18.8	12.6	34.9	11.6
EOG 4	2012–13*	110,147	21.6	32.9		37.9	7.6
	2013–14	103,553	24.3	18.5	11.3	38.8	7.0
	2014–15	113,959	23.3	17.9	11.6	39.9	7.3
EOG 5	2012–13*	109,702	22.2	36.7		33.0	8.1
	2013–14	111,175	22.4	22.3	13.8	32.7	8.7
	2014–15	106,589	25.1	22.1	10.8	33.6	8.5
EOG 6	2012–13*	111,575	15.1	36.4		36.0	12.4
	2013–14	110,955	16.1	25.2	11.4	34.9	12.4
	2014–15	114,459	19.2	23.6	10.6	34.0	12.7
EOG 7	2012–13*	110,784	14.2	36.0		38.1	11.8
	2013–14	113,012	15.0	25.7	10.0	37.3	11.9
	2014–15	114,661	19.0	24.9	9.5	35.0	11.6
EOG 8	2012–13*	108,855	18.6	38.9		33.2	9.3
	2013–14	111,946	18.4	25.9	12.1	33.5	10.2
	2014–15	116,751	21.5	25.1	11.8	31.6	10.0
English II	2012–13*	105,779	15.6	31.7		47.4	5.3
	2013–14	109,569	16.6	20.5	9.6	47.0	6.3
	2014–15	114,680	19.4	20.5	9.6	44.9	5.8

**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				
		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	50,888	18.0	32.7		36.0	13.2
		Male	52,160	22.5	33.5		33.1	10.9
	2013–14	Female	55,082	16.6	18.7	13.1	38.3	13.3
		Male	56,100	21.6	19.5	12.6	35.1	11.2
	2014–15	Female	56,926	18.9	18.5	12.8	37.0	12.8
		Male	59,450	25.4	18.9	12.4	33.0	10.4
EOG 4	2012–13*	Female	54,630	20.0	33.0		38.8	8.2
		Male	55,517	23.2	32.9		37.0	7.0
	2013–14	Female	50,912	21.7	18.0	11.5	40.9	8.0
		Male	52,641	26.9	19.0	11.1	36.9	6.1
	2014–15	Female	55,848	20.2	17.4	11.8	42.4	8.2
		Male	58,111	26.4	18.3	11.5	37.6	6.3
EOG 5	2012–13*	Female	54,482	20.4	36.8		34.0	8.9
		Male	55,220	24.0	36.6		32.0	7.4
	2013–14	Female	54,950	20.5	22.7	13.9	33.4	9.5
		Male	56,225	24.3	21.9	13.7	32.1	8.0
	2014–15	Female	51,929	22.4	21.9	10.9	35.3	9.5
		Male	54,660	27.7	22.2	10.7	32.0	7.4
EOG 6	2012–13*	Female	55,292	12.7	36.5		37.3	13.5
		Male	56,283	17.4	36.4		34.8	11.4
	2013–14	Female	54,630	13.7	25.2	11.7	36.1	13.3
		Male	56,325	18.5	25.2	11.1	33.7	11.6
	2014–15	Female	55,825	16.0	23.8	11.0	35.3	13.9
		Male	58,634	22.2	23.4	10.1	32.8	11.5
EOG 7	2012–13*	Female	55,006	11.9	36.2		39.4	12.5
		Male	55,778	16.4	35.7		36.8	11.1
	2013–14	Female	55,820	12.6	25.8	10.2	38.8	12.6
		Male	57,192	17.4	25.7	9.7	36.0	11.3
	2014–15	Female	55,939	15.8	24.9	10.0	36.6	12.7
		Male	58,722	22.1	24.9	9.1	33.5	10.5
EOG 8	2012–13*	Female	54,279	14.8	39.1		35.5	10.6
		Male	54,576	22.3	38.8		31.0	8.0
	2013–14	Female	55,395	14.4	25.1	12.5	36.1	11.9
		Male	56,551	22.4	26.6	11.8	30.9	8.5
	2014–15	Female	57,159	16.8	24.5	12.2	34.6	11.9
		Male	59,592	26.1	25.6	11.4	28.7	8.2

Table 9.5 EOC English II 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***
English II	2012–13*	Female	52,422	11.8	30.1		51.5	6.6
		Male	53,357	19.2	33.4		43.4	4.0
	2013–14	Female	53,936	12.6	18.6	9.6	51.5	7.7
		Male	55,633	20.4	22.4	9.7	42.6	4.9
	2014–15	Female	56,272	14.5	19.1	9.6	49.6	7.3
		Male	58,408	24.1	21.8	9.5	40.3	4.4

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

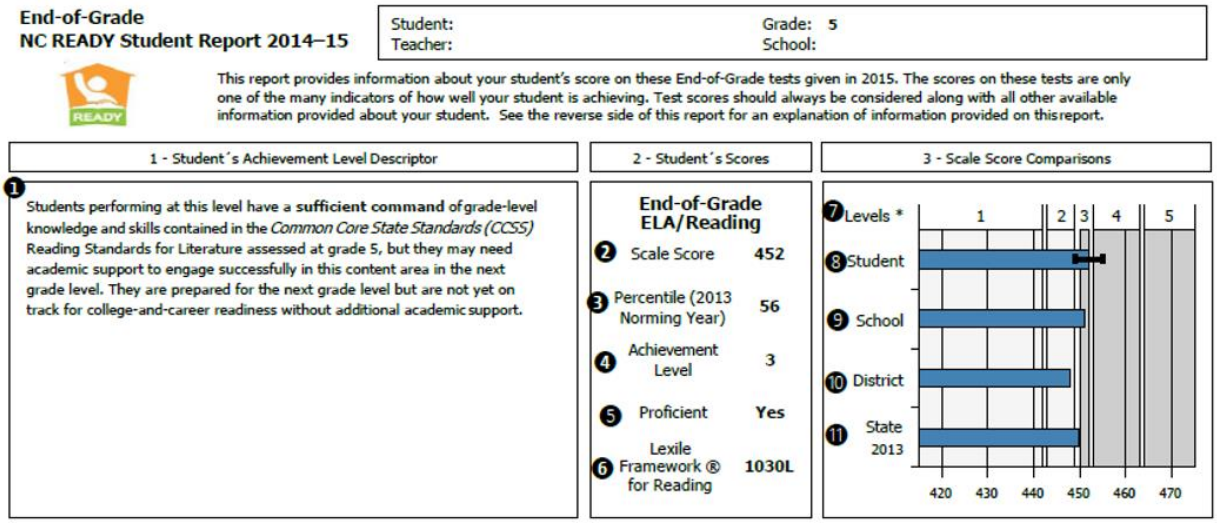
9.2 Sample Reports

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 of 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. A sample ISR report for Grade 5 ELA 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 Grade 5 EOG ELA/Reading 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 452 in Grade 5 EOG ELA and a percentile of 56 is said to have performed better than 55% 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.

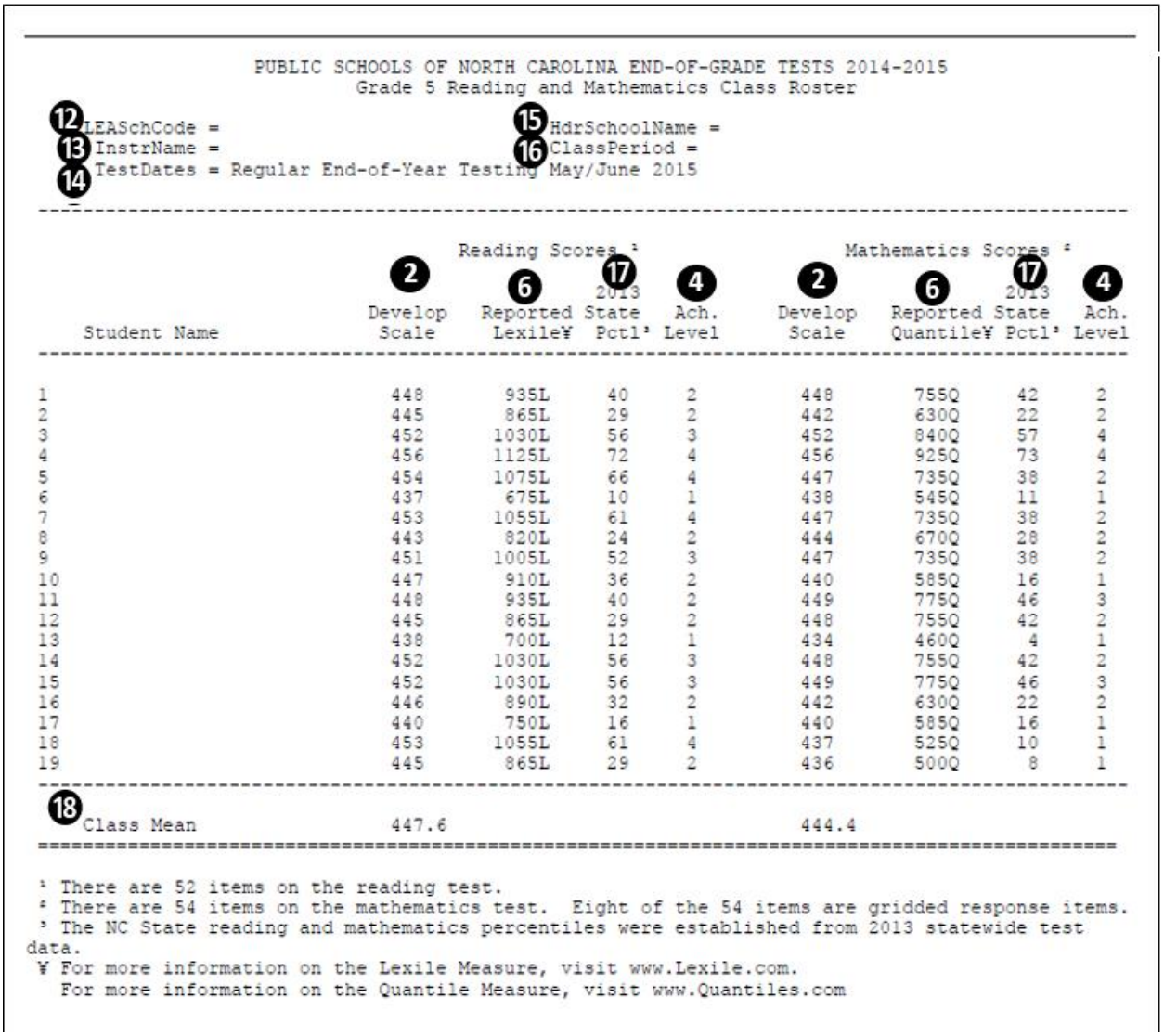
The Lexile Framework® for Reading (label 6) shows Lexile Framework® level that is associated with the EOG scale score. Additional information on Lexile can be found at http://www.lexile.com.

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 test a second time, the scale score would very likely fall 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 Pctl³	4 Ach. Level	2 Develop Scale	6 Reported Quantile¥	17 2013 State Pctl³	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

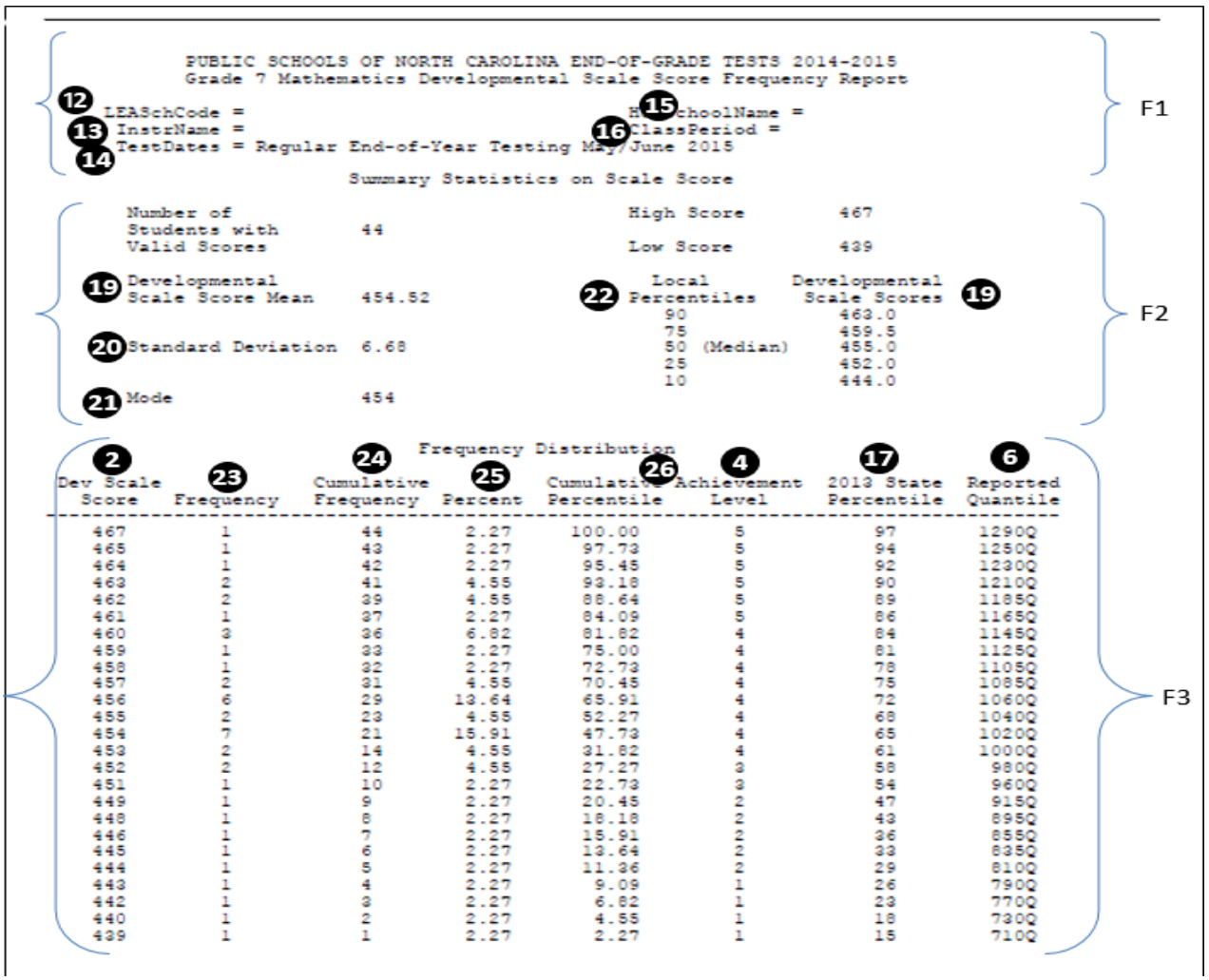
General information is reported from label 12 to label 16. LEASchCode (label 12) refers to the Local Education Agency (LEA) school code. InstrName (label 13) refers to the instructor's name. TestDates (label 14) refers to the time of year in which the exam was administered. HdrSchoolName (label 15) refers to the school name. ClassPeriod (label 16) refers to the class period. This report presents the same information as the ISR, but its main difference is that it displays theh score summary for all the students in a class. For mathematics (label 6), the Quantile® score is similar to Lexile score for ELA and shows the Quantile Framework® level that is associated with the EOG math scale score. 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 EOG Mathematics Assessment. The ELA report is similar just a different content.

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 HSchoolName =
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 developmental 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 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} 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 scores 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 that earned each scale score. 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.

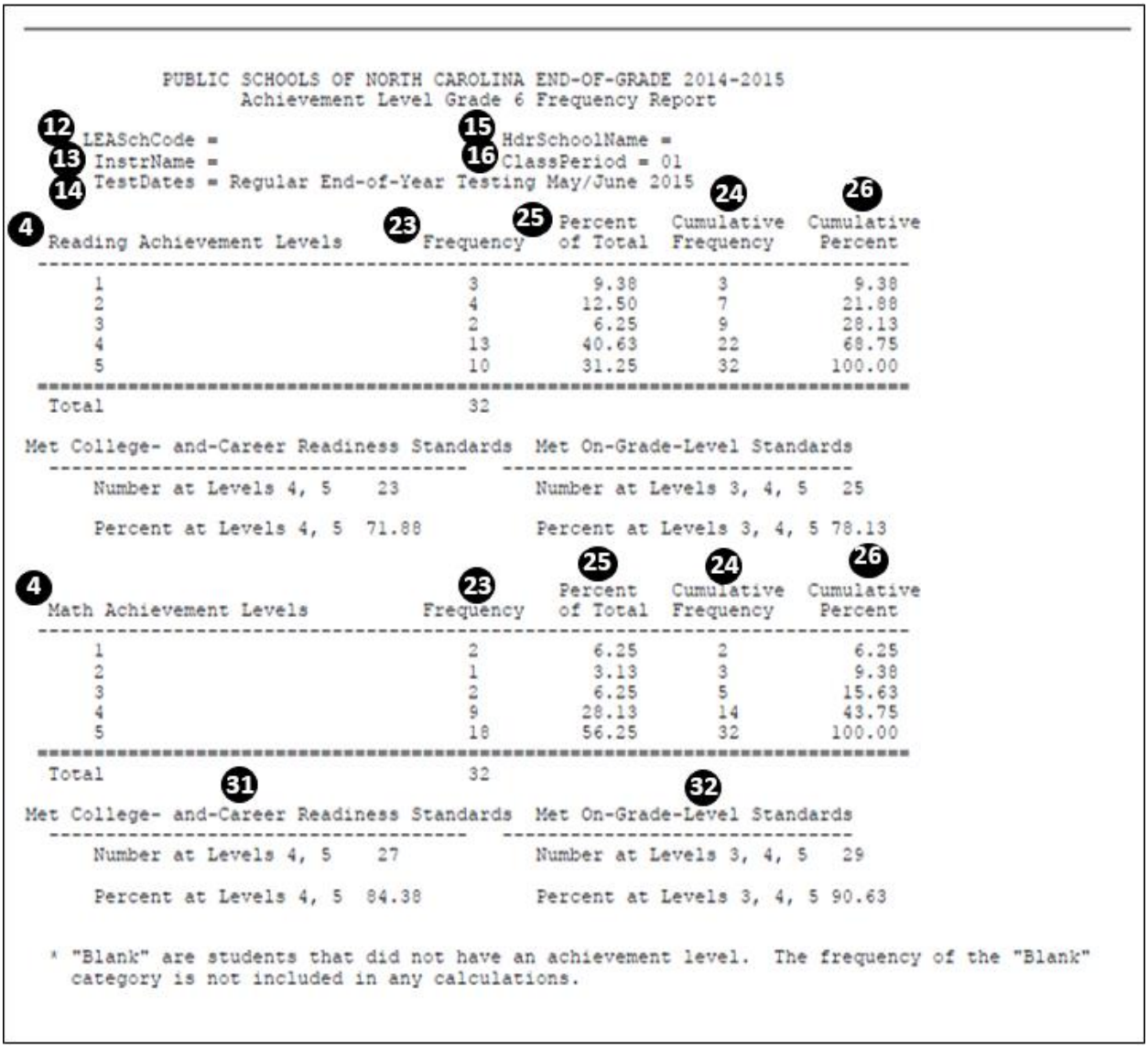
The 2013 State Percentile column (label 17) displays to the ELA/reading and mathematics percentiles that were established from 2013 statewide assessment data. This column shows that a scale score of 456 was in the 72nd percentile in 2013.

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. EOG ELA will display a corresponding Lexile column.

9.2.4 Achievement Level Frequency Reports

A sample Achievement Level Frequency Report for an EOG ELA/Reading 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			

31 Met College- and-Career Readiness Standards		32 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

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

In this sample, the exam was a regular administration (label 14). 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 / 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 the 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 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. The goal summary report contains goal-level score reporting and subscale scores which reflect items related to literary reading and items related to informational reading respectively. Subscales reported in the goal summary are 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 of 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 percentage 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 in 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, 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 documents validity evidence based content summarized from the alignment study and evidence based on relation to other variables summarized from the EOG/EOC Lexile linking study and the last part describes procedures used to ensure EOG and EOC assessments are accessible and fair to all students.

10.1 Reliability Evidence of ELA EOG and EOC English II

Internal consistency reliability estimates provide a sample base summary statistic that describes the proportion of reported scores which is the true score variance. In order to justify valid use of scores in large scale standardized assessments, evidence must be documented that shows test results are stable, consistent, and dependable across all subgroups of the intended population. A reliable test produces scores that are expected to be relatively stable if the test were administered repeatedly under similar conditions. Scores from a reliable test reflect 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 one refers to a perfectly reliable measures 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 accuracy of individual students' scores since they are sample based.

Table 10.1 ELA and English II reliabilities by Subgroup

EOG/EOC and Form		Gender		Ethnicity[1]			All
		Female	Male	Black	Hispanic	White	
ELA Grade 3	*A*	0.91	0.92	0.90	0.89	0.91	0.91
	B	0.91	0.92	0.91	0.91	0.91	0.92
	C	0.91	0.92	0.90	0.91	0.90	0.91
ELA Grade 4	*A*	0.89	0.89	0.87	0.88	0.88	0.89
	B	0.89	0.90	0.88	0.88	0.88	0.90
	C	0.87	0.88	0.86	0.86	0.86	0.88
ELA Grade 5	*A*	0.89	0.90	0.88	0.88	0.88	0.90
	B	0.87	0.89	0.86	0.87	0.87	0.88
	C	0.89	0.90	0.87	0.88	0.87	0.89
ELA Grade 6	*A*	0.89	0.90	0.87	0.87	0.88	0.89
	B	0.90	0.91	0.89	0.89	0.90	0.91
	C	0.89	0.90	0.88	0.88	0.89	0.89
ELA Grade 7	*A*	0.88	0.89	0.87	0.87	0.88	0.89
	B	0.89	0.90	0.87	0.88	0.88	0.90
	C	0.88	0.90	0.87	0.87	0.88	0.89
ELA Grade 8	*A*	0.87	0.88	0.86	0.86	0.87	0.88
	B	0.87	0.89	0.86	0.86	0.87	0.88
	C	0.87	0.88	0.85	0.86	0.88	0.88
English II	*A*	0.88	0.89	0.86	0.88	0.88	0.89
	B	0.88	0.89	0.87	0.88	0.87	0.89
	C	0.87	0.89	0.87	0.88	0.87	0.89
	M	0.89	0.90	0.87	0.89	0.89	0.89
	N	0.89	0.89	0.87	0.88	0.89	0.89
	O	0.88	0.89	0.87	0.88	0.88	0.89

Table 10.1 shows Cronbach alpha reliability estimates for all EOG and EOC ELA forms by grade and major demographic variables. Across all forms, reliability estimates from the 2012–

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

2013 population range from the high 0.88 to lower 0.91. Subgroups reliabilities are also consistent across forms and subgroups in the same range as the overall estimates. Exception to this general trend are recorded in subgroup (Black and Hispanic) reliabilities for some forms in grades 4, 7, and 8 in which the reported alpha is between 0.85 and 0.86.

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 helps understand the accuracy of examinees' classifications. It allows for a probabilistic statement to be made about the amount of precision on student's reported 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 ELA 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 scale points. For most of the EOG and EOC ELA scale scores, the standard error of measurement in the middle range of scores, particularly at the cut point between Level II and Level III, is generally around 3 points. Scores at the lower and higher ends of the scale (above the 97.5th percentile and below the 2.5th percentile) have standard errors of measurement of approximately 5 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 and Hoss/Loss by Form and Grade Level

		LOSS		Level 2		Level 3		Level 4		Level 5		HOSS	
ELA	Form	Loss	SE	Partial	SE	Sufficient	SE	Solid	SE	Superior	SE	Hoss	SE
EOG 3	A	406	5	432	3	439	3	442	3	452	4	462	6
	B	406	5	432	3	439	3	442	3	452	5	461	6
	C	406	5	432	3	439	3	442	3	452	5	461	6
EOG 4	A	412	5	439	3	445	3	448	3	460	4	468	6
	B	412	5	439	3	445	3	448	3	460	5	468	6
	C	412	5	439	3	445	3	448	4	460	5	468	6
EOG 5	A	419	5	443	3	450	3	453	3	464	4	472	6
	B	419	5	443	3	450	3	453	3	464	5	472	6
	C	418	5	443	3	450	3	453	3	464	4	476	6
EOG 6	A	418	6	442	4	451	3	454	3	465	4	478	6
	B	419	5	442	3	451	3	454	3	465	4	478	6
	C	416	5	442	3	451	3	454	3	465	4	478	6
EOG 7	A	419	6	445	4	454	3	457	4	469	5	482	6
	B	420	6	445	3	454	3	457	3	469	4	482	6
	C	421	5	445	4	454	3	457	3	469	4	483	6
EOG 8	A	422	6	449	4	458	4	462	4	473	5	487	6
	B	422	6	449	4	458	4	462	4	473	4	488	6
	C	423	6	449	4	458	4	462	4	473	4	487	6
English II	A	118	5	141	3	148	3	151	3	165	4	181	5
	B	121	5	141	4	148	3	151	3	165	4	180	5
	C	119	5	141	4	148	3	151	3	165	4	180	6
	M	118	5	141	3	148	3	151	3	165	4	179	5
	N	121	5	141	4	148	3	151	3	165	4	179	5
	O	119	5	141	3	148	3	151	3	165	4	179	5

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

The SEs at Level 2 and Level 3 across forms and grades ranged from 3 to 4, and Level 4 ranged from 4 to 5. One useful application of the conditional SE 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, on grade proficiency (Level 3) cut score for grade 3 ELA is 439. A student who took

Form A and scored 439 with a SE of 3 has a 68% probability that their true score or ability ranges from 436 to 442 (439±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 95% confidence interval of the student predicted ability will be from 433 to 445 (439±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. Classification consistency refers to "the agreement between classifications based on two non-overlapping, equally difficult forms of the test," and decision accuracy refers to "the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known" (Livingston & Lewis, 1995, P. 178). That is, classification consistency refers to the agreement between two observed scores, while classification accuracy refers to the agreement between observed and true scores.

The analyses are implemented using the computer program BB-Class.[m] The program provides results for both the Hanson and Brennan (1990 and Livingston and Lewis (1995) procedures. 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 where "(a) items are not equally weighted and/or (b) some or all of the items are polytomously scored" (Brennan, 2004, pp. 2-3), so the analyses for the EOG ELA Grade 3 to Grade 8 followed the HB procedures, and the analyses for EOC English II used LL procedures.

Table 10.3 presents the decision accuracy and consistency indexes for achievement levels at each grade. Overall, the values indicate good classification accuracy (ranging from 0.89 to 0.97) and consistency (from 0.84 to 0.96). For example, if Grade 3 ELA students who were classified as Level 2 take a non-overlapping, equally difficult form a second time, 92% 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	0.94	0.92	0.91	0.88	0.91	0.87	0.89	0.86
Grade 4	0.93	0.90	0.89	0.85	0.89	0.84	0.93	0.89
Grade 5	0.92	0.89	0.89	0.85	0.89	0.84	0.94	0.92
Grade 6	0.94	0.92	0.90	0.87	0.90	0.86	0.91	0.88
Grade 7	0.94	0.92	0.90	0.86	0.89	0.85	0.92	0.89
Grade 8	0.93	0.90	0.89	0.85	0.89	0.85	0.93	0.91
English II	0.94	0.91	0.91	0.87	0.91	0.87	0.97	0.96

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

10.4 EOG and EOC Dimensionality Analysis

Evidence of overall dimensionality for ELA, EOG, and EOC assessments was explored using Principal Component Analysis (PCA). PCA is an exploratory technique that seeks to summarize observed variables into fewer linear dimensions referred to as components. The

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

primary question in a PCA analysis is to determine the fewest number of reasonable dimensions or components that can explain most of the observed variance in the data. Two 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 a 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 form 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 14 -17 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 7 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 ELA Grade 3 Scree Plot of Operational Forms

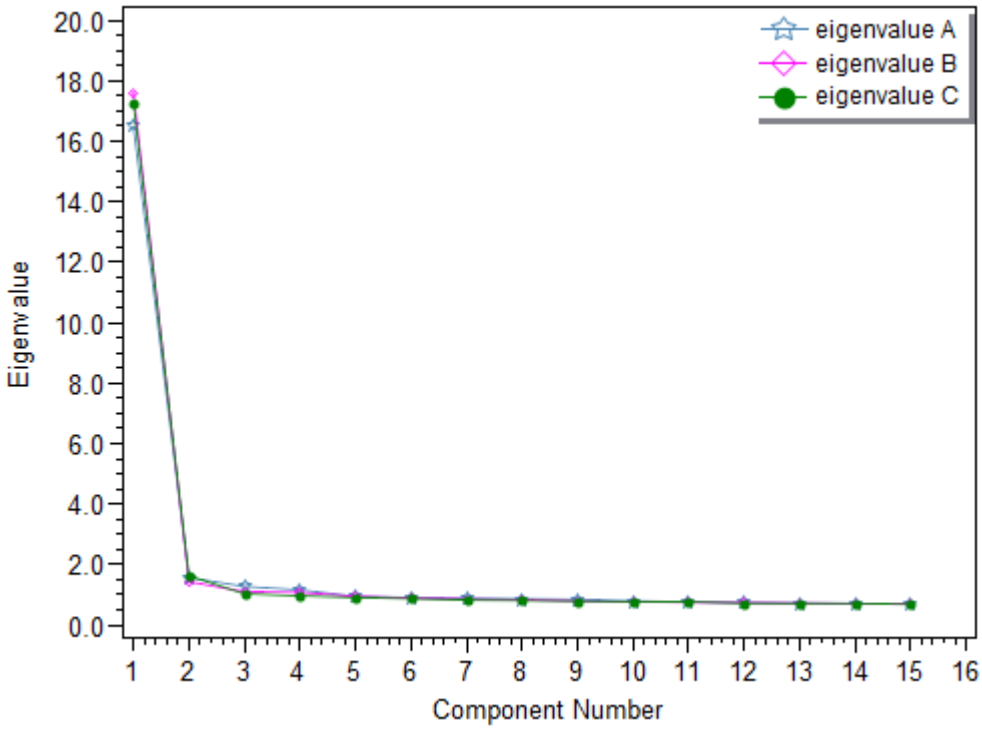


Figure 10.2 ELA Grade 4 Scree Plot of Operational Forms

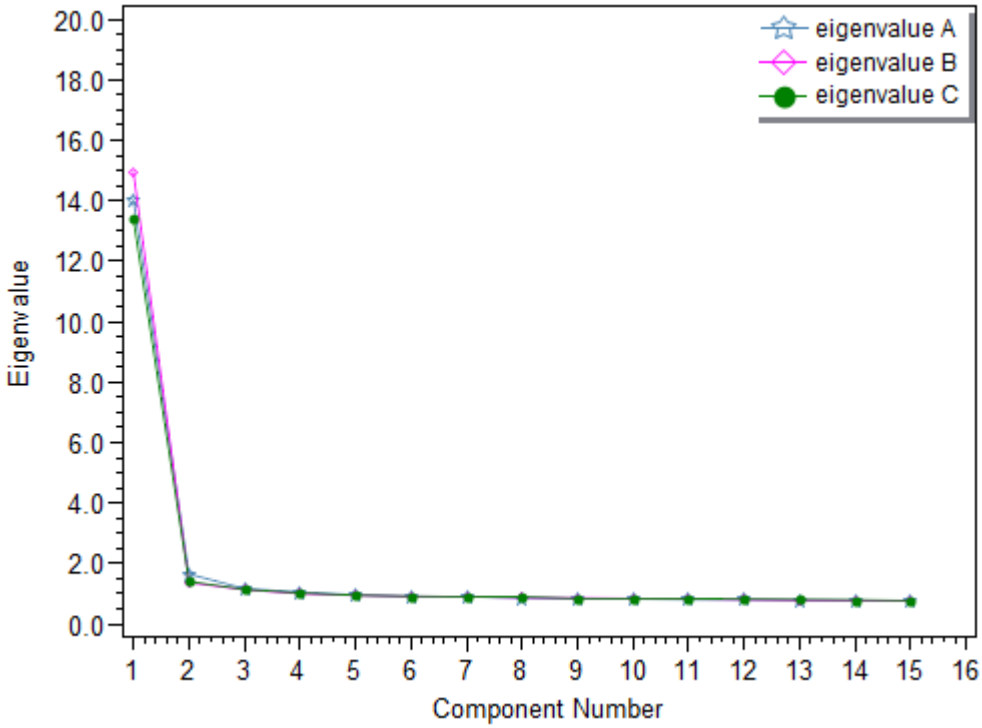


Figure 10.3 ELA Grade 5 Scree Plot of Operational Forms

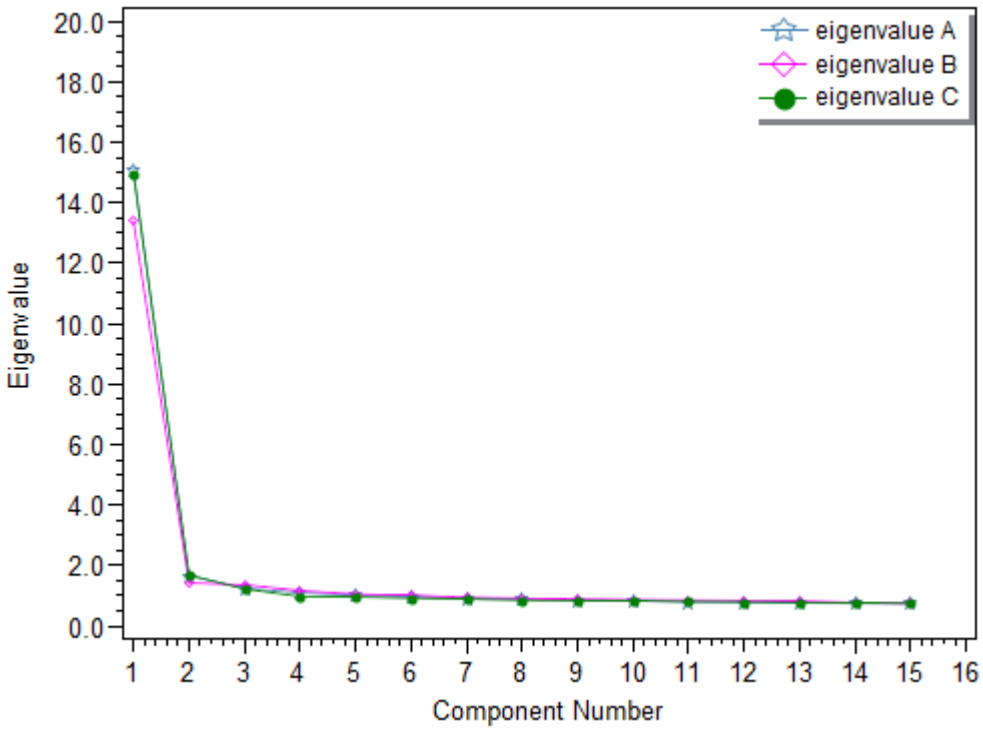


Figure 10.4 ELA Grade 6 Scree Plot of Operational Forms

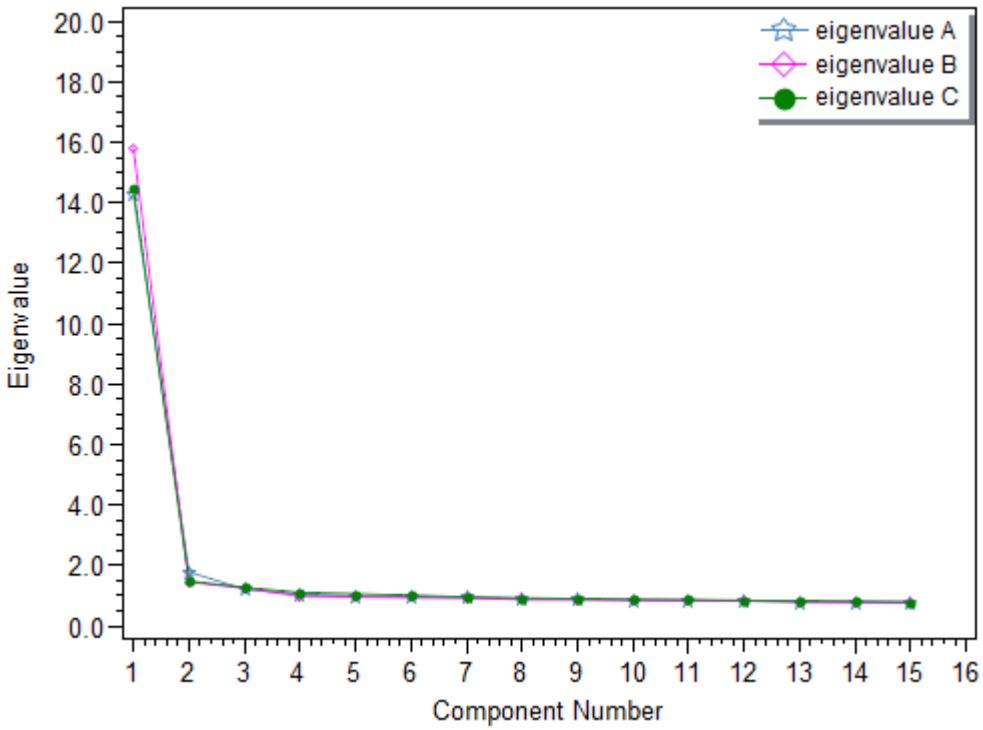


Figure 10.5 ELA Grade 7 Scree Plot of Operational Forms

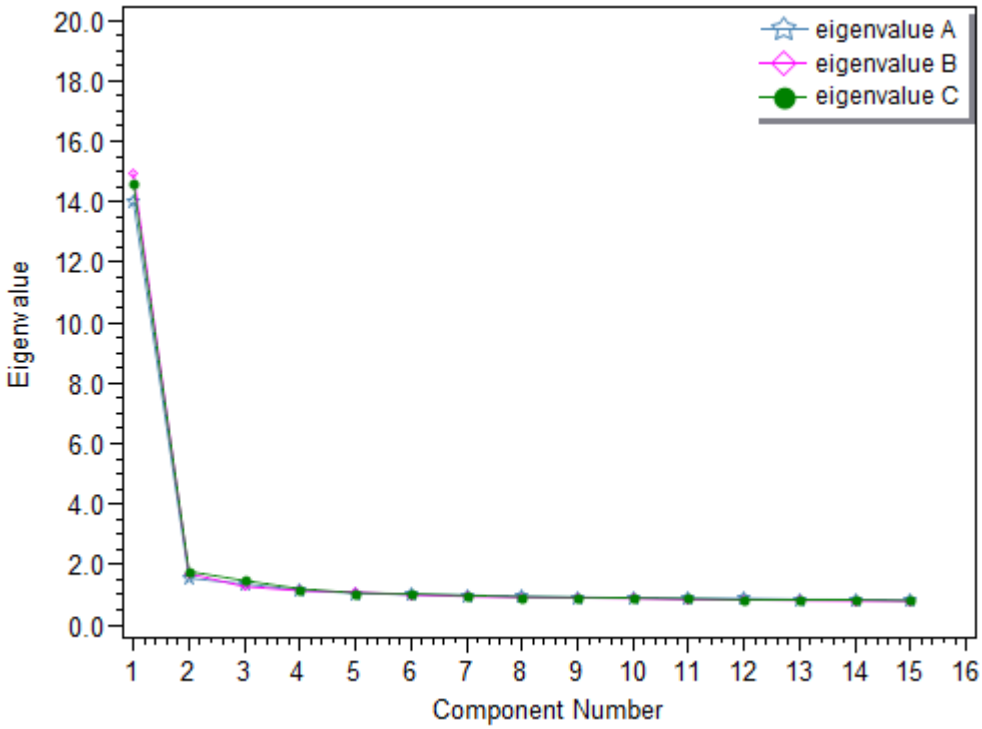


Figure 10.6 ELA Grade 8 Scree Plot of Operational Forms

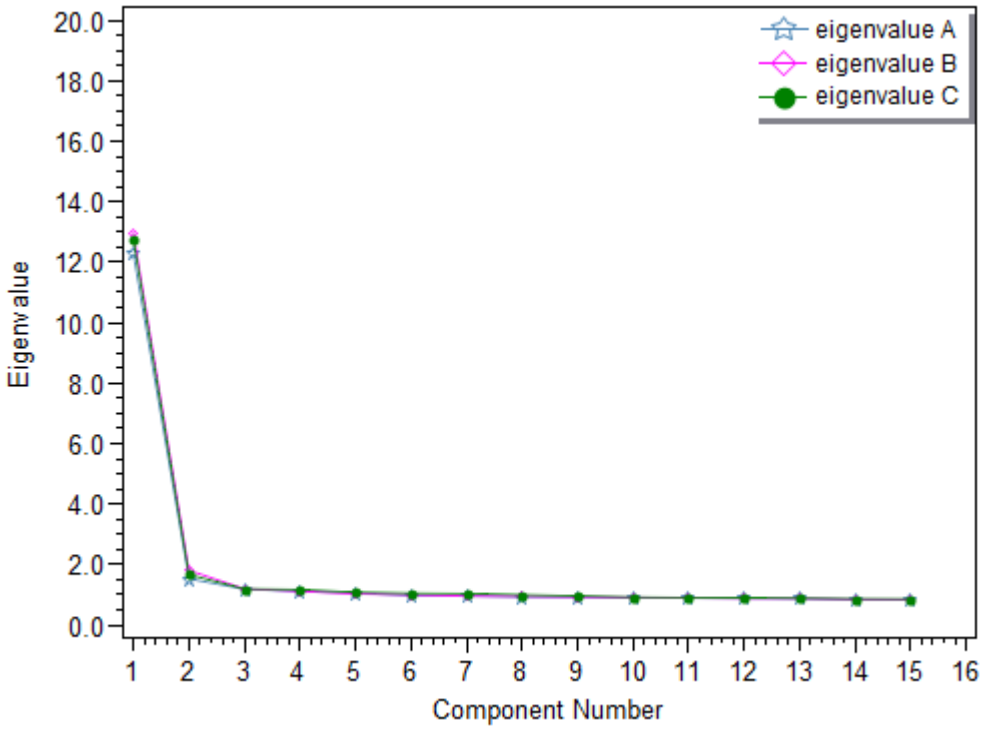
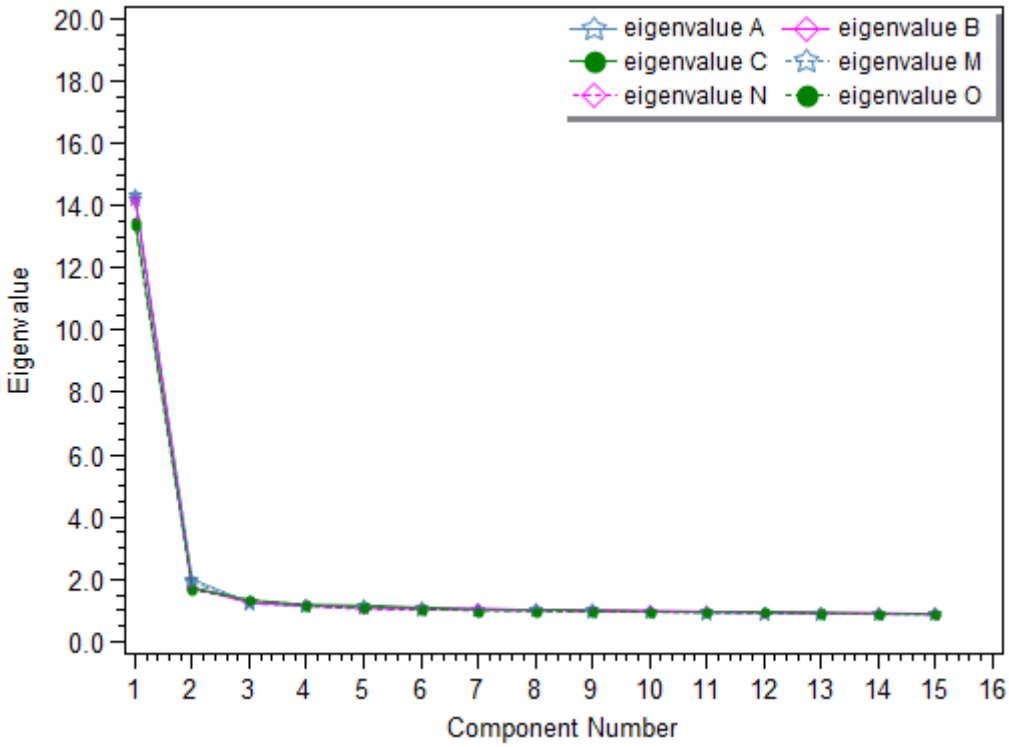


Figure 10.7 English II Scree Plot of Operational Forms



10.5 Alignment Study

In September, 2014 the North Carolina Department of Public Instruction commissioned the Wisconsin Center for Education Research (WCER) 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

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.

10.5.4 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.

10.5.5 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	English II
BR	0.59	0.71	0.70	0.66	0.64	0.67	0.70

Table 10.4 displays BR index by grade for the NC End-of-Grade assessments for grades 3–8 and the End-of-Course assessments of English II. 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.

10.5.6 Topic Coverage

The first dimension considered in most, if not all alignment analyses, regardless of the methodology employed, concerns what Norman Webb (1997) calls categorical concurrence. For convenience, and to better fit the SEC terminology, this indicator is simply referred to as topic coverage (TC) and measures a seemingly simple question; does the topic or 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	English II
TC	0.65	0.64	0.64	0.72	0.86	0.81	0.88

Table 10.5 provides the summary alignment results for TC for each of the assessed grades in mathematics and reading 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.

10.5.7 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	English II
PE	0.86	0.59	0.67	0.83	0.66	0.64	0.65

Table 10.6 reports the performance expectations measured across assessed grade levels for reading. It is expressed as an index with a range of 0 to 1, with 0.50 indicating acceptable alignment. As can be seen, all subjects/grades surpass this threshold.

10.5.8 Alignment Results

While the SEC approach to alignment allows reporting and consideration of the results along each of these three dimensions, the most powerful alignment measure 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. Grade 4 reading appears more borderline because each of the sub-measures are above 0.5, but the PE measure for both is noticeably lower than TC and BR, again suggesting that any alignment issues related to these assessments will likely center on performance expectations.

Table 10.7 Overall Alignment Index by Grade

Grade	**EOG 3**	**EOG 4**	**EOG 5**	**EOG 6**	**EOG 7**	**EOG 8**	**English II**
OAI	0.58	**0.47**	0.52	0.62	0.64	0.62	0.57

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

For example, *Figure 10.9* indicates that the topics with the strongest emphasis in North Carolina's grade 4 assessable standards (map to the right "Target Content Areas") are comprehension and critical reading, particularly at the performance level of "analyze and generate" (equivalent to DOK levels 2 and 3). A careful visual review of the two maps in *Figure 10.9* in terms of the three alignment dimensions indicates the following:

- Balance of Representation (BR): The two figures are shaped similarly which indicates a good balance of representation for EOG grade 4 assessments. This is also confirm by a BR index of 0.71 see *Table 10.4*.

- Topic Coverage (TC): topics with the strongest emphasis are comprehension and critical reading, where the contour lines are closer together. This indicate the assessment blueprint is aligned to the content standards with respect to TC. The TC index for EOG grade 4 is 0.64 above the threshold of 0.50 see *Table 10.5*.

- Performance Expectation (PE): PE focuses on what students should "*be able to do*" more generally summarized by DOK levels. From the grade 4 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 "analyze and generate" (DOK 3 and 4). Analysis from the content map suggest that the weak alignment in grade 4 EOG is likely centered on performance expectations.

Figure 10.8 EOG Grade 3 Assessment and Standard content map

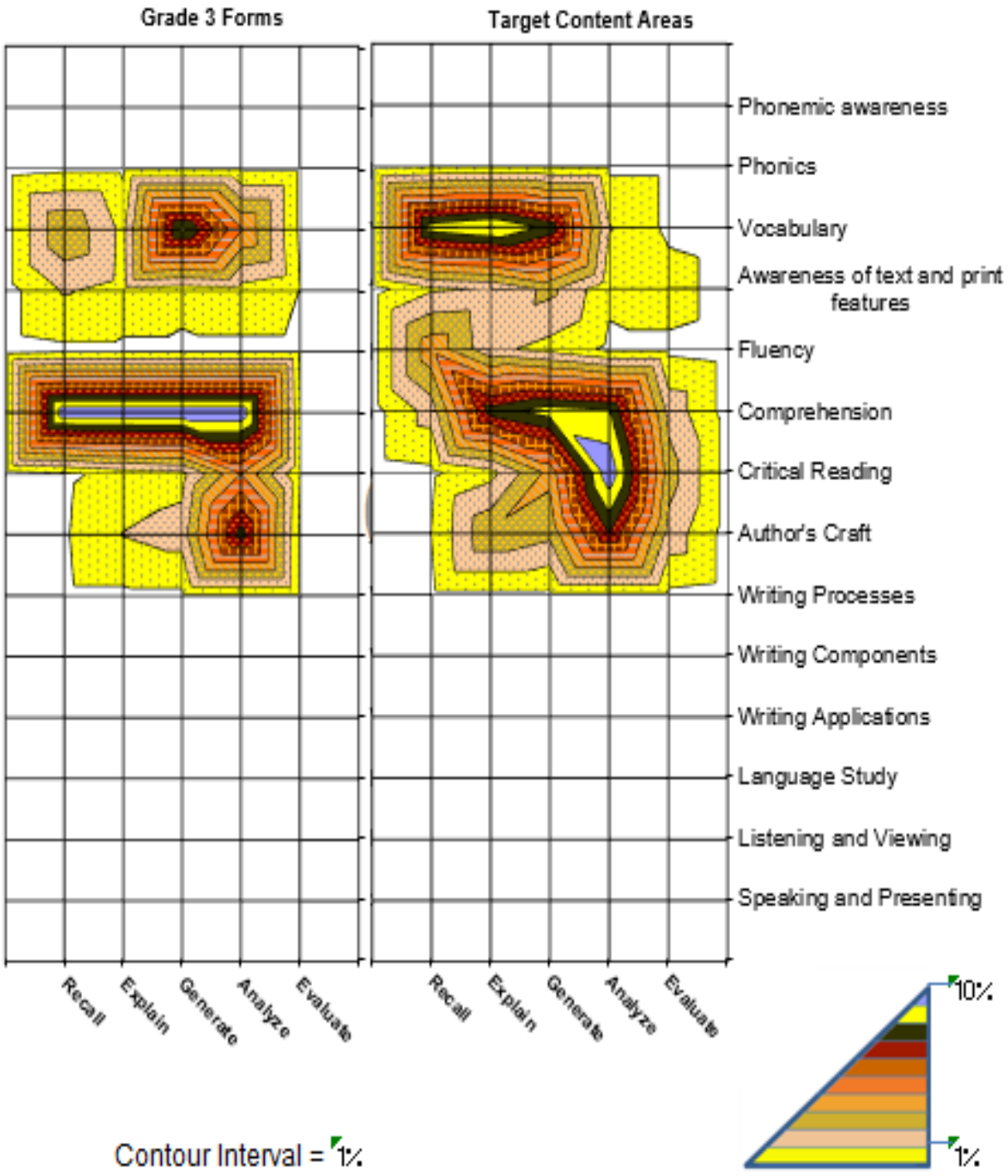


Figure 10.9 EOG Grade 4 Assessment and Standard content map

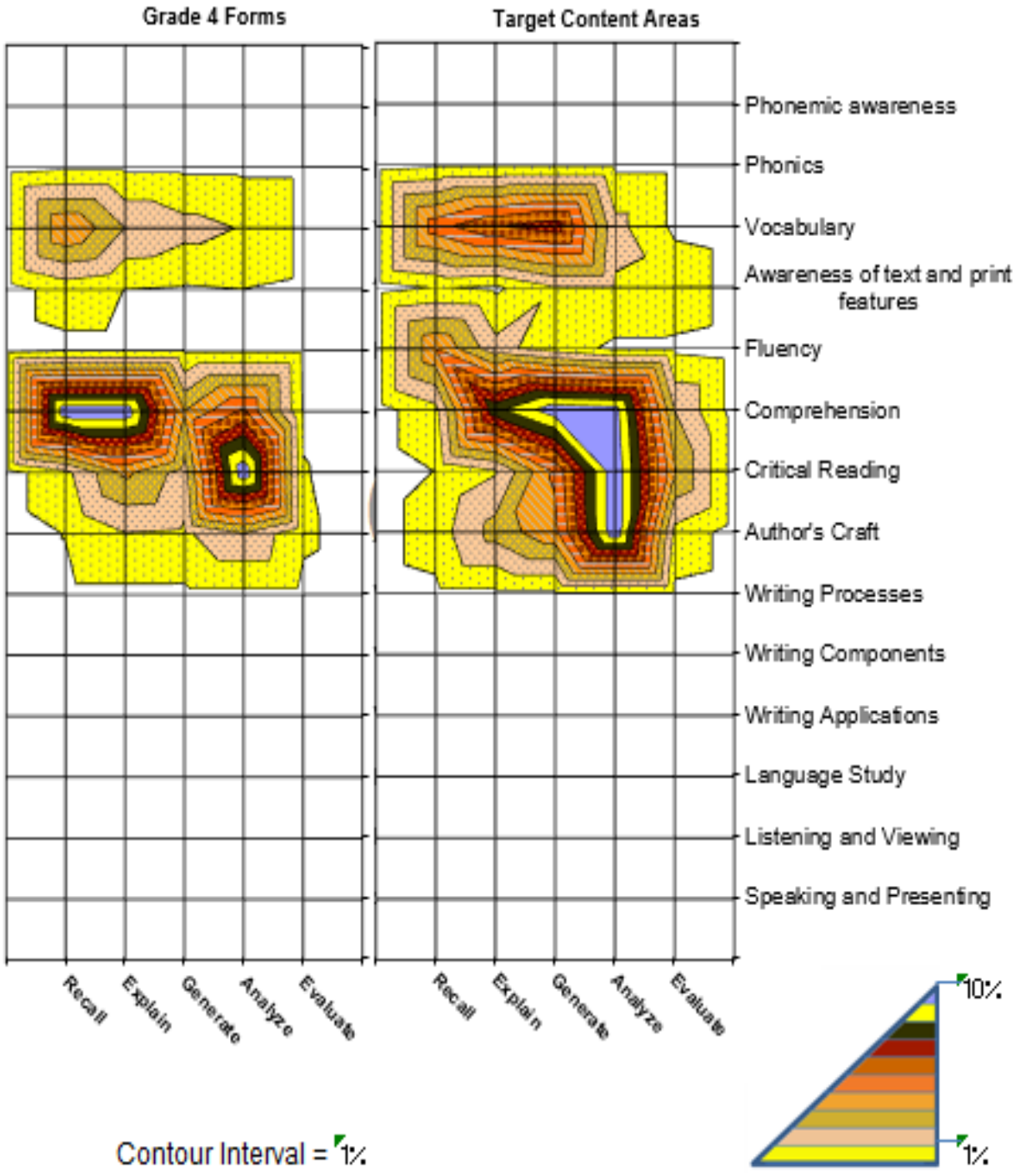


Figure 10.10 EOG Grade 5 Assessment and Standard content map

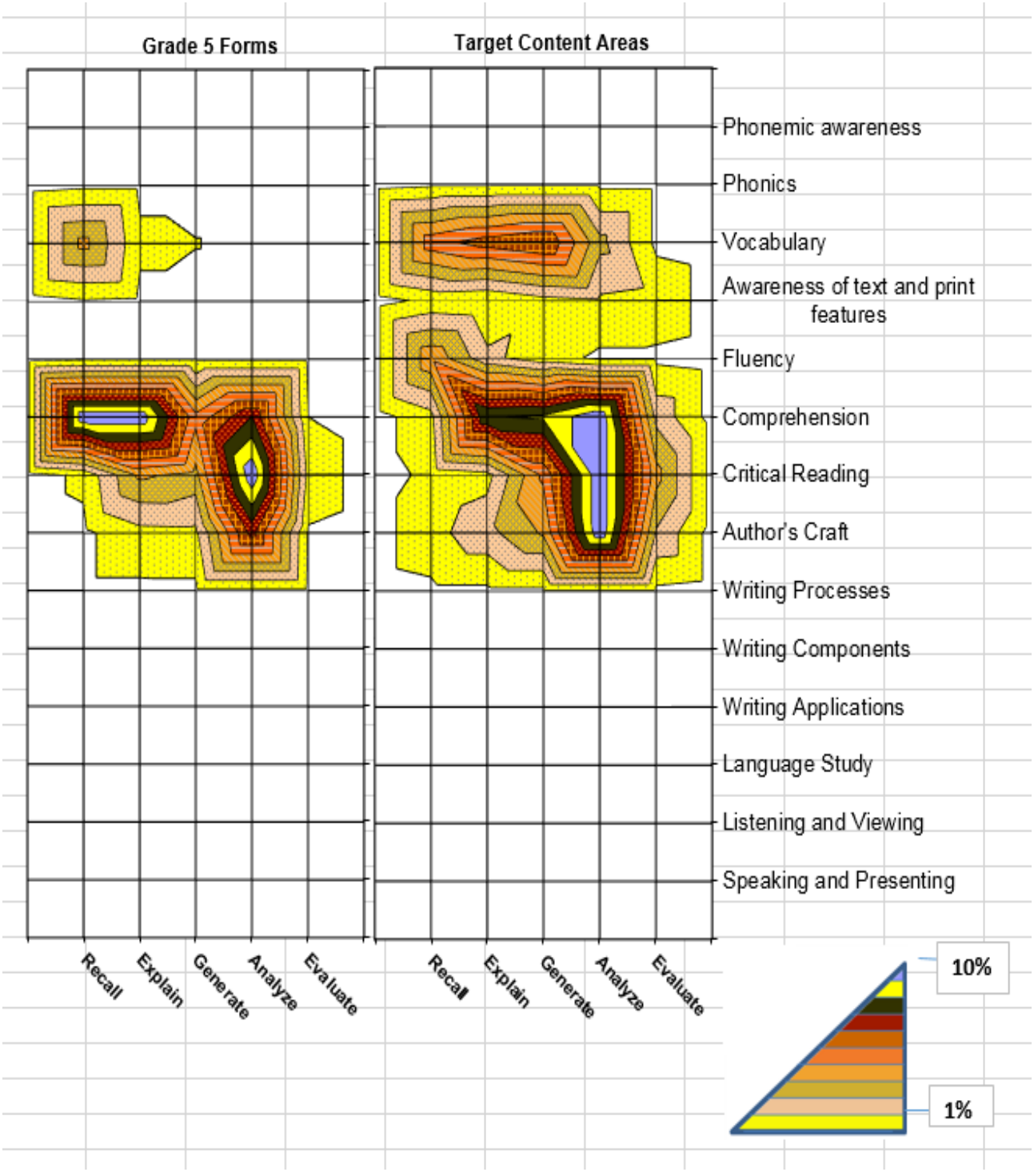


Figure 10.11 EOG Grade 6 Assessment and Standard content map

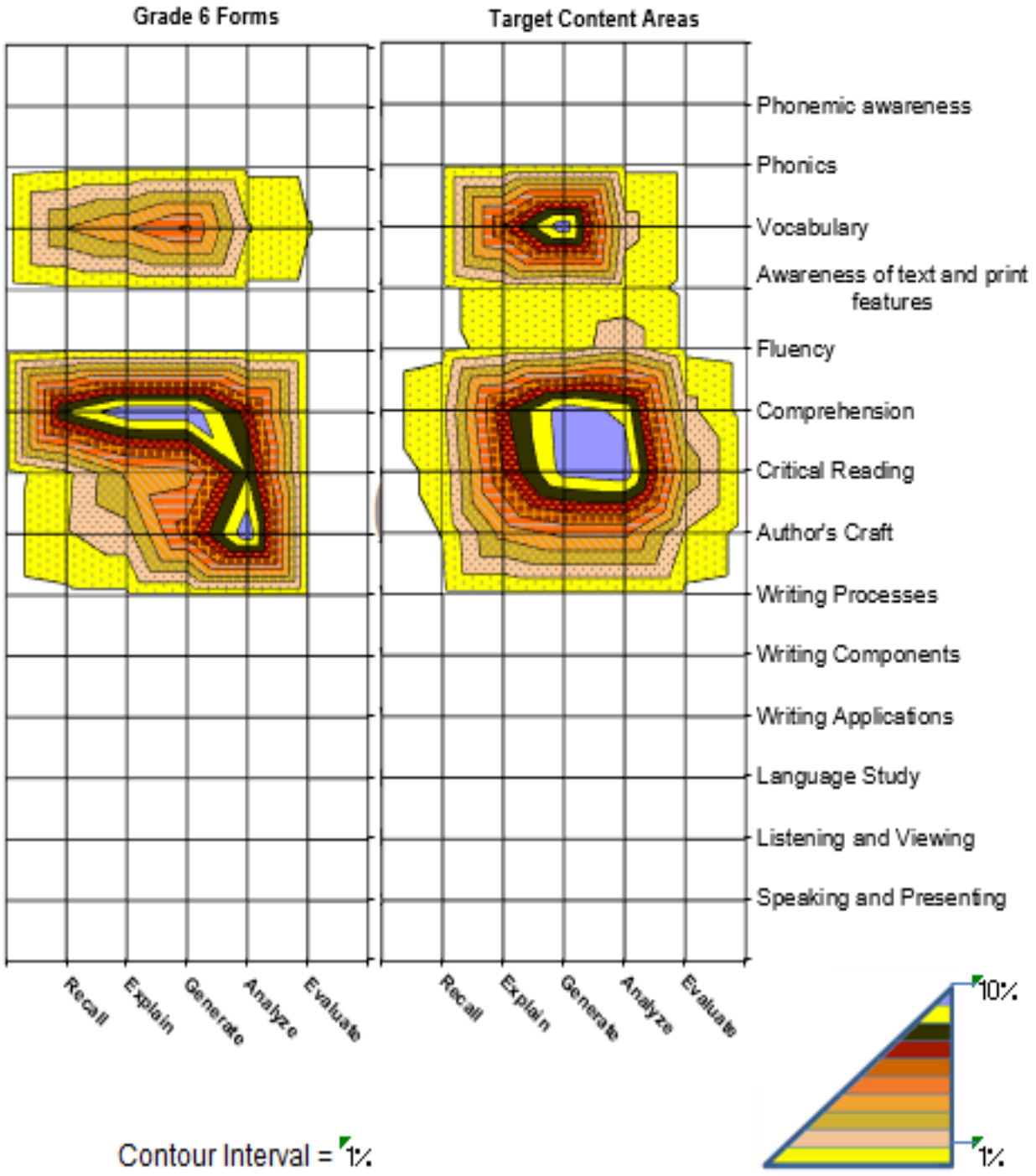


Figure 10.12 EOG Grade 7 Assessment and Standard content map

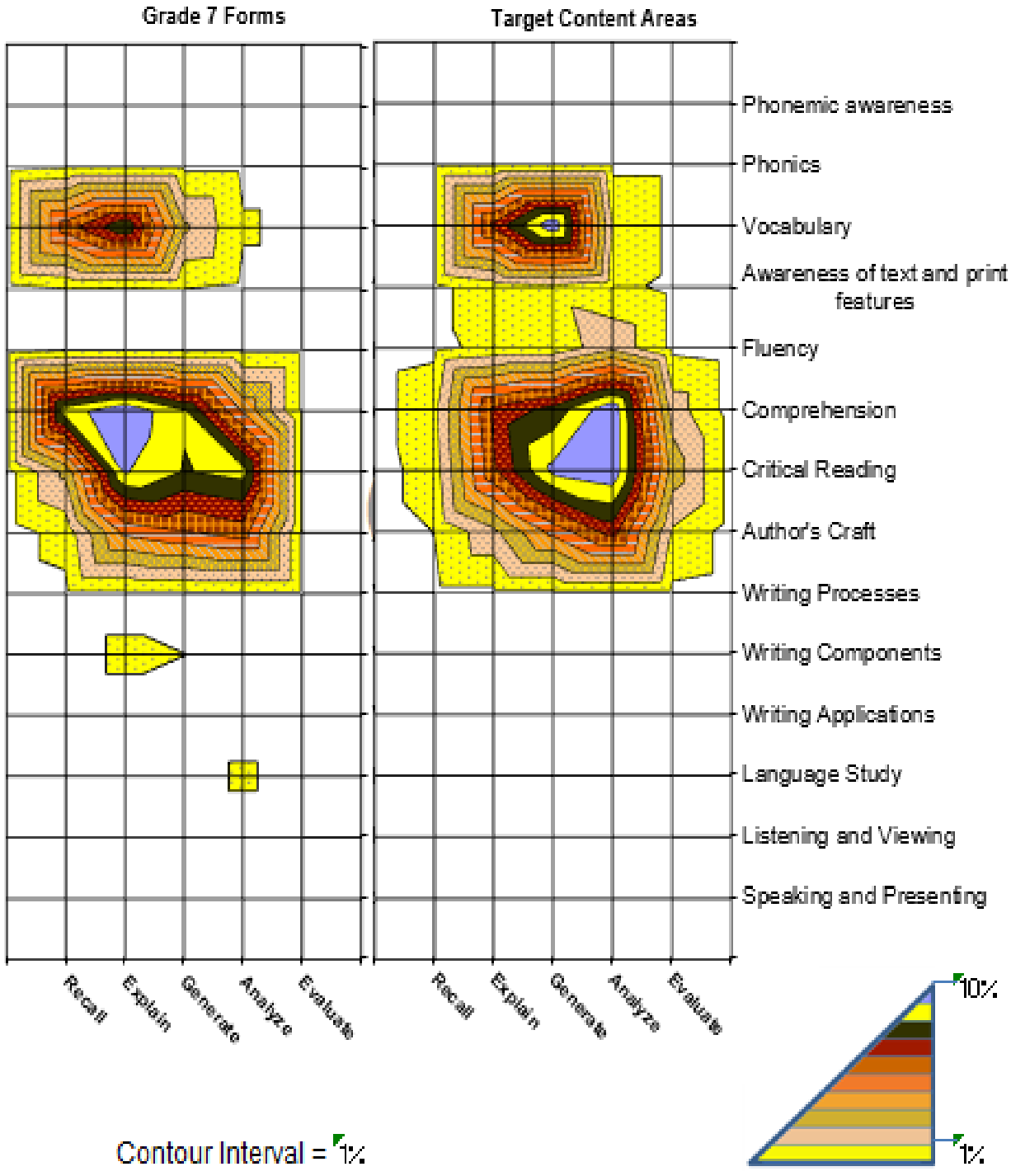


Figure 10.13 EOG Grade 8 Assessment and Standard content map

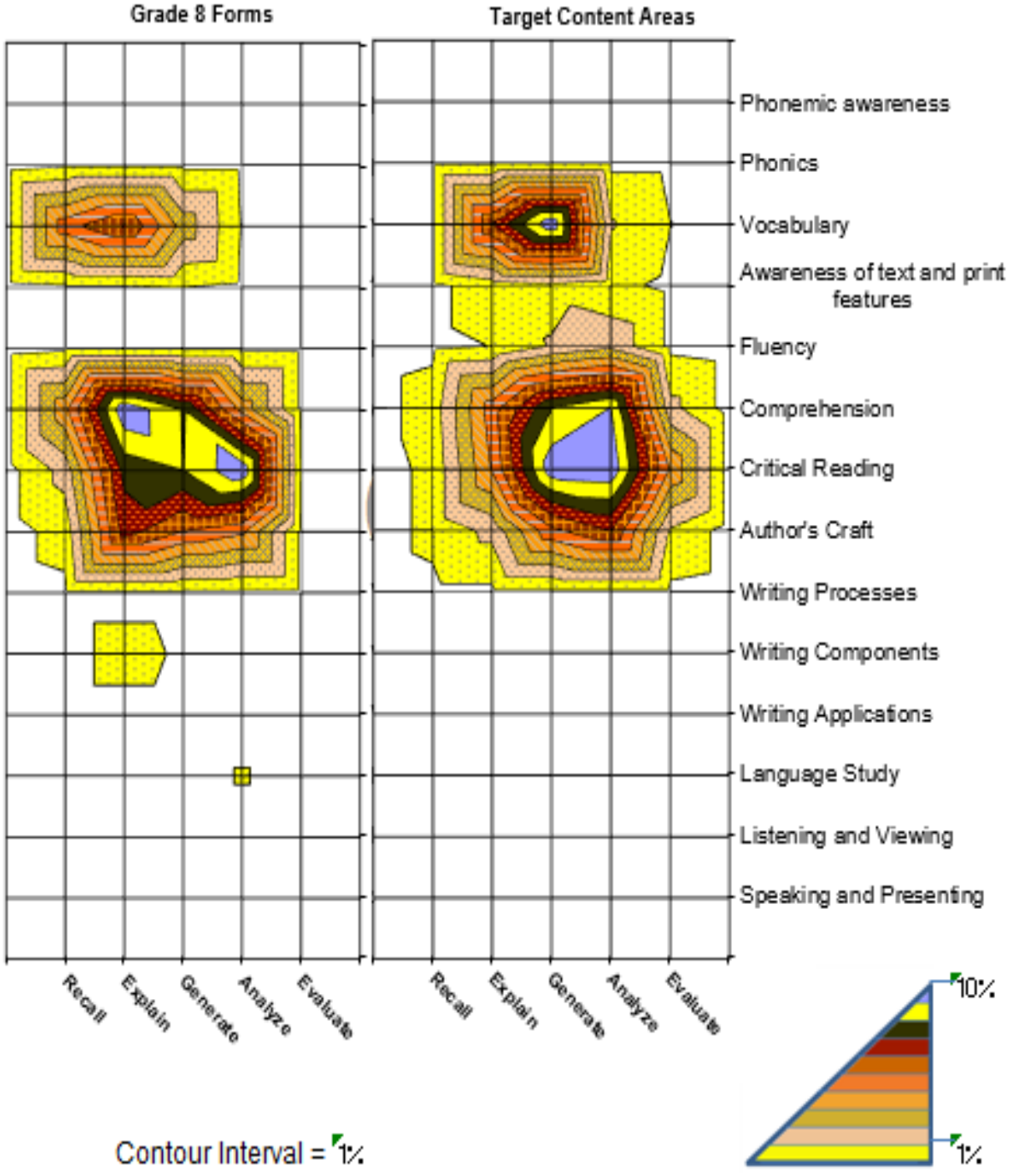
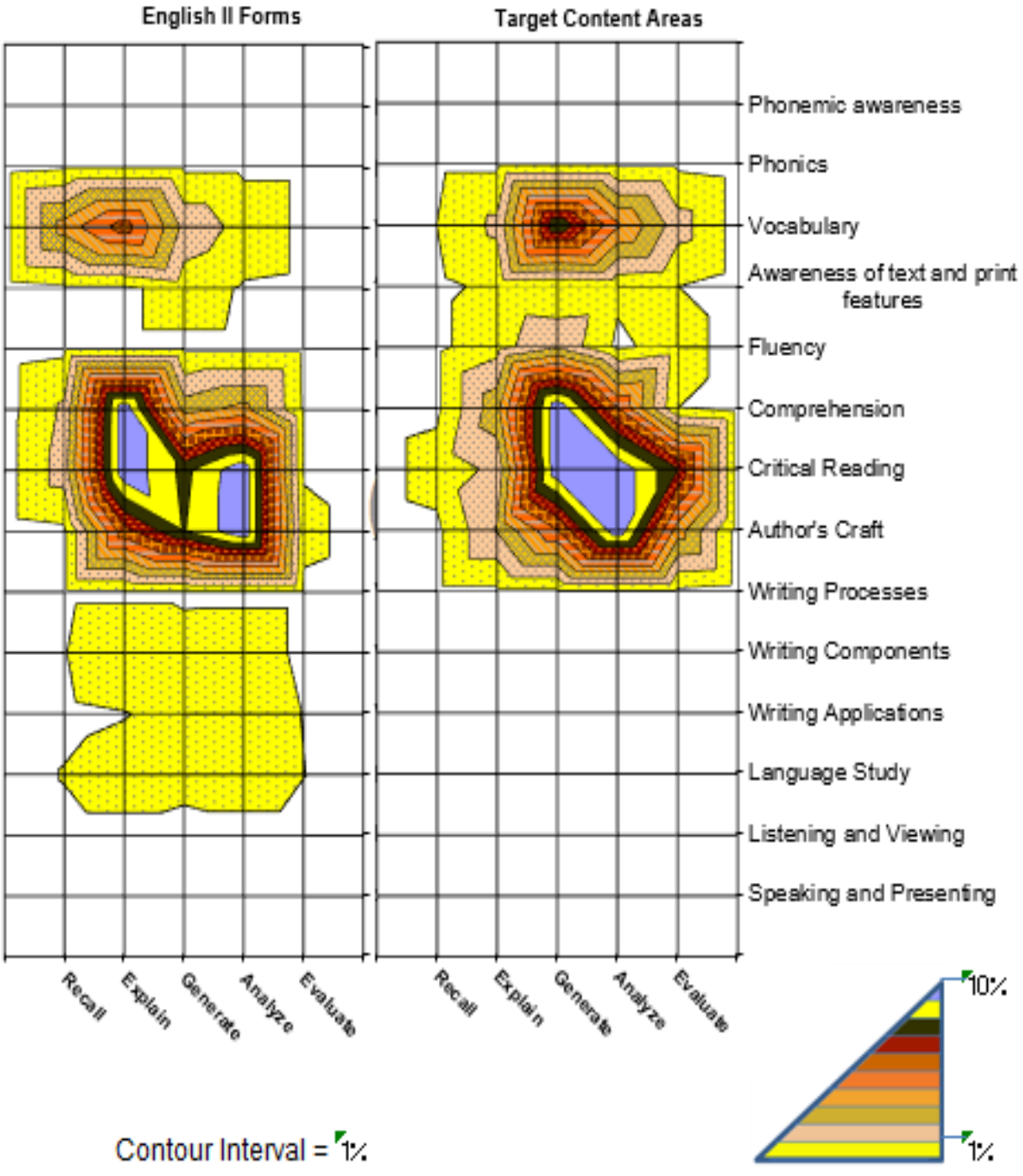


Figure 10.14 EOC English II Assessment and Standard content map



10.5.9 Discussion of Findings

As indicated by the results presented above, with the exception of grade 4 ELA, the assessments used by the state across the grades covered by this study reveal 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 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 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.

The reading assessment alignment results are very strong, with 27 of 28 indicators across all grade levels easily exceeding the 0.5 threshold. The one exception is for OAI at grade 4. Fine grain results summarized using the content maps presented in *Figure 10.8* through *Figure 10.14* indicate two separate alignment issues related to the grade 4 assessment. One concerning the breadth of sub-topics assessed within Vocabulary (a topic coverage issue), and the other concerns the performance expectations targeted for reading content associated with Comprehension (a performance expectation issue). Within Vocabulary, results indicate that the assessment touches on only one Vocabulary topic among 13 touched on by the grade 4 standards. Within content associated with Comprehension, fine grain results indicate that alignment would be improved with a shift in performance expectations from Recall and Explain to Use and Analyze.

These can be challenging performance expectations to address in a standardized multiple-choice assessment format, and while other formats are possible, they are expensive and present their own challenges, including scoring reliability and validity.

Once student performance data has been collected (Phase III of the study), additional information will be available regarding the impact of the assessments' alignment characteristics

on student performance, controlling for students' opportunities 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 reading in grades 3 through 8, as well as one end of course assessment for high school reading. 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 reading assessments administered by the state are for the most part very well aligned. Marginally low alignment measures were noted for grade 4 reading.

10.6 Evidence Regarding Relationships with External Variables

One of the primary intended uses of the EOG and EOC ELA assessments is to provide data to measure students' achievement and progress relative to readiness as defined by College- and Career-Readiness standards. For the ELA assessments to provide evidence of this type of achievement, it is important that reading passages are an appropriate measure of college and career readiness. To examine the level of reading required by the NC EOG and EOC ELA assessments, NCDPI commissioned MetaMetrics, Inc. to examine the relationship

of the ELA assessments to the Lexile Framework® for reading (Contract No. NC10025818 dated December 17, 2012).

The primary purpose of this linking study was to provide parents and teachers with reading levels (i.e., Lexile score) to predict the books and texts a student should be matched with for successful reading experiences, given their performance on the NC READY EOG Reading/EOC English II assessment. A secondary purpose was to examine the reading level of the NC READY EOG Reading/EOC English II assessments to determine if there is support for the claim that the assessments are a measure of college and career readiness. This section summarizes important evidence from the report. The full report may be found in Appendix 10-A Lexile Linking Technical Report 2013.

10.6.1 The Lexile Framework for Reading

The Lexile Framework is a tool that can help teachers, parents, and students locate challenging reading materials. Text complexity (difficulty) and reader ability are measured in the same unit—the Lexile. Text complexity is determined by examining such characteristics as word frequency and sentence length. Items and text are calibrated using the Rasch model. The typical range of the Lexile Scale is from 200L to 1600L, although actual Lexile measures can range from below zero (BR) to above 2000L.

MetaMetrics, Inc. has collected a good amount of validity evidence over the past three decades to show that the Lexile Framework measures reading comprehension and text difficulty. This evidence includes demonstrating strong relationships between (1) the Lexile Framework and other measures of reading comprehension (e.g., other standardized assessments); (2) the Lexile Framework and Basal readers; and (3) the Lexile Framework and the difficulty of reading test items.

10.6.2 Linking the Lexile Framework to the NC Assessments

The Lexile Framework was linked to the NC Assessments through linking tests designed to be as similar as possible to the NC READY EOG Reading/EOC English II assessments, including the number of operational items per test and the difficulty of the items. The items for the Lexile Linking Tests were chosen to optimize the match to the target test. The IRT difficulty values associated with the NC READY EOG Reading/EOC English II items were converted to

Lexile measures using a computer program developed by MetaMetrics, Inc. Details of the linking are provided in the full report (see Appendix 10-A).

Table 10.8 presents the achievement level cut scores on the NC READY EOG Reading/EOC English II assessments and the associated Lexile measures based on the liking study. The North Carolina Department of 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 for each category.

Table 10.8 NC READY EOG Reading/EOC English II performance level cut scores and the associated Lexile measures[n].

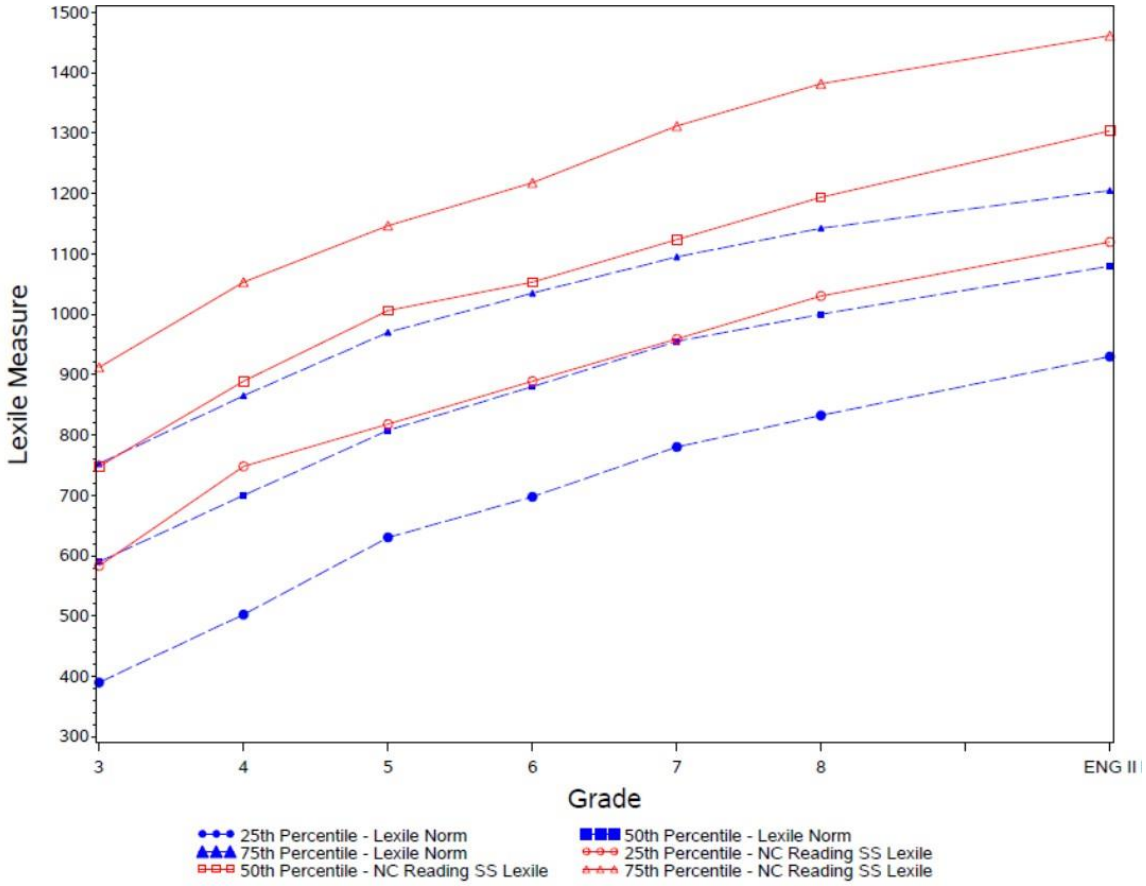
Grade	Level 3		Level 4		Level 5	
	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure
3	439	725L	442	795L	452	1030L
4	445	865L	448	935L	460	1220L
5	450	985L	453	1055L	464	1310L
6	451	1005L	454	1075L	465	1335L
7	454	1075L	457	1145L	469	1430L
8	458	1170L	462	1265L	473	1525L
E II	148	1225L	151	1305L	165	1670L

Figure 10.15 shows the Lexile measures for the NC READY EOG Reading/EOC English II assessment as compared to the norms that have been developed for use with The Lexile Framework for Reading. These norms were created based on linking studies conducted with the Lexile Framework. Overall, it can be seen that the NC READY EOG Reading/EOC English II Lexile measures are higher across the grades at each percentile. The 25th percentile for the NC READY EOG Reading/EOC English II Lexile measures is closer to the 50th

[n] 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

percentile of the Lexile measures. The 50th percentile for the NC READY EOG Reading/EOC English II Lexile measures is closer to the 75th percentile of the Lexile measures. Therefore, the NC READY EOG Reading/EOC English II scores were higher than the Lexile norms. This translates to the statement that the students in North Carolina were more able than the Lexile norms for a national population.

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



10.6.3 The Lexile Framework and College- and Career-Readiness

As noted above, one purpose of this study was to examine the reading level associated with the NC READY EOG Reading/EOC English II Assessments. If these assessments are to provide information about college- and career-readiness, then the reading level of the

assessments must be an appropriate measure of college- and career-readiness. It would undermine the credibility of the NC assessments ability to measure college- and career-readiness if the reading levels of the Reading and English assessments were below grade level. If, however, they align to Lexile measures associated with college- and career-readiness, then this is evidence supporting the use of the NC assessments.

Table 10.9 shows the Lexile ranges aligned to college- and career-readiness (NC Level 3 and 4 (2013) and Level 4 and 5 (2014). This continuum can be "stretched" to describe the reading demands expected of students in Grades 1–12 who are "on track" for college and career (Sanford-Moore and Williamson, 2012). *Table 10.9* also shows the Lexile levels of the Level 4 cut score for each NC Reading and English assessment. The Lexile score associated with the Level 4 cut score is either at the upper limit or above the Lexile ranges for college- and career-readiness.

Table 10.9 *Lexile ranges aligned to college- and career-readiness expectations, by grade.*

Grade	2012 "Stretch" Text Measure	Lexile Associated with Level 4 Cut score
1	190L to 530L	
2	420L to 650L	
3	520L to 820L	795L
4	740L to 940L	935L
5	830L to 1010L	1055L
6	925L to 1070L	1075L
7	970L to 1120L	1145L
8	1010L to 1185L	1265L
9	1050L to 1260L	
10	1080L to 1335L	
11-12	1185L to 1385L	

Figure 10.16 shows the relationship between the "Old Level 3" performance standard for each grade level established on the NC READY EOG Reading/EOC English II Assessment and the "stretch" reading demands. This shows that the NC READY EOG Reading/EOC English II performance standards for "Level 3" at each grade level is set at a level that is consistent with being "on track" for college- and career-readiness at the end of Grade 12.

Figure 10.16 Comparison of NC READY EOG Reading/EOC English II "Old Level 3" standards with college and career reading levels described by the CCSS.

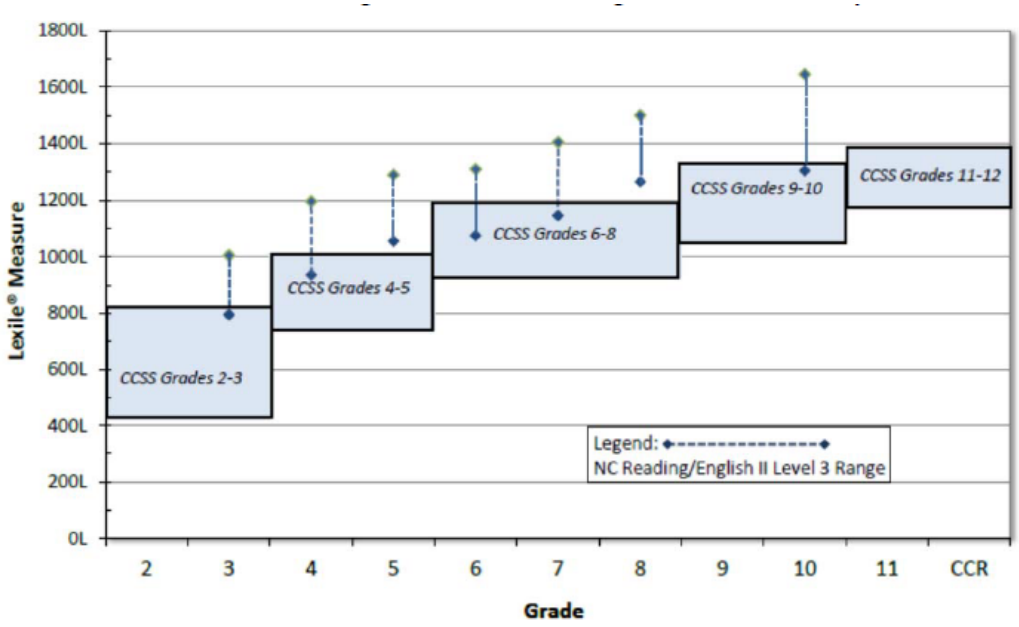
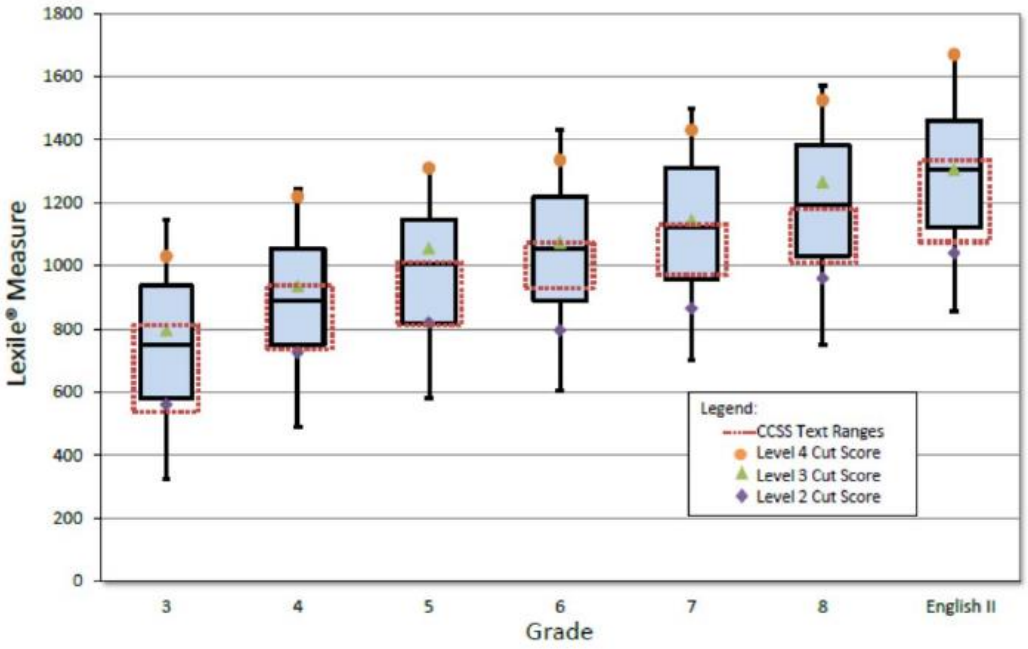


Figure 10.17 shows that the spring 2013 student performance on the NC READY EOG Reading/EOC English II assessments at each grade level is "on track" for college- and career-readiness. Students can be matched with reading materials that are at or above the recommendations in Appendix A of the CCSS for ELA for each grade level.

Figure 10.17 NC READY EOG Reading/EOC English II 2012-2013 student performance expressed as Lexile measures.



In 2008, MetaMetrics and the North Carolina Department of Public Instruction conducted a study to link the NC EOG Reading Test with the Lexile scale (MetaMetrics, 2008). The minimum score considered "proficient" (Level 3 or current Level 4) at each grade level on the NC EOG Reading is presented in *Table 10.8*. In 2013, NCDPI transitioned their assessment program to the NC READY EOG Reading Assessment to align with the Common Core State Standards in English/Language Arts and to describe student reading performance in relation to college- and career-readiness. One outcome of this change was to set the performance standards for NC READY EOG Reading at a higher level. The Lexile scale can be used as an external "yardstick" to evaluate this change in reading demand on the North Carolina reading assessment. The information in *Table 10.10* shows that the NC READY EOG Reading standards demanded more of students in terms of reading ability in 2013.

Table 10.10 Minimum "Level 3" Lexile measure on NC EOG Reading (2008) and NC READY EOG Reading (2013).

Grade	"Proficient" Level 3 Cut Score (2008)	"Proficient" Level 3[o] Cut Score (2013)
3	665L	795L
4	790L	935L
5	940L	1055L
6	990L	1075L
7	1115L	1145L
8	1165L	1265L

10.6.4 Conclusions

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

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

[o] Level 4 beginning 2014 using the 5 achievement level scale

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[p]
    - 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
    - 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

[p] See discussions on bias review in Chapter 4

    - 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

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 a 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.

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 pvalue and reliability, and they had closely aligned TCCs and 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 includes:

- 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		Test forms built using the same blueprint and match on difficulty and content.
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.

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

Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, *35*, 93-107.

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

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

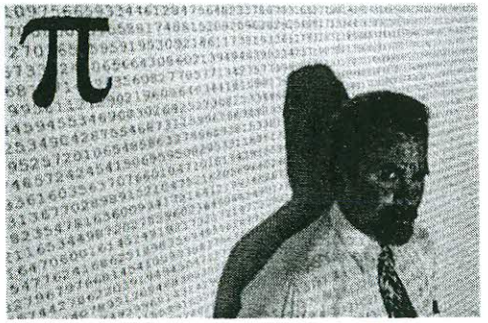
1) 190
2) 200
3) 290
4) N

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

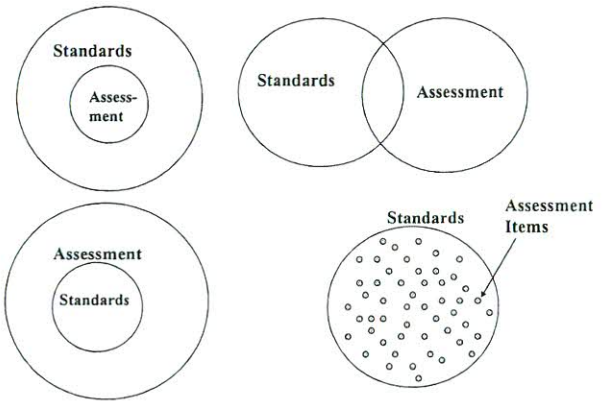
Bloom's Revised Taxonomy of Cognitive Process Dimensions	Webb's Depth-of-Knowledge (DOK) Levels			
	Level 1 Recall & Reproduction	Level 2 Skills & Concepts	Level 3 Strategic Thinking/ Reasoning	Level 4 Extended Thinking
Remember				
Understand				
Apply				
Analyze				
Evaluate				
Create				

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

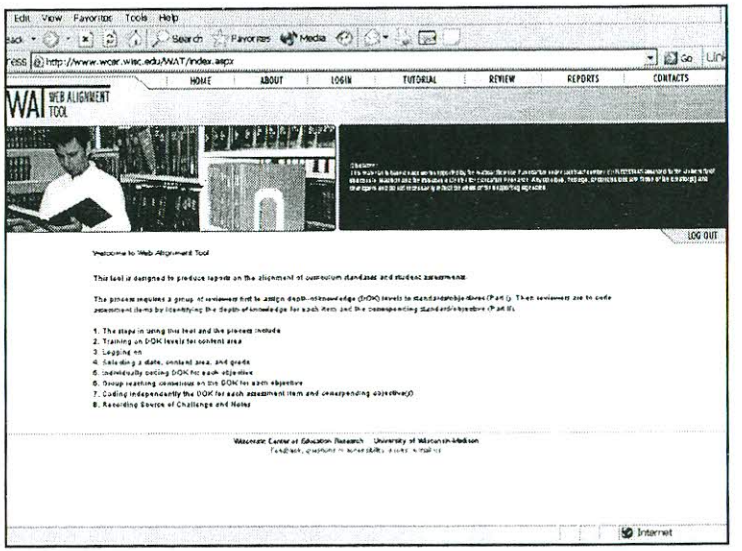
D. Balance of Representation

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 English Language Arts

READY EOG Assessments, Grades 3–8
READY EOC English II Assessments

North Carolina Assessment Specifications Summary

Purpose of the Assessments

- Edition 4 Grades 3–8 English Language Arts (ELA) assessments and the high school English II assessments will measure students' proficiency on the NC *Standard Course of Study* (NC*SCS*) for English Language Arts, 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 Department of Public Instruction K-12 English Language Arts wiki site at http://elaccss.ncdpi.wikispaces.net/Common+Core+State+Standards.
- Every grade has a set of content standards that define what all students are expected to know and be able to do by the end of the grade.
- The ELA NC *Standard Course of Study* is divided into 4 strands: reading, writing, speaking and listening, and language.

Prioritization of Standards

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

Table 1
Weight Distributions for Grades 3–5

Domain	Grade 3	Grade 4	Grade 5
Reading for Literature	32–37%	30–34%	36-40%
Reading for Information	41–45%	45–49%	37-41%
Reading Foundation Skills	NA	NA	NA
Writing	NA	NA	NA
Speaking and Listening	NA	NA	NA
Language	20–24%	19–21%	21-25%
Total	**100%**	**100%**	**100%**

Table 2
Weight Distributions for Grades 6–8

Standard	Grade 6	Grade 7	Grade 8
Reading for Literature	32–36%	34–38%	31–35%
Reading for Information	41–45%	41–45%	42–46%
Writing	NA	NA	NA
Speaking and Listening	NA	NA	NA
Language	21–25%	19–23%	20–24%
Total	**100%**	**100%**	**100%**

Table 3
Weight Distributions for High School English II

Standard	English II
Reading for Literature	30–34%
Reading for Information	32–38%
Writing	14–18%
Speaking and Listening	NA
Language	14–18%
Total	**100%**

- Appendices A–G show the number of operational items for each standard assessed for the 2014-15 forms. Note that future coverage of standards could vary within the constraints of the content category weights in *Tables 1–3*.

Cognitive Rigor and Item Complexity
Assessment items has been 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 English Language Arts. These items will require students to not only recall information, but also apply concepts and skills, make decisions, and explain or justify their thinking.

Types of Items

- The Grades 3–8 English Language Arts assessments consist of four-response-option multiple-choice items. Multiple-choice items will be worth one point each.
- The high school English II assessment consists of four-response-option multiple-choice items, technology-enhanced items (online mode only), and constructed-response items.
- The English II assessment includes four constructed response items. One constructed response item is an embedded field test item and will not be included in the student's score but will be used for purposes of developing items for future test forms. Three constructed response items are operational and will be included in the student's score.
- The constructed response items will be short answer and can typically be answered well in a paragraph or less. These short answer items will be worth two points each. Students will write their responses on the lines provided on the answer sheet. Students must not write beyond the end of the lines or in the margins. Words written in the margins or unlined areas of the answer sheet will not be scored. Students must not add more lines to the answer sheet. Words written on extra lines will not be scored. Scorers only review for the specific criteria as stated in the item. Additional information not required in the answer does not increase the student's score.
- Released forms are available at http://www.ncpublicschools.org/accountability/testing/releasedforms. Released forms may be used by school systems to help acquaint students with items. These materials must not be used for personal or financial gain.
- The ***NCEXTEND1*** ELA alternate assessment consists of fifteen performance-based, multiple-choice items.

Delivery Mode

- Grades 3–8 ELA assessments are designed for a paper-and-pencil administration. The Grade 7 English Language Arts/Reading assessment will be available for online administration effective with the 2014–15 spring administration. The Grade 8 English Language Arts/Reading assessment will be available for online administration effective with the 2015–16 spring administration.
- The high school English II assessment has been designed for an online administration but will also be available in a paper version.
- ***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*** ELA and high school English II assessments has been designed for paper/pencil administrations with online data entry by the assessor.
- End-of-grade and end-of-course assessments are only provided in English. Native language translation versions are not available.

Appendix A
Grade 3 English Language Arts
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 test specification weights. 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.

Grade 3 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	3–5
RL.2	1–2
RL.3	4–5
RL.4	4–6
RL.5	–
RL.6	–
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	6–9
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	1–3
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	6
RI.2	3
RI.3	3
RI.4	2–4
RI.5	–
RI.6	–
RI.7	2–5
RI.8	2
RI.9	–
RI.10	–

* 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 English Language Arts
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 test specification weights. 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.

Grade 4 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	4–6
RL.2	1–2
RL.3	2–3
RL.4	4–5
RL.5	–
RL.6	–
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	5–7
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	2–4
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	3–6
RI.2	3–4
RI.3	4
RI.4	3
RI.5	2–3
RI.6	–
RI.7	–
RI.8	4–5
RI.9	–
RI.10	–

* 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 English Language Arts
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 test specification weights. 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.

Grade 5 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	4–7
RL.2	1–5
RL.3	2–7
RL.4	3–6
RL.5	–
RL.6	2–3
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	2–4
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	0–4
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	5–7
RI.2	2–4
RI.3	3
RI.4	4–5
RI.5	–
RI.6	–
RI.7	–
RI.8	2–3
RI.9	–
RI.10	–

* 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 English Language Arts
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 test specification weights. 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.

Grade 6 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	3
RL.2	2–3
RL.3	2–4
RL.4	4–5
RL.5	3–4
RL.6	–
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	6–7
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	4
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	3–5
RI.2	3–4
RI.3	2–3
RI.4	3–4
RI.5	2–4
RI.6	1–4
RI.7	–
RI.8	1–3
RI.9	–
RI.10	–

* 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 English Language Arts
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 test specification weights. 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.

Grade 7 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	4–5
RL.2	2–3
RL.3	1–4
RL.4	6–7
RL.5	1–2
RL.6	1–2
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	4
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	4
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	4
RI.2	2–3
RI.3	2–5
RI.4	2–4
RI.5	3
RI.6	2
RI.7	–
RI.8	3
RI.9	–
RI.10	–

* 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 English Language Arts
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 test specification weights. 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.

Grade 8 Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	3–4
RL.2	4
RL.3	1–4
RL.4	4
RL.5	–
RL.6	0–3
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3	–
L.4.a	3–5
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	5–7
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	4
RI.2	1–3
RI.3	2–3
RI.4	1–2
RI.5	4–5
RI.6	4–5
RI.7	–
RI.8	3–4
RI.9	–
RI.10	–

* 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
English II
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 test specification weights. 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.

English II Standard	Number of Operational Items by Standard*
RL.1 (Reading: Literature)	3–9
RL.2	2–4
RL.3	0–4
RL.4	3–6
RL.5	1–3
RL.6	1–3
RL.7	–
RL.9	–
RL.10	–
L.1 (Language)	–
L.2	–
L.3.a	–
L.4.a	5
L.4.b	–
L.4.c	–
L.4.d	–
L.5.a	3–5
L.5.b	–
L.6	–
RI.1 (Reading: Informational Text)	3–7
RI.2	2–5
RI.3	1–4
RI.4	5–7
RI.5	2–5
RI.6	3–6
RI.7	–
RI.8	–
RI.9	–
RI.10	–
W.1 (Writing)	–
W.4	–
W.9.a	–
W.9.b	–

* 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:

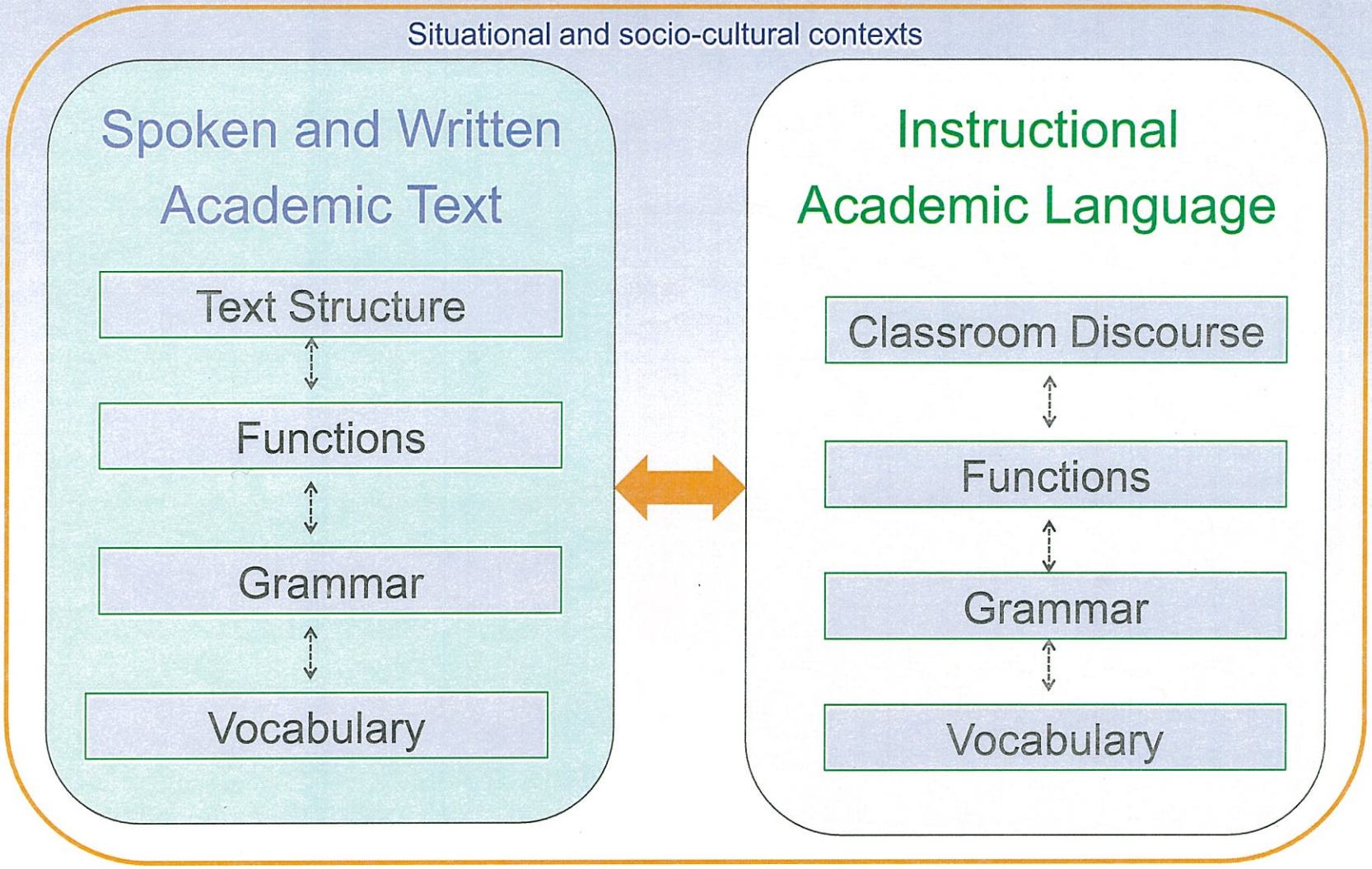
- to help analyze the content and language in standards, assessment tasks, and instructional materials;
- to help make explicit the expectations (cognitive, language) of students;
- to help inform instructional planning and practice so that they are intentional and appropriate in supporting students'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

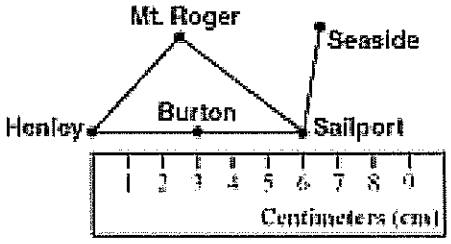
A. 2 days

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?

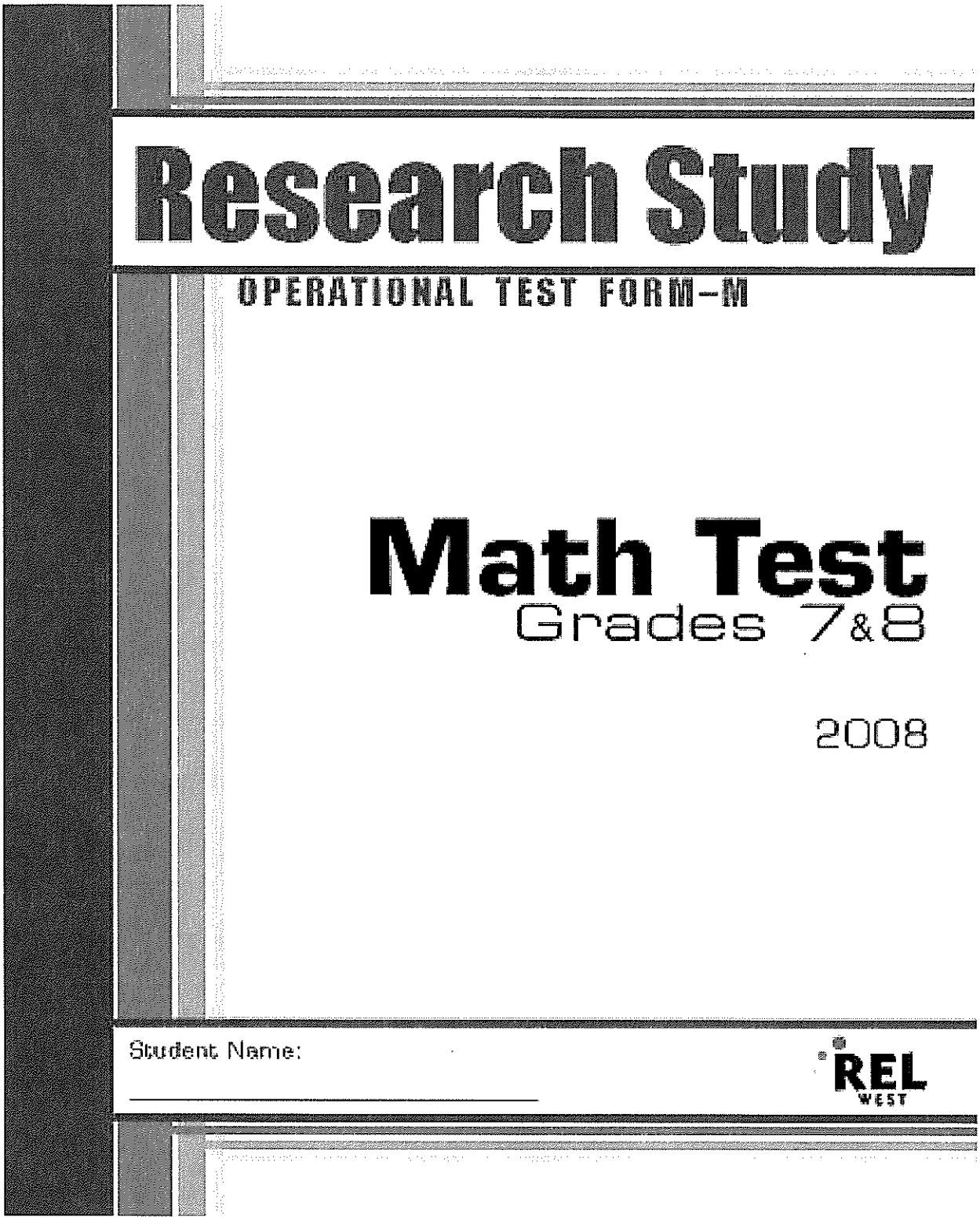
A. 1 cm = 6 km

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?

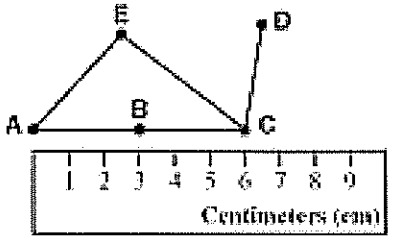
A. 2 days

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?

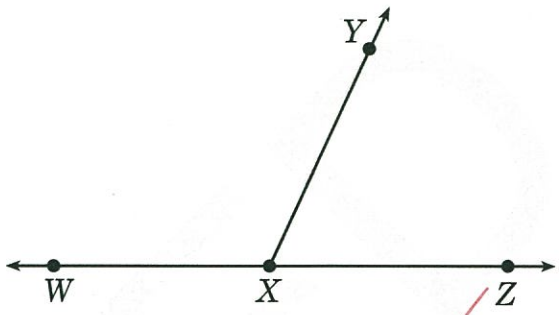
A $\frac{1}{12}$

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?

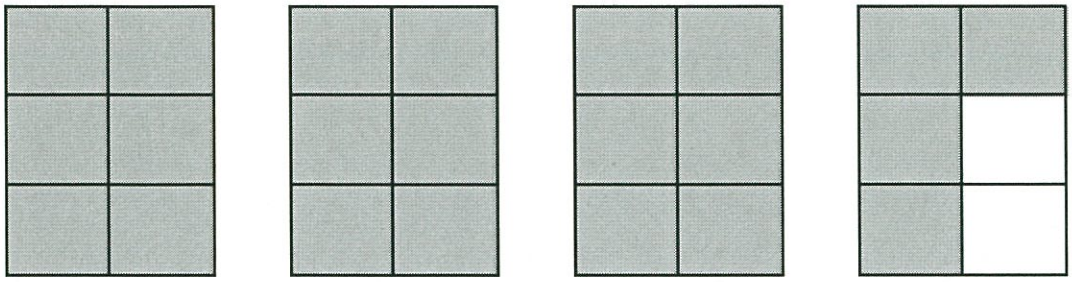
A between 50 and 60 pounds

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 $3\frac{4}{6}$

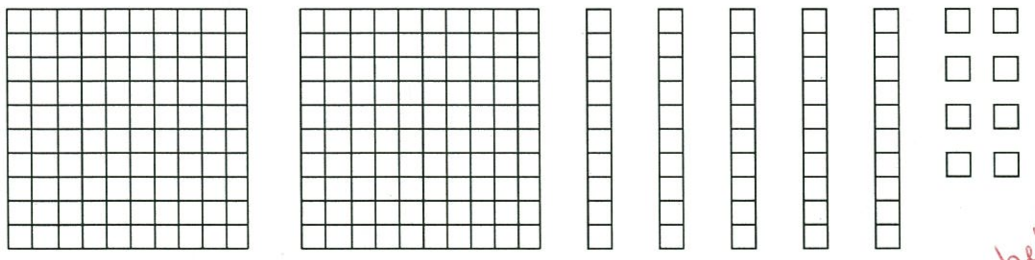
C $4\frac{2}{6}$

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

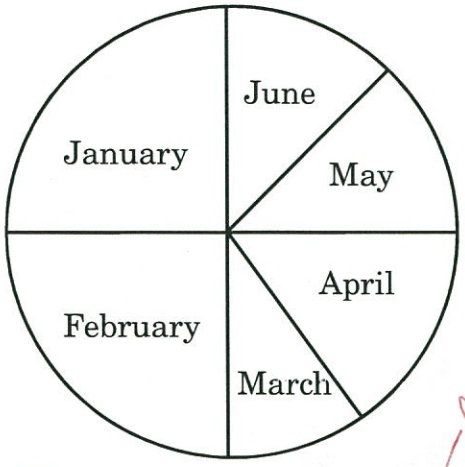
C 358

D 385

Is there better placement for the stem?

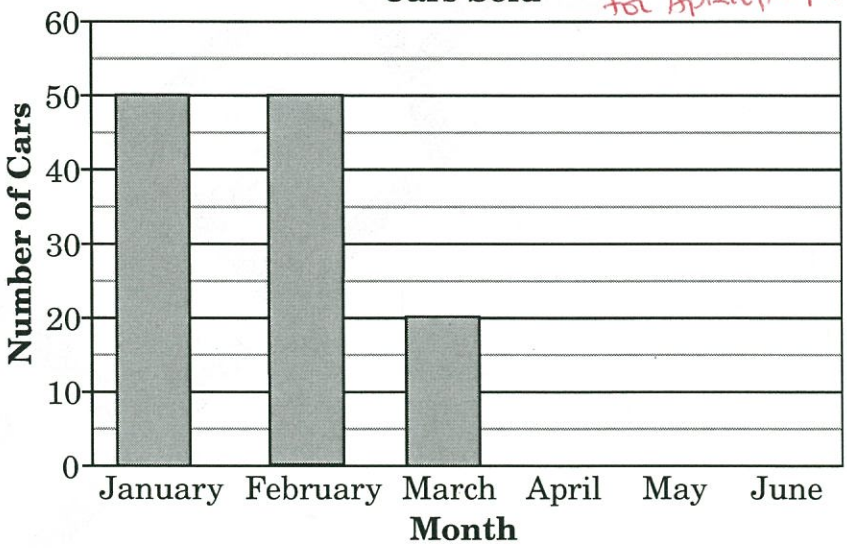
30. A dealership sold 200 cars in a six-month period. The circle graph below displays the distribution of sales by month.

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).

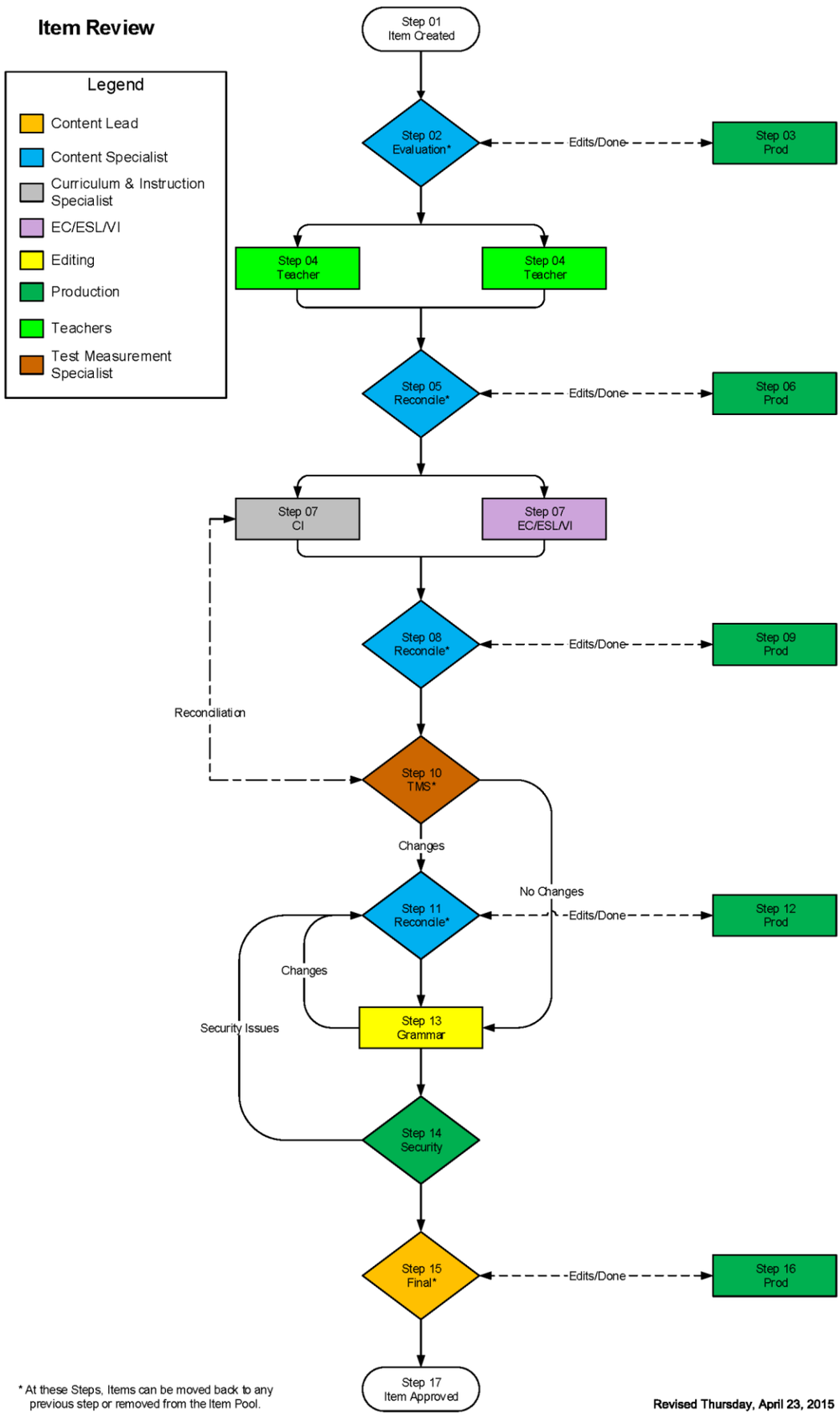
Step 14: Security Check
Production staff checks to make sure no duplicate copy of the item exists in the test development databases. If there is a duplicate copy of the item or a requested revision was not made, then the item is flagged and sent back to **Step 11**.

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.

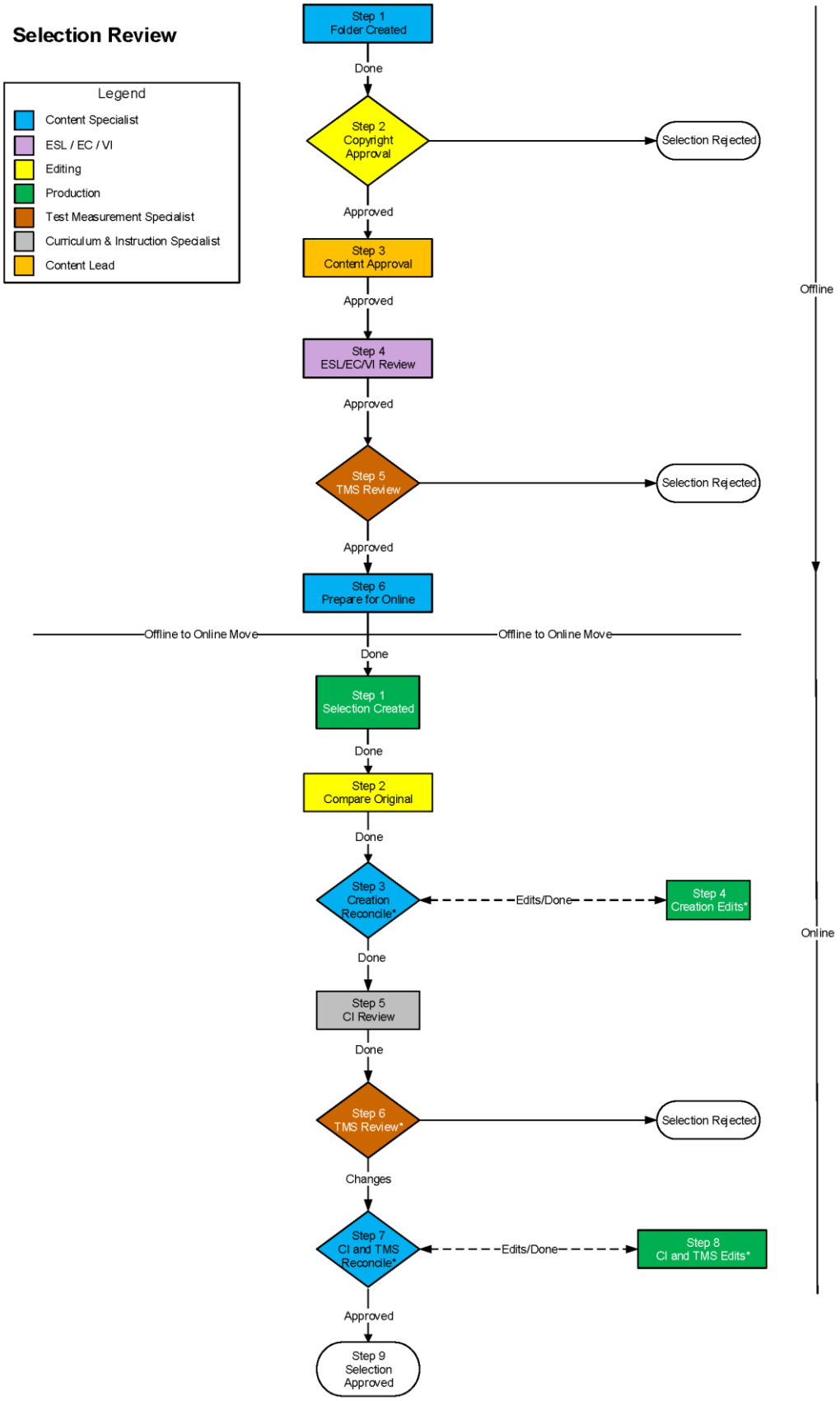
Step 7: Reconcile Curriculum and Instruction Review and Test and Measurement Specialist Review
A Content Specialist reviews any comments/changes requested by Curriculum and Instruction or by the Test and Measurement Specialist, and sends changes to **Step 8** (Production) to be made if necessary. Once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

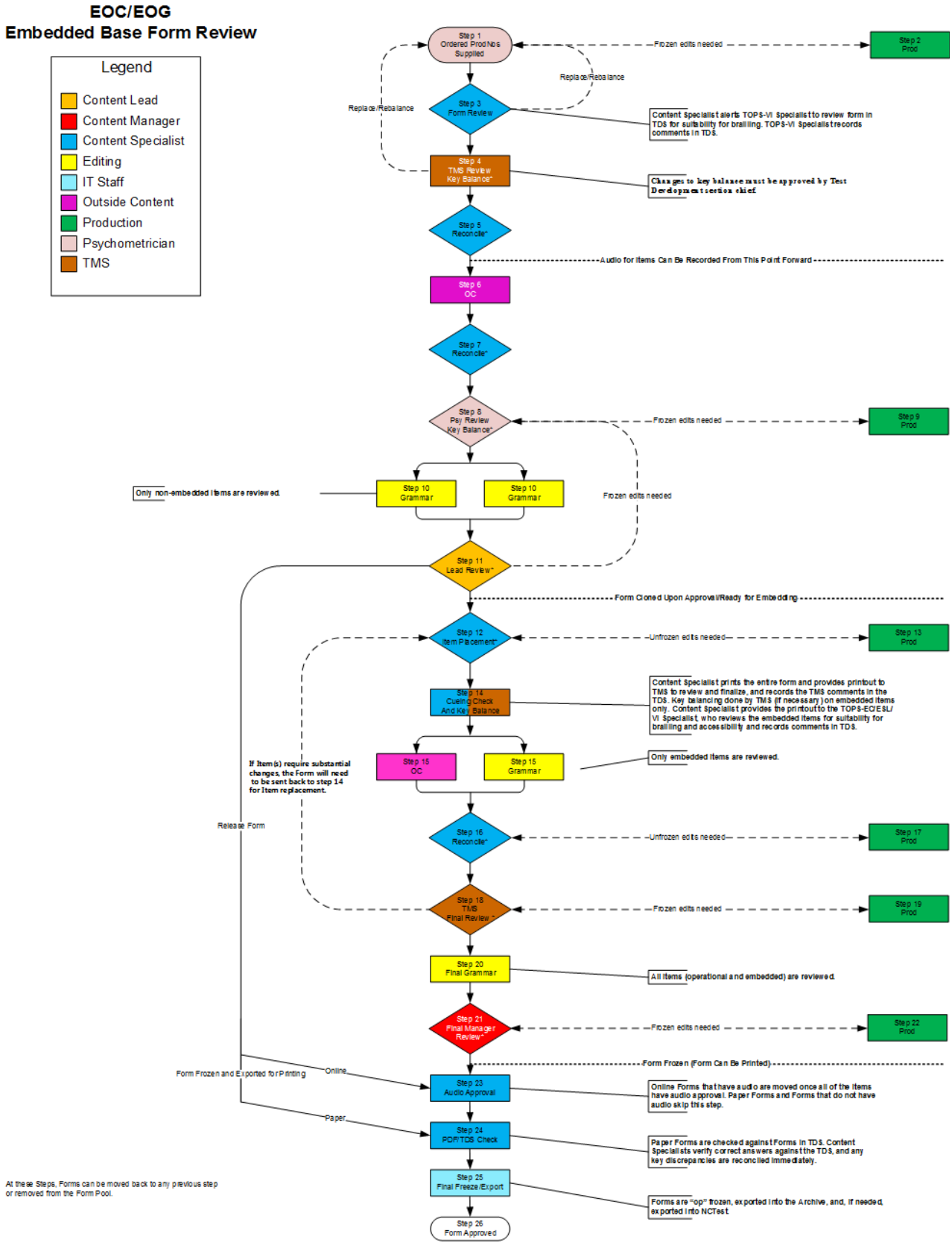
Step 23: Audio Approval
A Content Specialist reviews the audio for each item and either approves the audio or indicates it needs correction. After all items' audio have been approved, the form is sent to **Step 24** (PDF/Online Check).

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.

APPENDIX 4-C TIF & CSE PLOTS BASED ON FIELD TEST PARAMETERS-ELA

Figure 1 EOG ELA Grade 3 Test Information with associated Standard Errors

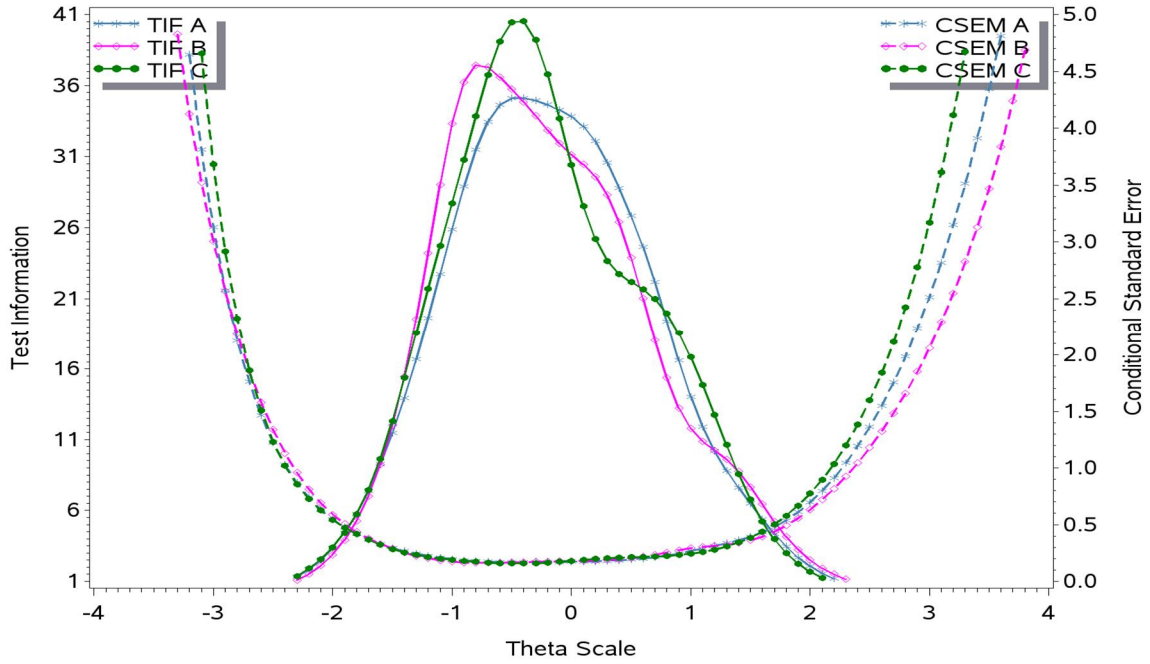


Figure 2 EOG *ELA Grade 4 Test Information with associated Standard Errors*

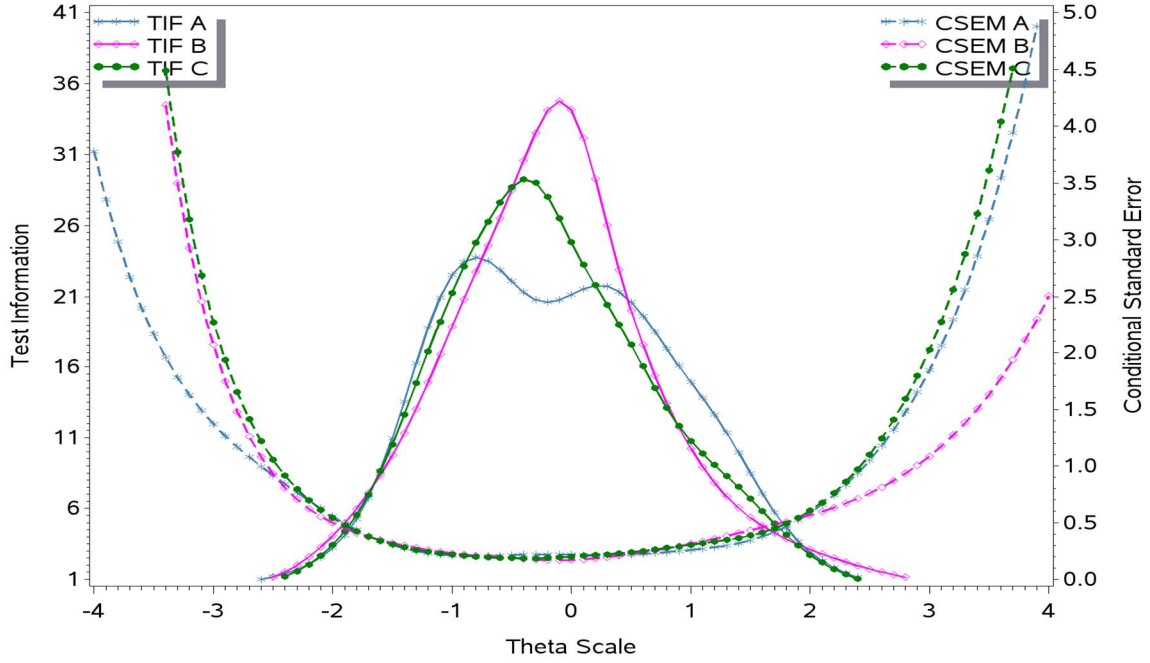


Figure 3 EOG *ELA Grade 5 Test Information with associated Standard Errors*

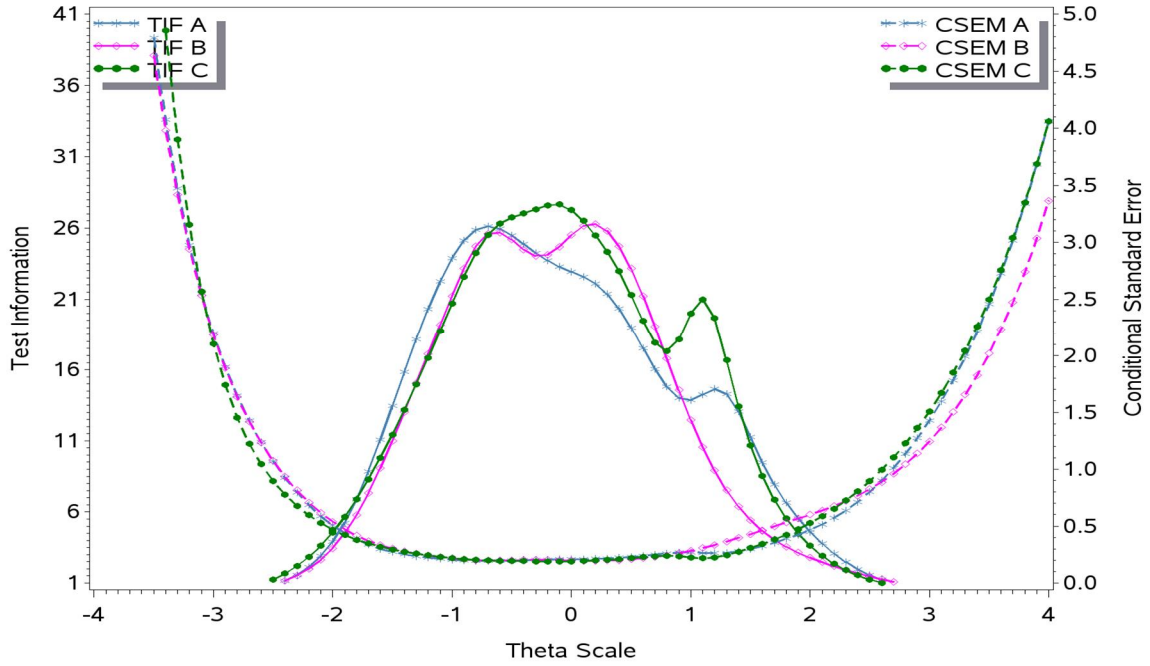


Figure 4 EOG *ELA Grade 6 Test Information with associated Standard Errors*

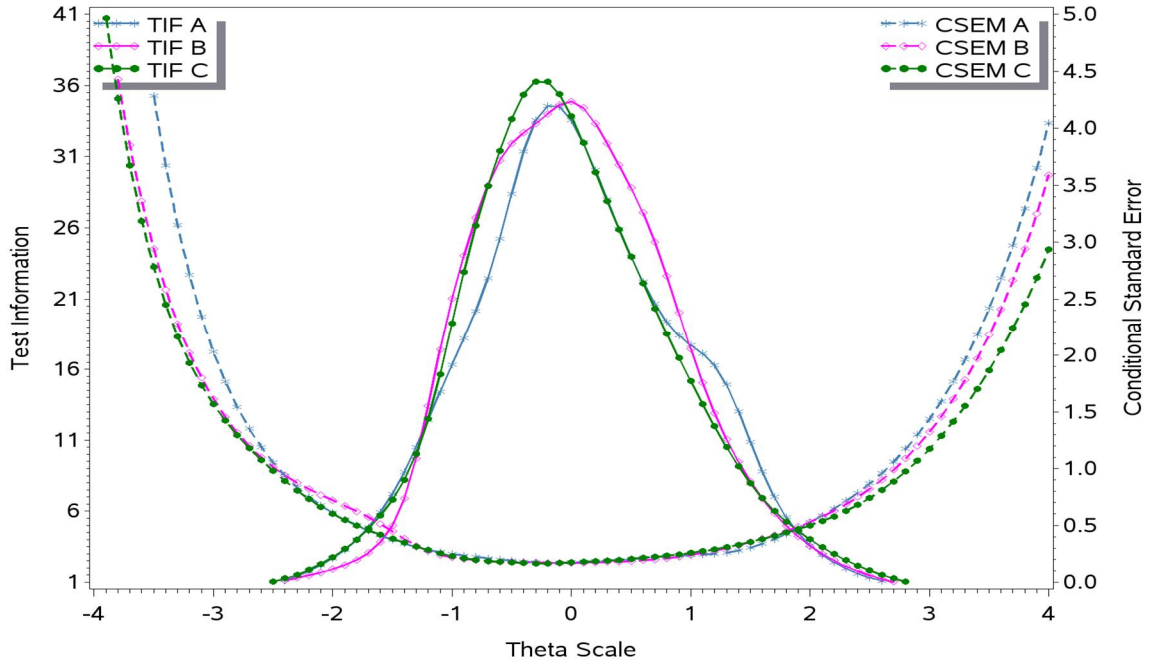


Figure 5 EOG *ELA Grade 7 Test Information with associated Standard Errors*

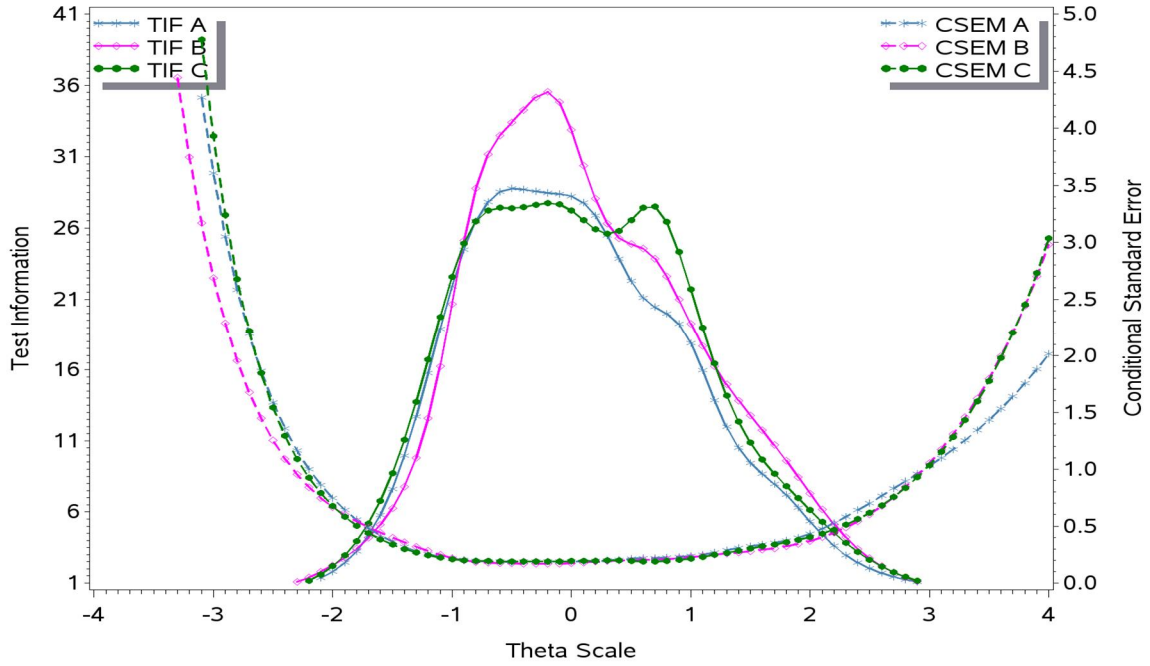


Figure 6 EOG *ELA Grade 8 Test Information with associated Standard Errors*

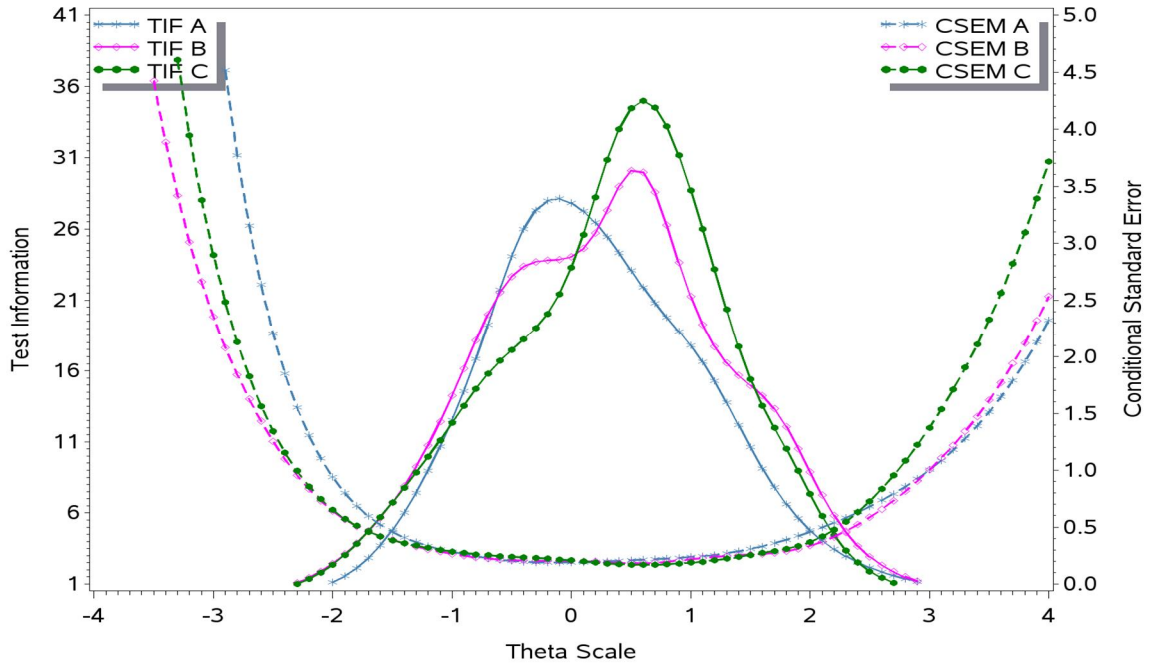
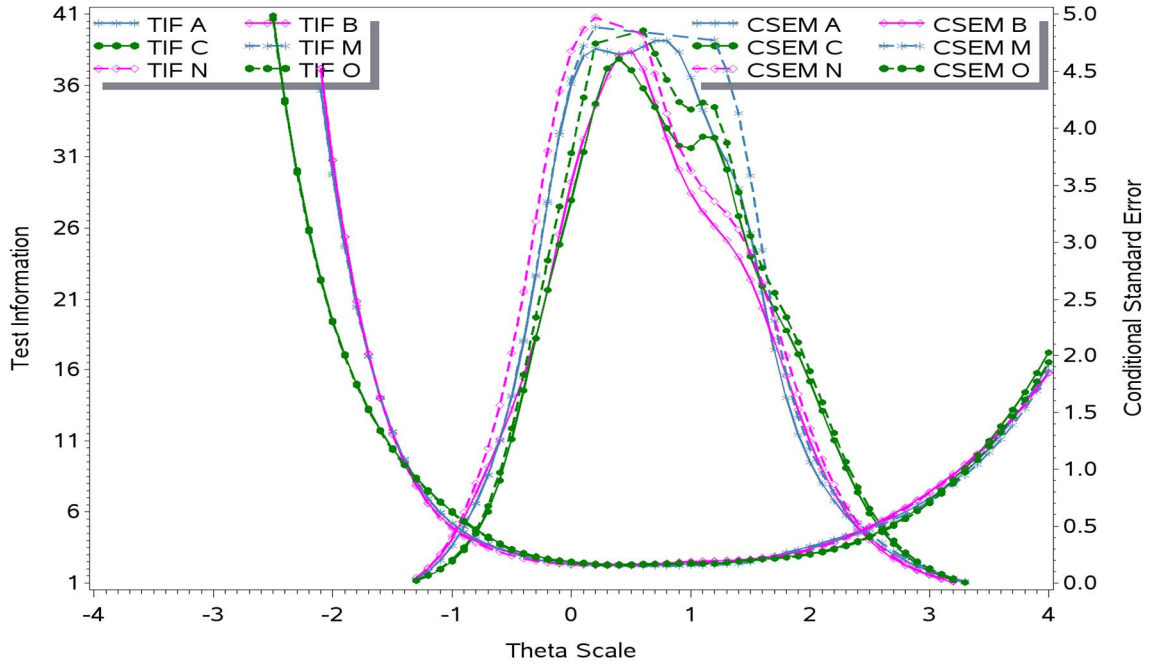


Figure 7 EOC English II *Test Information with associated Standard Errors*

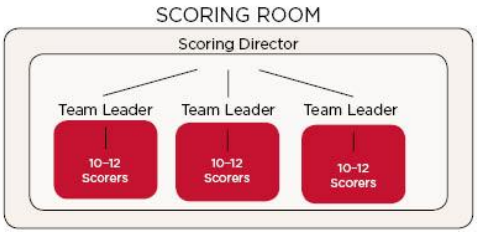


North Carolina Scoring Process: English II

Project Staffing

Questar uses a number of different models for scoring. For North Carolina we will be using a hierarchical structure of:

- Scoring Director
- Team Leaders
- Scorers



Scoring Directors are chosen by for a project based on the following qualifications:

- 4-year degree
- Content Expertise
- Previous project experience
- Experience with ScorePoint
- Ability to work under pressure to meet deadlines
- Ability to travel, facilitate, and interact with client
- Possesses good work ethic and integrity
- Good verbal and written communication skills
- Evaluations
- Schedule Flexibility

The Scoring Directors have the overall responsibility for the training of the project and content as well as the scoring expectations. They undergo extensive specialized training to prepare them for their roles as scoring experts and monitors, by working with QAI or department content specialists, as well as attending a workshop on managing a project and seasonal staff which includes:

Questar Philosophy

- Mission Statement and Core Values
- Expectations
- Roles and Responsibilities
    - HR Training
    - Scoring Director, Team Leader and Reader Training

Tools to Help with Success

- Outlook/Email
- ADP/Timeclock
- OpsPath
- Validity
- ScorePoint (our proprietary scoring engine)

Team Leaders report directly to the Scoring Directors and are typically in charge of a team of 10–12 scorers, depending on the item(s) and content area. They are specifically trained on the requirements and processes for scorer monitoring and intervention, including interpreting ScorePoint reports such as, Reader Reliability (RR) and Score Point Distribution (SPD) reports, conducting read behinds, holding one-on-one discussions, and scoring.

Team Leaders (TLs) are selected based on:

- 4-year degree
- Content knowledge
- Previous project experience
- Experience with ScorePoint
- Evaluations

Scorers must have fulfilled the following requirements:

- 4-year degree (in a related field in the content area for which they will be scoring as appropriate)
- Attend an open house for an introduction to Questar philosophy
- Complete an application process, complete with references
- Complete a sample of the content area for which they are applying
- Complete a one-on-one interview with Questar scoring staff

Training and Qualifying

Training Materials

Training materials for North Carolina include responses scored during rangefinding that represent the full range of score points as determined by the rangefinding committees, including responses that exemplify the nuances of the rubric (e.g., differentiation of a low "3" from a high "2"). All materials will be provided to the client before use in operational scorer training.

Training materials are organized by item and will consist of the following:

- One **Passage**
- One **Prompt and Rubric**
- One **Scoring Guide (or Guide Set) -** contains approximately 10 items with a minimum of 3 anchor responses (1 for each score point). During training, the Scoring Guide is discussed response by response within the group setting to identify any nuances of individual responses that have been selected as exemplary. This phase also includes a discussion of often seen acceptable and non-acceptable details for each item.
- A **Training Set -** contains 10 responses, representing a variety of score points in random order. The training set is scored independently by each scorer and each response is discussed by the group. This set is used as a learning tool to assess whether the scorer understands the nuances as discussed in the Scoring Guide.
- A **Qualifying Set –** contains 10 responses, representing a variety of score points in random order. The qualifying set is scored independently by each scorer and each response is discussed by the group. This set is used to determine whether a scorer is eligible to continue on to scoring. Meeting the qualification standards on this set demonstrates that the scorer will be able to apply the necessary skills to score.

Training the Team Leaders

Since effective scorer training relies largely on having knowledgeable, flexible Team Leaders, the Scoring Directors carefully select and train only the most qualified people to be Team Leaders. Our Team Leaders are trained prior to scorers, so they will be familiar with all of the training materials and the scoring procedures prior to scorer training.

Scorers will be divided into teams, and each scorer will be assigned a unique scorer identification number. That identification number allows for the tracking of scorer performance via the scorer quality control reports throughout the online scoring.

Once the training staff is confident that the scorers understand and have an awareness of the need to be sensitive to the performances of students, nondisclosure forms are signed and training begins. Scorers, like Team Leaders, must meet the requirements of the qualification standards before scoring student responses. Any scorer who is unable to meet the qualifying standards is dismissed, a stipulation understood by all scorers when they are hired. Since North Carolina does not require qualifying and this is an internal QAI process, it is possible to consider additional training in order to qualify.

Scorers who have been assigned to this project will be led by our experts through a rigorous training process which includes the following prior to actual scoring:

- Signing of a nondisclosure agreement
- Acknowledgement of the QAI harassment policy
- Review of the customer expectations and goals
- Scorers are reminded they must set aside any biases they may have about students, student work, and the scoring criteria presented
- Training to use the ScorePoint online scoring system

Once scorers have been instructed on the above, individual training begins with the following process:

- Training the Scoring Guide: this includes discussing the rubric, presenting the task or item (i.e., graphics and all related assets), reviewing the eligible score points, followed by group participation and discussion of each response using examples and annotations as appropriate. Questions by scorers are addressed as a group for consistent messaging and decisions.
- Scorers then complete a training set independently to assess their grasp of the scoring thus far.
- Each response in the training set is reviewed with the group with an explanation and examples as needed to ensure scorer consistency on the nuances of each response and score point.
- Scorers complete a qualifying set independently. Results using the qualification criteria will determine if they are allowed to score that particular task type.
- In addition, each nonscoreable code is explained and examples are provided as available.
- Protocol for “alerting” responses that require attention is discussed at this time.

Following the successful completion of training and qualifying, scoring center staff activate individual scorers in the system, allowing them to score student responses.

Qualification Standard

- 80% exact agreement on rubrics

Note: Exact agreement rates listed above for qualifying are the lowest acceptable percentages.

Scoring Rules

Nonscoreable codes are used to identify responses that cannot receive a numeric score based on the item's rubric (e.g., blank responses). Refer to the "Score_and_NonscoreCodes" business requirements document for scoring details. The following table shows the possible nonscoreable codes available for assignment:

	Blank	Illegible	Foreign Lang.	Repeating Prompt	Off Topic	Incoherent	Other Reason
READING	BL	IL	FL	RP	OT	IC	OR

In some cases, scorers will utilize functionality in Questar's ScorePoint system to submit responses that potentially require nonscoreable assignment to the TLs for review. A TL must either accept the nonscoreable code provided by the scorer or provide a numeric score or different nonscoreable code. If the TL does not accept the nonscoreable code provided by the scorer, the modified score/code is entered under the TL's scorer ID, and the response is discussed with the scorer as a retraining step.

Scoring Agreement Rate

For scoring there will be a 80% exact agreement rate for all rubrics. The exact agreement rates given for scoring are the lowest acceptable percentages. We expect the exact and adjacent agreement rates to be significantly higher during scoring. If scorers do not meet the qualification standards they will be prohibited from scoring that item. We require our reader reliability to meet the qualifying standard(s). If scorers are not meeting this, they may be removed from scoring that item or from the project altogether based on the scorer monitoring results as detailed below.

Alerted Responses

Scorers are directed to send up for review any student response that may suggest the possibility of teacher interference, plagiarism, or use of inappropriate content. Similarly, scorers are also instructed as to what kinds of things should trigger a review for disturbing content (e.g., possible physical or emotional abuse, suicidal ideation, threats of harm to themselves or others, etc.). When a scorer identifies a response that fits these criteria it is scored and then marked as an alert in the ScorePoint system. The scorer also selects the reason for the alert and includes any comments to explain the need for the alert.

All alerts are reviewed by the SD to ensure the responses are properly flagged. After the SD reviews all alerted responses, a file of those alerted responses is generated. Alert files will be sent to the required individuals weekly, or at any other interval requested by NCDPI.

Scorer Monitoring

Scorers are monitored daily throughout the scoring window by means of ScorePoint system reports, validity responses (when quantity of responses permits), ongoing training, one-on-one discussions, and read behinds. Monitoring activities are described in greater detail below.

Read Behinds

Read behinds is a term used in handscoring to describe the process through which TLs and/or SDs review assigned scores in order to confirm each scorer's ability to score accurately.

Questar uses two kinds of read behinds:

- **Random read behinds** are done throughout the day for all scorers, regardless of whether an issue or concern has been noted. Random read behinds are part of the ongoing monitoring process.
- **Prescribed read behinds** represent an increased number of read behinds due to some issue that may have come to the attention of the scoring leader through a ScorePoint system monitoring report, a comment or question from the scorer, or during a random read behind.

Once scoring begins, read behinds become an integral part of the scoring leaders' responsibilities for the duration of the scoring window.

Scores Changed in Read Behind

During a read behind, whether random or prescribed, a TL or SD may encounter a score that has been assigned by a scorer erroneously. It is during the read behind process that the incorrect score is changed, which then results in a series of actions taking place—the score is corrected, the response is discussed with the scorer one-on-one, and the number of read behinds is increased to ensure the scorer is scoring accurately based on the rubric.

There are several important notes regarding score changes:

- A score changed in read behind results in the new score from the TL or SD. The new score becomes the score of record.
- A score is changed only if there is no rubric justification for the score given by scorer 1. Borderline score changes or "preferential" score changes should not be made.
- For all task types: Should scorer reliability percentages fall below the proposed minimum exact agreement rate, or should a TL or SD have any concern about a scorer's scoring accuracy, prescribed read behinds will increase and appropriate actions will be taken. The scorer will not read unsupervised until the TL or SD is satisfied that he/she is scoring accurately.

Note: We typically use a 10% read-behind rate for each scorer. If the Team Leader or Scoring Director needs to increase the read behinds based on monitoring metrics, and modifications are required to scores, daily one-on-one discussions are required.

One-on-One Discussions

A one-on-one discussion may be held with a scorer in the context of a score changed in a read behind. A discussion may also take place to address questions or issues brought up by the scorer or as a training tool using specific exemplar responses from scoring to point out problems or scoring tendencies a scorer may exhibit.

Note: If one-on-one discussions are required and performance does not improve, the scorer would be removed from scoring that item based on qualitative and quantitative data.

Paired Scoring

Paired scoring is an effective tool that Questar utilizes for ongoing training and clarification to maintain and strengthen a scorer's understanding as they apply item specific rubrics. Paired scoring can be employed in a few ways:

- **Paired scoring at the beginning of the project**
  - Paired scoring helps to ensure consistent application of the rubric throughout the room and serves as a springboard for discussion.
  - Scorers are instructed to discuss the issue with another scorer before submitting a score if a disagreement or question arises. Any inconsistencies or misunderstandings are brought to light and addressed with the group.
  - Paired scoring is often used at the beginning of the project or after weekends. Scorers may be intentionally paired (e.g., experienced scorers with newer scorers) in order to discuss responses and talk through the rubric as a score is assigned.
- **Paired scoring as a group**
  - Scorers may be engaged in paired scoring with the SD as time allows at the beginning of scoring.
  - The responses will be projected on a screen or read aloud.
  - Paired scoring in this context allows the SD/TL to describe the rubric application in detail with student responses, and ensures the full group of scorers understands how to apply the rubric consistently and accurately beyond the examples in the training materials.

Recalibration Sets

Recalibration sets may be created from responses scored during rangefinding (if available) or during the field test scoring for use during operational scoring. The responses chosen are exemplar responses that will be instructive to the scorers. Sets can include 3–5 responses to exemplify the nuances of the rubric(s). Recalibration sets are typically used after an extended break from scoring.

Validity

Validity responses are pre-scored responses strategically interspersed in the pool of responses during operational scoring. These responses are not distinguishable to the scorer and scores are only accepted for monitoring purposes, not in replacement of the score of record. The use of validity responses is an objective process that helps ensure that scorers are applying the same standards throughout the project. This procedure offers feedback on the accuracy and consistency of individual scorers and groups of scorers assigned to a given item. The frequency of Questar's validity process can be adjusted as appropriate throughout scoring (e.g., initial scoring of item, weekend breaks, or clarifications on line responses).

Removal from Scoring an Item

A scorer may be removed from scoring an item in the event that they are not scoring correctly and consistently on one or more tasks. This is determined when interventions have not resulted in the required improvement based on our daily scorer monitoring.

Dismissal from Scoring the Project

A scorer will be dismissed if retraining does not elicit satisfactory results and it is determined that a scorer cannot accurately score student responses.

Note: Should a scorer be removed from scoring an item or dismissed from scoring the project, their scored responses would be reviewed and potentially rescored based on our monitoring process.

Monitoring Reports

Overview

Questar's ScorePoint system features a variety of system-generated reports on scoring metrics.

Reports can be filtered using different parameters to monitor scorers, such as by teams, individuals, and individual items. The Scoring Director (SD) uses these reports to monitor each team and the group as a whole to ensure consistent scoring across all teams. Team Leaders (TL) use these reports to closely monitor the scorers on their team, both in terms of productivity and reliability.

In addition to our internal monitoring efforts, we provide the Item Reliability and Score Point Distribution report described below.

Item Reliability and Score Point Distribution Report (IRSPD)

The IRSPD report displays the inter-rater reliability of the distribution of scores for each item for the project for the entire group. This report includes the number of responses scored, and can also be used to monitor production.

Content Area / Item / Domain	Responses Read			% Read	%	Agreement Rate %			Score Distribution %							
	Once	Twice	Third	Twice	Adjudicated	Exact	Adjacent	Non	0	1	2	3	4	5	6	Non Score

For North Carolina, 10% of the responses will receive two readings:

- Scorer 1 score will be the score of record.
- Scorer 2 score will be used to calculate scorer reliability.
- Responses are randomly chosen and redistributed throughout the day to be scored independently by a different scorer in the room (scorer 2).
- Scorers do not know if they are doing a first scoring or a second scoring.
- There will be resolution scorings.

Developmental Scale for North Carolina End-of-Grade/End-of-Course ELA/Reading and English II Tests, Fourth Edition

Alan Nicewander, Ph.D.
Tia Sukin, Ed.D.
Josh Goodman, Ph.D.
Huey Dodson, B.S.
Matthew Schulz, Ph.D.
Susan Lottridge, Ph.D.
Phoebe Winter, Ph.D.

Submitted to the
North Carolina Department of Education

December 2, 2013

Pacific Metrics Corporation
1 Lower Ragsdale Drive
Building 1, Suite 150
Monterey, CA 93940

Developmental Scale for North Carolina End-of-Grade/End-of-Course ELA/Reading and English II Tests, Fourth Edition

This technical report describes the methods used and results found by Pacific Metrics Corporation in deriving the developmental scale for the North Carolina End-of-Grade/End-of-Course ELA/Reading and English II Tests, Fourth Edition. To create the vertical scale, Pacific Metrics used the methods already in place by North Carolina Department of Public Instruction (NCDPI) as described in the *North Carolina Reading Comprehension Tests Technical Report* (Bazemore & Van Dyke, 2004). For the ELA/Reading and English II scale, Pacific Metrics used Appendix C (Thissen, Edwards, Coon, & Woods, 2004) of that report. The article by Williams, Pommerich, and Thissen (1998) was also used as a reference.

Grade levels included in the Fourth Edition developmental scale slightly differ from those included in the First through Third editions. While First through Third edition scales include grades Pre 3 through 8, the Fourth Edition scale includes grades 3 through 8. The corresponding End-of-Course assessment, English II, was also included in the initial scale, but was dropped due to a North Carolina team decision.

Fourth Edition Developmental Scale

Table 1 presents the Fourth Edition developmental scale for the population for ELA/Reading and English II. Grade 5 was the base grade for the developmental scale, using a mean of 450 and standard deviation of 10. To create the developmental scale, the same items (called a linking set) were administered to students in adjacent grades. Both above- and below-grade links were used for the ELA/Reading and English II scale. Items were operational when on-grade level but served as embedded (e.g., did not contribute toward student scores) when placed off-grade level.

Table 1. Developmental Scale Means and Standard Deviations Derived from Spring 2013 Item Calibration for North Carolina End-of-Grade/End-of-Course Tests of Reading Comprehension/English II, Fourth Edition

	Population	
Grade	Mean	Standard Deviation
3	440.01	10.90
4	446.00	10.33
5	450.00	10.00
6	452.70	10.99
7	455.97	11.12
8	458.66	11.35
English II	461.82	11.75

As shown in table 1 and as expected, the mean scores increased between grades, with growth ranging from 3 to 6 scale score points. The smallest increase occurred between grade 6 and grade 7; the largest increase occurred between grade 3 and grade 4.

The values for the developmental scales are based upon item response theory (IRT) estimates of differences between adjacent-grade mean thetas (θ) and ratios of adjacent-grade standard deviations of θ. The three-parameter logistic model was used to estimate item and person parameters. flexMIRT™ version 1.88 (Cai, 2012) was used. In flexMIRT™, the below grade was considered the reference group; its population mean and standard deviation were set to 0 and 1, respectively. The above-grade mean and standard deviation were estimated using the scored data and the IRT parameter estimates. These parameters were provided in the flexMIRT™ output and did not require independent calculation.

Individual runs in flexMIRT™ were conducted for each of the grade-pair links. For ELA/Reading, each grade pair for grades 3 through 8 had twelve links (six below-grade and six above-grade), and grade-pair 8–English II had thirty links (fifteen below-grade and fifteen above-grade). The linking sets varied between six and eight items, and each linking set was associated with a reading passage.

Under the assumption of equivalent groups, the form results were averaged within grade pairs to produce one set of values per adjacent grade. Outlying values were dropped if they were greater than two standard deviations from the mean. For ELA/Reading, three sets of values were dropped as outliers—one each from the 3–4, 6–7, and 7–8 grade pairs. Table 2 displays the average difference in adjacent-grade means and standard deviation ratios for Reading. Table 3 presents the mean difference and standard deviation ratio for each adjacent-grade link for Reading.

Table 2. Average Mean Difference in Standard Deviation Units of Lower Grade and Average Standard Deviation Ratios Derived from Spring 2013 Item Calibrations for North Carolina End-of-Grade/End-of-Course Tests of ELA/Reading and English II, Fourth Edition

Grades	Average Mean Difference	Average Standard Deviation Ratio	Number of Grade-Pair Forms
3–4*	0.550	0.948	11
4–5	0.387	0.968	12
5–6	0.270	1.099	12
6–7*	0.298	1.011	11
7–8*	0.242	1.021	11
8–English II	0.278	1.035	30

Note: An asterisk (*) denotes that one outlier was removed from the average for this grade pair.

Table 3. Values for Adjacent-grade Means in Standard Deviation (SD) Units of Lower Grade and Standard Deviation Ratios, Derived from Spring 2013 Item Calibrations for North Carolina End-of-Grade/End-of-Course Tests of ELA/Reading and English II, Fourth Edition

Grades 3–4		Grades 4–5		Grades 5–6		Grades 6–7		Grades 7–8		Grade 8–Eng II	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.375	1.103	0.318	1.271	0.070	1.185	0.203	1.083	0.084	1.219	0.444	1.265
0.596	0.859	0.388	0.717	0.386	1.002	0.426	0.884	0.211	1.037	0.354	1.017
0.515	0.906	0.403	0.937	0.188	1.096	0.319	1.102	0.231	1.030	0.107	1.375
0.608	1.130	0.427	1.000	0.262	1.119	0.266	1.064	0.155	1.310	0.460	1.002
0.480	1.065	0.294	0.887	0.235	1.350	0.243	1.116	0.155	1.043	0.532	0.853
0.519	0.928	0.365	0.919	0.282	0.974	0.289	0.805	0.328	1.005	0.269	1.020
0.682	0.774	0.535	0.755	0.421	0.858	0.391	0.720	0.303	0.797	0.583	0.922
0.588	0.950	0.498	0.987	0.355	0.953	0.411	1.021	0.193	1.113	0.643	0.696
0.561	0.908	0.308	1.095	0.300	1.160	0.257	1.040	0.363	0.995	–0.036	1.429
0.533	0.878	0.457	0.831	0.329	1.117	0.323	0.912	0.376	1.005	–0.133	1.245
0.506	1.016	0.346	1.038	0.303	1.153	0.277	1.043	0.264	0.949	0.400	1.163
0.465	1.014	0.302	1.176	0.103	1.221	0.268	1.056	0.151	1.034	0.292	0.868
										0.383	1.019
										0.310	0.968
										–0.025	1.346
										0.477	0.829
										0.441	0.822
										0.090	0.846
										0.426	0.953
										0.552	0.726
										–0.150	1.368
										–0.093	1.227
										0.182	1.229
										0.268	0.818
										0.227	1.070
										0.487	0.866
										0.216	1.086
										0.411	0.788
										0.119	1.277
										0.115	0.969

Note: Means and standard deviations in shaded cells were dropped from analyses as outliers.

Comparison of Fourth Edition Developmental Scale to First through Third Edition Scales

Table 4 presents the mean scale scores by grade for the First, Second, Third, and Fourth editions for ELA/Reading and English II. To facilitate comparison of the growth between grades among the First, Second, Third, and Fourth editions, figure 1 presents the mean scores plotted together for ELA/Reading and English II. To place the First, Second, Third, and Fourth edition scores on similar scales, a value of 300 was added to the First Edition scores, a value of 200 was added to the Second Edition scores, and a value of 100 was added to the Third Edition scores.

For ELA/Reading and English II, greater average growth between grades 3–8 occurred in the Third Edition (19.72) than in the First, Second, and Fourth editions (13.96, 14.14, and 18.65, respectively). As shown in figure 1, the First through Fourth editions exhibited similar growth in mean scores between grades 3–8.

Table 4. Comparison of Population Means and Standard Deviations for First through Fourth Editions of North Carolina End-of-Grade/End-of-Course Tests of ELA/Reading and English II

	First Edition (1992)		Second Edition (2002)		Third Edition (2008)		Fourth Edition (2013)	
Grade	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Pre 3	139.02	8.00	236.66	11.03	326.56	14.64		
3	145.59	9.62	245.21	10.15	338.62	12.57	440.01	10.90
4	149.98	9.50	250.00	10.00	345.20	10.79	446.00	10.33
5	154.74	8.21	253.92	9.61	350.00	10.00	450.00	10.00
6	154.08	9.44	255.57	10.41	352.86	10.12	452.70	10.99
7	157.81	9.09	256.74	10.96	355.63	9.79	455.97	11.12
8	159.55	8.96	259.35	11.13	358.34	9.49	458.66	11.35
English II							461.82	11.75

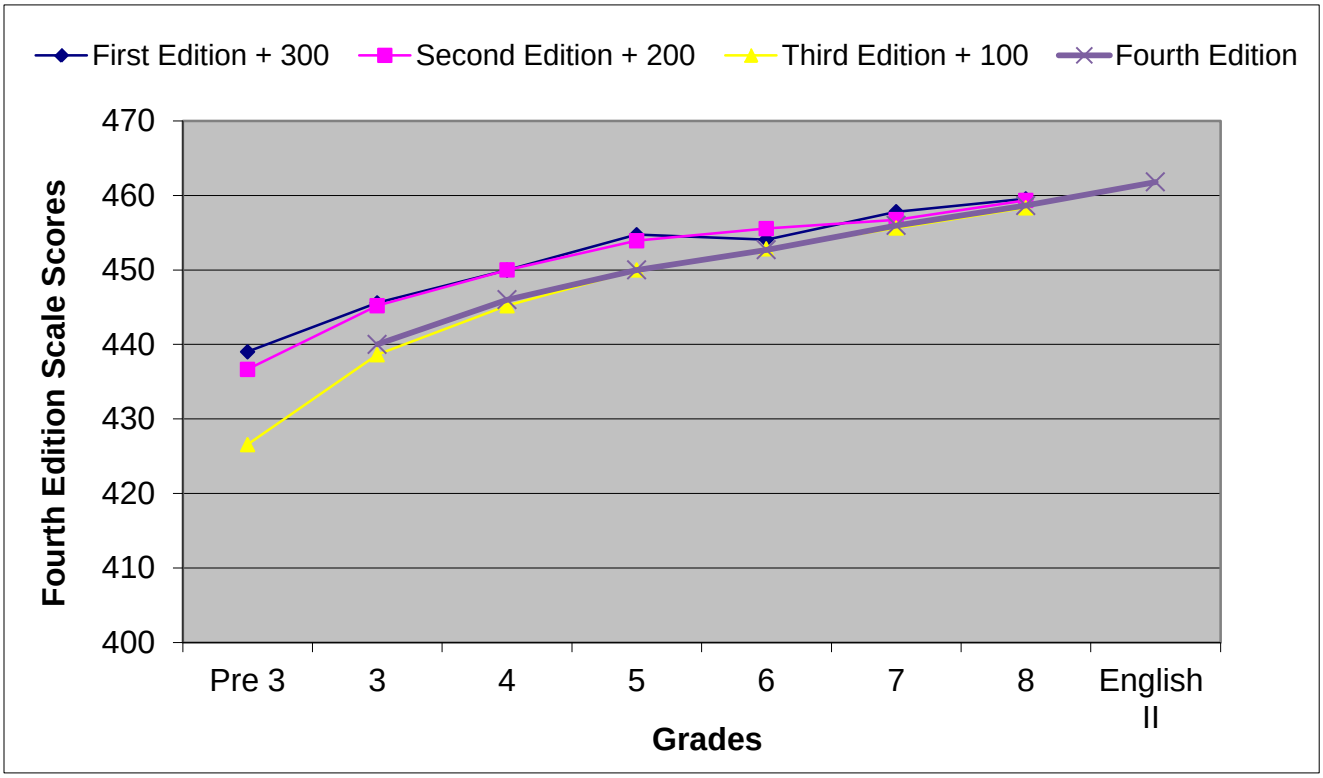


Figure 1. Comparison of Growth Curves between First, Second, Third, and Fourth Editions of North Carolina End-of-Grade/End-of-Course Tests of ELA/Reading and English II.

Quality Assurance Procedures

The authors have applied a variety of analyses and procedures to ensure that the results of the scaling and linking studies are correct. For the vertical scale, the mean difference and standard deviation ratios for the grades and forms were compared to the classical test theory *p*-values of the linking items. The data provided evidence that the mean difference and standard deviation ratios were accurate in both direction and magnitude (see table 5). Also, previous work using the described statistical method to create the vertical scale was applied to the Second Edition data to ensure that it reproduced the scale correctly.

Table 5. Average Mean Difference in Standard Deviation Units of Lower Grade and Standard Deviation Ratios, and Average Difference in *p*-values (Higher Minus Lower Grade) of Linking Sets, for North Carolina End-of-Grade/End-of-Course Tests of ELA/Reading and English II, Fourth Edition

Grade Pair	Average Mean Difference	Mean *p*-value Difference for Linking Items
3–4*	0.550	0.097
4–5	0.387	0.068
5–6	0.270	0.044
6–7*	0.298	0.049
7–8*	0.242	0.046
8–English II	0.278	0.050

Note: An asterisk (*) denotes that one grade-pair link was dropped from analyses as an outlier.

Additionally, IRT parameters provided separately by the North Carolina Department of Education were correlated with the flexMIRT™ calibrated item parameters within grade pairs and averaged across grades. For Reading, the average correlation for discrimination parameters was 0.97 with a standard deviation of 0.01 across grade and form pairs. The average correlation for difficulty or step parameters (for English II multi-point items) was 0.97 with a standard deviation of 0.02. The average correlation for guessing parameters was 0.93 with a standard deviation of 0.02.

Psychometrics Underlying the Developmental Scale

The procedure for creating the developmental scale is based upon that described in Williams, Pommerich, and Thissen (1998). The procedure is divided into four steps, described below.

Step 1. flexMIRT™ was used to calibrate the End-of-Grade and End-of-Course Reading tests' item and population parameters for adjacent grades. This process was described in the section entitled "Fourth Edition Developmental Scale" of this report and resulted in average mean difference and average standard deviation ratios (m_n and s_n) for each grade n (see table 2).

Step 2. A (0,1) growth scale anchored at grade 3 was constructed to yield the following means (M_n) and standard deviations (S_n):

$M_n = M_{n-1} + m_n S_{n-1}$, mean for Grade n on (0,1) growth scale anchored at the lowest grade (with grade 3 indexed as n=3),

$S_n = s_n S_{n-1}$, standard deviation for grade n on (0,1) growth scale anchored at the lowest grade (with grade 3 indexed as n=3),

where $M_2 \equiv 0$, and $S_2 \equiv 1$. This (0,1) growth scale was generated recursively upwards to the End-of-Course (English II).

Step 3. The scale was re-centered (re-anchored) at grade 5, yielding

$$M^*_n = \frac{(M_n - M_5)}{S_5}$$
$$S^*_n = \frac{S_n}{S_5}$$

as the means (M^*_n) and standard deviations (S^*_n).

Step 4. The final step in constructing the growth scale was the application of a linear transformation in order to produce a growth scale with the grade 5 mean and standard deviations equal to 450 and 10, respectively, *viz.*,

$$\mu_n = 450 + 10 M^*_n$$
$$\sigma_n = 10 S^*_n \ ,$$

where μ_n is the mean of the final growth scale in grade n and σ_n is the standard deviation for the growth scale in grade n.

References

Bazemore, M., & Van Dyke, P. (2004). *North Carolina Reading Comprehension Tests Technical Report.* Raleigh, NC: North Carolina Department of Public Instruction.

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.

Thissen, D., Edwards, M., Coon, C. & Woods, C. (2002). *North Carolina Reading Comprehension Tests Technical Report, Appendix C.* Raleigh, NC: North Carolina Department of Public Instruction.

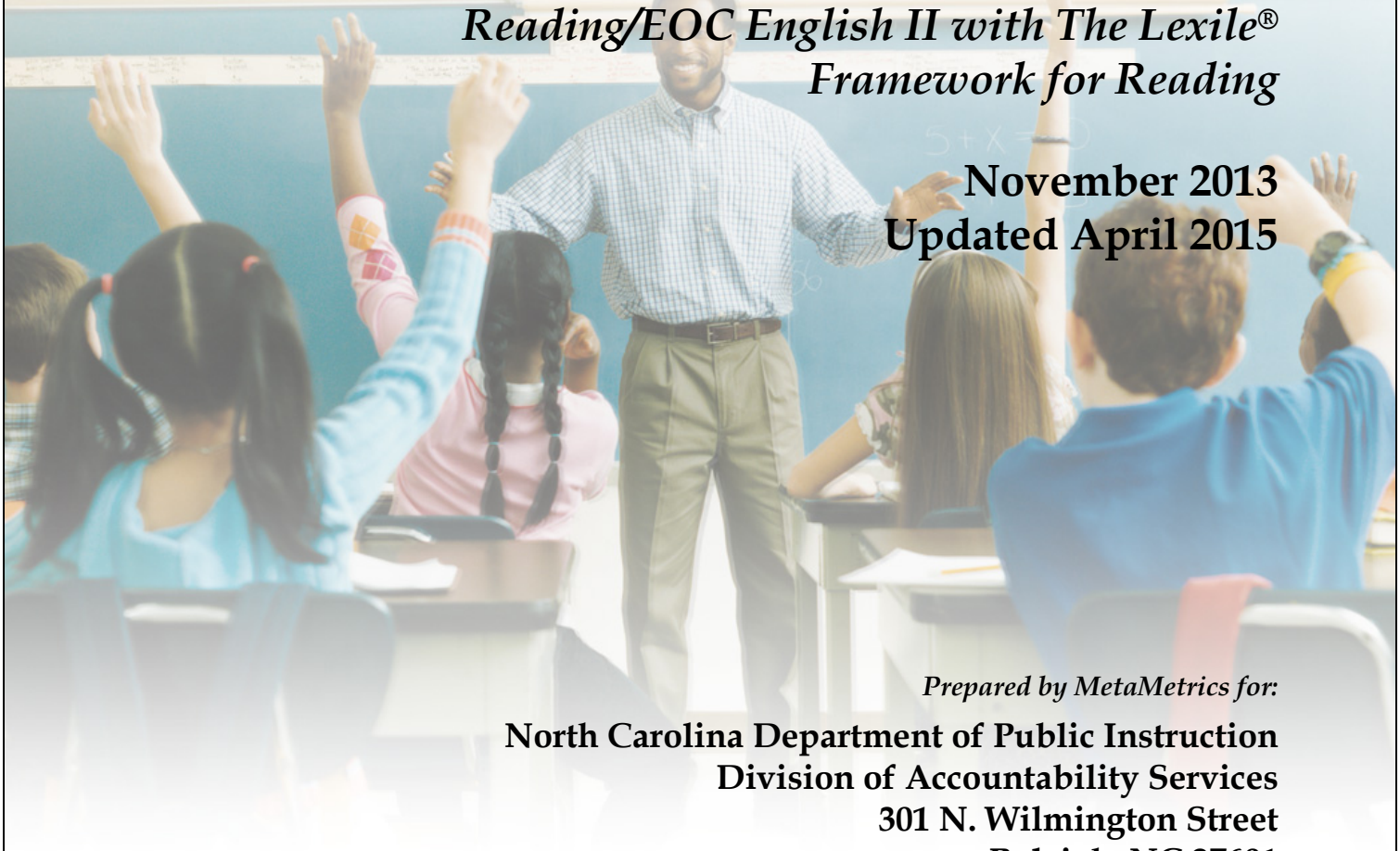
Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based upon Thurstone methods and item response theory. *Journal of Educational Measurement, 35,* 93-107.

APPENDIX 10-A LEXILE LINKING TECHNICAL REPORT

Linking the NC READY EOG Reading/EOC English II with the Lexile® Framework

A Study to Link the North Carolina READY EOG Reading/EOC English II with The Lexile® Framework for Reading

November 2013
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.Lexile.com

Preface
Lexile Scale Enhancements

The Lexile® Framework for Reading is a scientific approach to measuring reading ability and the complexity of reading materials. The Lexile Framework includes a Lexile measure and the Lexile scale. A Lexile measure represents both the complexity of a text, such as a book or article, and an individual's reading ability. Lexile measures are expressed as numeric measures followed by an "L" (e.g., 850L), and are placed on the Lexile scale. (There is no space between the measure and the "L.") The Lexile scale is a developmental scale for reporting reader ability and text complexity, ranging from below 200L for emergent readers and emergent-reader texts to above 1600L for advanced readers and texts. Lexile measures of one thousand or greater are reported without a comma (e.g., 1050L). All Lexile reader measures should be rounded to the nearest 5L to avoid over-interpretation of the measures. As with any test score, uncertainty in the form of measurement error is present. If the Lexile reader measure is xxx2.5 or higher or xxx7.5 or higher, it is rounded up to the next highest 5L; below those points, the measure is rounded down to the next lowest 5L. For example, if a computed Lexile reader measure is 772.51, it should be reported as 775L. If the computed Lexile reader measure is 777.42, is should be reported as 775L.

Prior to May 1, 2014, all Lexile reader measures at or below 0L were reported as BR (Beginning Reader). Starting in spring 2014, Lexile reader measures below 0L may be reported with a more specific measure. These BR measures are shown as "BRxxxL." For example, a Lexile reader measure of -150 is reported as BR150L where "BR" stands for "Beginning Reader" and replaces the negative sign in the number. The Lexile scale is like a thermometer, with numbers below zero indicating decreasing reading ability as the number moves away from zero. The smaller the number following the BR code, the more advanced the reader is. For example, a BR150L reader is more advanced than a BR200L reader. Above 0L, measures indicate increasing reading ability as the numbers increase. For example, a 200L reader is more advanced than a 150L reader.

Lexile 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. A computed Lexile measure of 772.51 would be represented as 773L. If the purpose is instructional, then the Lexile measures should be capped at the upper bound of measurement error (e.g., at the 95th percentile point of the national Lexile norms) to ensure developmental appropriateness of the material. MetaMetrics expresses these measures used for instructional purposes as "Reported Lexile Measures" and recommends that they be used on individual score reports. In instructional environments where the purpose of the Lexile measure is to

appropriately match readers with text, all scores below 0L should be reported as "BRxxxL." No student should receive a negative Lexile measure on a score report. The lowest reported value below 0L is BR400L.

Table i. Maximum reported Lexile measures by grade.

Grade	Lexile Caps
	850L
1	900L
2	1100L
3	1200L
4	1300L
5	1400L
6	1500L
7	1600L
8	1700L
9	1725L
10	1750L
11	1800L
12	1825L

Some assessments report a Lexile range for each student rather than a specific Lexile reader measure. The Lexile range is 50L above to 100L below the student's actual Lexile measure. For example, the Lexile range for a specific reader measure of 700L is 600L to 750L. This range represents the boundaries between relatively easy reading material for the student and the level at which the student will be more challenged, yet can still read successfully.

Text within the Technical Report has been updated to correspond with the language of the enhanced Lexile 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. Two examples arise in large-scale assessment. In one situation, one test can provide a unique type of information (such as national comparisons available from NAEP) but is not administered very often. At the same time another test is administered more often, but is not able to provide the breadth of information (such as a state assessment). An auxiliary score scale for a test can be established to provide this additional information through assessment scale linkages. Once linkages are established between the two assessments, then the results of the more-frequently-administered assessment can be translated in terms of the scale for the other assessment.

In another situation, the linkage between two score scales can be used to provide a context for understanding the results of one of the assessments. For example, sometimes it is hard to explain what a student can read based on the results of a reading comprehension test. Parents typically ask the questions "If my child is in the fourth grade and scores 450 on the NC READY EOG Reading assessment, what does this mean?" or "Based on my child's test results, what can he or she read and how well?" or "Is my child well prepared to meet the reading demands of grade level materials?" Once a linkage is established with an assessment that is related to specific book or text titles, then the results of the assessment can be explained and interpreted in the context of the specific titles that a student should be able to read.

Auxiliary score scales can be used to "convey additional normative information, test-content information, and information that is jointly normative and content based. For many test uses, an auxiliary scale conveys information that is more crucial than the information conveyed by the primary score scale. In such instances, the auxiliary score is the one that is focused on, and the primary scale can be viewed more as a vehicle for maintaining interpretability over time" (Petersen, Kolen, and Hoover, 1989, p. 222). One such auxiliary scale is The Lexile® Framework for Reading, which was developed to appropriately match readers with text at a level that provides challenge but not frustration.

Linking assessment results with the Lexile Framework provides a mechanism for matching each student's reading ability with text on a common scale. It serves as an anchor to which texts and assessments can be connected allowing parents, teachers, and administrators to speak the same language. In addition, the Lexile Framework provides a common way to monitor if students are "on track" for the reading demands of various postsecondary endeavors. By using the Lexile Framework, the same metric is applied to

the books students read, the tests they take, and the results that are reported. Parents often ask questions like the following:

- How can I help my child become a better reader?
- How do I challenge my child to read so that she is ready for various college and career options?

Questions like these can be challenging for parents and educators. By linking the NC READY EOG Reading/EOC English II assessment with The Lexile Framework for Reading, educators and parents will be able to answer these questions and will be better able to use the results from the test to improve instruction and to develop each student's level of reading comprehension.

This research study was designed to determine a mechanism to provide reading levels that can be matched to text based on the NC READY EOG Reading/EOC English II test scores. 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:

- present a solution for matching readers with text;
- provide North Carolina with Lexile measures on the NC READY EOG Reading/EOC English II assessment;
- develop tables for converting NC READY EOG Reading/EOC English II scale scores to Lexile measures; and
- produce a report that describes the linking analysis procedures.

The Lexile Framework for Reading

All symbol systems share two features: a semantic component and a syntactic component. In language, the semantic units are words. Words are organized according to rules of syntax into thought units and sentences (Carver, 1974). In all cases, the semantic units vary in familiarity and the syntactic structures vary in complexity. The comprehensibility or difficulty of a message is dominated by the familiarity of the semantic units and by the complexity of the syntactic structures used in constructing the message.

The Semantic Component

As far as the semantic component is concerned, it is clear that most operationalizations are proxies for the probability that an individual will encounter a word in a familiar context and thus be able to infer its meaning (Bormuth, 1966). This is the basis of exposure theory, which explains the way receptive or hearing vocabulary develops (Miller and Gildea, 1987; Stenner, Smith, and Burdick, 1983). Klare (1963) hypothesized that the semantic component varied along a familiarity-to-rarity continuum. This concept was further developed by Carroll, Davies, and Richman (1971), whose word-frequency study examined the reoccurrence of words in a five-million-word corpus of running text. Knowing the frequency of words as they are used in written and oral communication provided the best means of inferring the likelihood that a word would be encountered by a reader and thus become a part of that individual's receptive vocabulary.

Variables such as the average number of letters or syllables per word have been observed to be proxies for word frequency. There is a strong negative correlation between the length of words and the frequency of word usage. Polysyllabic words are used less frequently than monosyllabic words, making word length a good proxy for the likelihood that an individual will be exposed to a word.

In a study examining receptive vocabulary, Stenner, Smith, and Burdick (1983) analyzed more than 50 semantic variables in order to identify those elements that contributed to the difficulty of the 350 vocabulary items on Forms L and M of the *Peabody Picture Vocabulary Test*—Revised (Dunn and Dunn, 1981). Variables included part of speech, number of letters, number of syllables, the modal grade at which the word appeared in school materials, content classification of the word, the frequency of the word from two different word counts, and various algebraic transformations of these measures.

The first word frequency measure used was the raw count of how often a given word appeared in a corpus of 5,088,721 words sampled from a broad range of school materials (Carroll, Davies, and Richman, 1971). For example, the word "accident"

appears 176 times in the 5,088,721-word corpus. The second word frequency measure used was the frequency of the "word family." A word family included: (1) the stimulus word; (2) all plurals (adding "-s" or "-es" or changing "-y" to "-ies"); (3) adverbial forms; (4) comparatives and superlatives; (5) verb forms ("-s," "-d," "-ed," and "-ing"); (6) past participles; and (7) adjective forms. For example, the word family for "accident" would include "accidental," "accidentally," "accidentals," and "accidents," and they would all have the same word frequency of 334. The frequency of a word family was based on the sum of the individual word frequencies from each of the types listed.

Correlations were computed between algebraic transformations of these means (mean frequency of the words in the test item and mean frequency of the word families in the test item) and the rank order of the test items. Since the items were ordered according to increasing difficulty, the rank order was used as the observed item difficulty. The log of the mean word frequency provided the strongest correlation with item rank order (r = -0.779) for the items on the combined form.

The Lexile Framework currently employs a 600-million-word corpus when examining the semantic component of text. This corpus was assembled from the more than 15,000 texts that were measured by MetaMetrics for publishers from 1998 through 2002. When text is analyzed by MetaMetrics, all electronic files are initially edited according to established guidelines used with the Lexile Analyzer software. These guidelines include the removal of all incomplete sentences, chapter titles, and paragraph headings; running of a spell check; and re-punctuating where necessary to correspond to how the book would be read by a child (for example, at the end of a page). The text is then submitted to the Lexile Analyzer that examines the lengths of the sentences and the frequencies of the words and reports a Lexile measure for the book. When enough additional texts have been analyzed to make an adjustment to the corpus necessary and desirable, a linking study will be conducted to adjust the calibration equation such that the Lexile measure of a text based on the current corpus will be equivalent to the Lexile measure based on the new corpus.

The Syntactic Component

Klare (1963) provides a possible interpretation for how sentence length works in predicting passage difficulty. He speculated that the syntactic component varied with the load placed on short-term memory. Crain and Shankweiler (1988), Shankweiler and Crain (1986), and Liberman, Mann, Shankweiler, and Westelman (1982) have also supported this explanation. The work of these individuals has provided evidence that sentence length is a good proxy for the demand that structural complexity places upon verbal short-term memory.

While sentence length has been shown to be a powerful proxy for the syntactic complexity of a passage, an important caveat is that sentence length is not the underlying causal influence (Chall, 1988). Researchers sometimes incorrectly assume that manipulation of sentence length will have a predictable effect on passage difficulty. Davidson and Kantor (1982), for example, illustrated rather clearly that sentence length can be reduced and difficulty increased and vice versa.

Based on previous research, it was decided to use sentence length as a proxy for the syntactic component of reading difficulty in the Lexile Framework.

Calibration of Text Difficulty

The research study on semantic units (Stenner, Smith, and Burdick, 1983) was extended to examine the relationship of word frequency and sentence length to reading comprehension. In 1987(a), Stenner, Smith, Horabin, and Smith performed exploratory regression analyses to test the explanatory power of these variables. This analysis involved calculating the mean word frequency and the log of the mean sentence length for each of the 66 reading comprehension passages on the *Peabody Individual Achievement Test* (Dunn and Markwardt, 1970). The observed difficulty of each passage was the mean difficulty of the items associated with the passage (provided by the publisher) converted to the logit scale. A regression analysis based on the word-frequency and sentence-length measures produced a regression equation that explained most of the variance found in the set of reading comprehension tasks. The resulting correlation between the observed logit difficulties and the theoretical calibrations was 0.97 after correction for range restriction and measurement error. The regression equation was further refined based on its use in predicting the observed difficulty of the reading comprehension passages on 8 other standardized tests. The resulting correlation between the observed logit difficulties and the theoretical calibrations across the 9 tests was 0.93 after correction for range restriction and measurement error.

Once a regression equation is established linking the syntactic and semantic features of text to the difficulty of text, the equation can be used to calibrate test items and text.

The Lexile Scale

In developing the Lexile Scale, the Rasch model (Wright and Stone, 1979) was used to estimate the difficulties of the items and the abilities of the 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 group 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 group. If two different items are administered to the second group, 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 a scale with a fixed zero was to identify two anchor points for the scale. The following criteria were used to select the two anchor points: they should be intuitive, easily reproduced, and widely recognized. For example, with most thermometers the anchor points are the freezing and boiling points of water. For the Lexile Scale, the anchor points are text from seven basal primers for the low end and text from *The Electronic Encyclopedia* (Grolier, Inc., 1986) for the high end. These points correspond to the middle of first grade text and the midpoint of workplace text.

The next step was to determine the unit size for the scale. For the Celsius thermometer, the unit size (a degree) is $1/100^{th}$ of the difference between freezing (0 degrees) and boiling (100 degrees) water. For the Lexile Scale the unit size (a Lexile) was defined as $1/1000^{th}$ of the difference between the mean difficulty of the primer material and the mean difficulty of the encyclopedia samples. Therefore, a Lexile by definition equals $1/1000^{th}$ of the difference between the difficulty of the primers and the difficulty of the encyclopedia.

The third step was to assign a value to the lower anchor point. The low-end anchor on the Lexile Scale was assigned a value of 200.

Finally, a linear equation of the form

$$[(\text{Logit} + \text{constant}) \times \text{CF}] + 200 = \text{Lexile text measure} \qquad \text{Equation (1)}$$

was developed to convert logit difficulties to Lexile calibrations. The values of the conversion factor (CF) and the constant were determined by substituting in the low-end anchor point and then solving the system of equations.

The Lexile Scale ranges from below 200L to above 1600L. There is a not an explicit bottom or top to the scale, but rather two anchor points on the scale (described above) that describe different levels of reading comprehension. The Lexile Map, a graphic representation of the Lexile Scale from 200L to 1500L+, provides a context for understanding reading comprehension.

Validity of The Lexile Framework for Reading

Validity refers to the "degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). In other words, does the test measure what it is supposed to measure? For the Lexile Framework, which measures a skill, the most important aspect of validity that should be examined is construct validity. The validity of the Lexile Framework can be evaluated by examining how well Lexile measures relate to other measures of reading comprehension and text difficulty.

Lexile Framework and other Measures of Reading Comprehension. *Table 1* presents the results from studies where students were administered a Lexile assessment and another assessment of reading comprehension. There is a strong relationship between reading comprehension ability as measured by the Lexile Framework and reading comprehension ability as measured by other assessments.

Table 1. Results from linking studies conducted with The Lexile Framework for Reading.

Standardized Test	Grades in Study	N	Correlation Between Test Score and Lexile Measure
Gates-MacGinitie Reading Test	2, 4, 6, 8, 10	4,644	0.90
Metropolitan Achievement Test (8th ed.)	2, 4, 6, 8, 10	2,382	0.93
Texas Assessment of nowledge and Skills (TA S)	3, 5, 8	1,960	0.60 to 0.73
The Iowa Tests (Iowa Tests of Basic Skills and Iowa Tests of Educational Development)	3, 5, 7, 9, and 11	4,666	0.88
Stanford Achievement Test (Tenth Edition)	2, 4, 6, 8, and 10	3,064	0.93
Oregon Reading/Literature nowledge and Skills Test	3, 5, 8, and 10	3,180	0.89
Mississippi Curriculum Test	2, 4, 6, and 8	7,045	0.90
Georgia Criterion Referenced Competency Test (CRCT and GHSGT)	1 8, and 11	16,363	0.72 to 0.88
Wyoming Performance Assessment for Wyoming Students (PAWS)	3, 5, 7, and 11	3,871	0.91
Arizona Instrument to Measure Progress (AIMS)	3, 5, 7, and 10	7,735	0.89
South Carolina Palmetto Achievement Challenge Tests (PACT)	3 8	15,559	0.87 to 0.88
Comprehensive Testing Program (CPT 4 ERB)	2, 4, 6, and 8	924	0.83 to 0.88
Oklahoma Core Competency Tests (OCCT)	3 8	10,691	0.71 to 0.75
TOEFL iBT	NA	2,906	0.63 to 0.67
TOEIC	NA	2,799	0.73 to 0.74
entucky Performance Rating for Educational Progress (-PREP)	3 8	6,480	0.71 to 0.79
North Carolina ACT	11	3,472	0.84
North Carolina READY End-of-Grades/End-of-Course Tests (NC READY EOG/EOC)	3, 5, 7, 8, and E2	12,356	0.88 to 0.89

Notes: Results are based on final samples used with each linking study.
*Not vertically equated; separate linking equations were derived for each grade.

Lexile Framework and the Difficulty of Basal Readers. In a study conducted by Stenner, Smith, Horabin, and Smith (1987b) Lexile calibrations were obtained for units in 11 basal series. It was presumed that each basal series was sequenced by difficulty. So, for example, the latter portion of a third-grade reader is presumably more difficult than the first portion of the same book. Likewise, a fourth-grade reader is presumed to be more difficult than a third-grade reader. Observed difficulties for each unit in a basal series were estimated by the rank order of the unit in the series. Thus, the first unit in the first book of the first grade was assigned a rank order of one and the last unit of the eighth-grade reader was assigned the highest rank order number.

Correlations were computed between the rank order and the Lexile calibration of each unit in each series. After correction for range restriction and measurement error, the average disattenuated correlation between the Lexile calibration of text comprehensibility and the rank order of the basal units was 0.995 (see *Table 2*).

Table 2. Correlations between theory-based calibrations produced by the Lexile equation and rank order of unit in basal readers.

Basal Series	Number of Units	r_{OT}	R_{OT}	R'_{OT}
Ginn Rainbow Series (1985)	53	.93	.98	1.00
HBJ Eagle Series (1983)	70	.93	.98	1.00
Scott Foresman Focus Series (1985)	92	.84	.99	1.00
Riverside Reading Series (1986)	67	.87	.97	1.00
Houghton-Mifflin Reading Series (1983)	33	.88	.96	.99
Economy Reading Series (1986)	67	.86	.96	.99
Scott Foresman American Tradition (1987)	88	.85	.97	.99
HBJ Odyssey Series (1986)	38	.79	.97	.99
Holt Basic Reading Series (1986)	54	.87	.96	.98
Houghton-Mifflin Reading Series (1986)	46	.81	.95	.98
Open Court Headway Program (1985)	52	.54	.94	.97
Total/Means	660	.839	.965	.995

r_{OT} = raw correlation between observed difficulties (O) and theory-based calibrations (*T*).

R_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction.

R'_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction and measurement error.

*Mean correlations are the weighted averages of the respective correlations.

Based on the consistency of the results in *Table 2*, the Lexile theory was able to account for the unit rank ordering of the 11 basal series even with numerous differences in the series—prose selections, developmental range addressed, types of prose introduced (i.e., narrative versus expository), and purported skills and objectives emphasized.

Lexile Framework and the Difficulty of Reading Test Items. In a study conducted by Stenner, Smith, Horabin, and Smith (1987a), 1,780 reading comprehension test items appearing on nine nationally-normed tests were analyzed. The study correlated empirical item difficulties provided by the publishers with the Lexile calibrations specified by the computer analysis of the text of each item. The empirical difficulties were obtained in one of three ways. Three of the tests included observed logit difficulties from either a Rasch or three-parameter analysis (e.g., NAEP). For four of the tests, logit difficulties were estimated from item *p*-values and raw score means and standard deviations (Poznanski, 1990; Wright, and Linacre, 1994). Two of the tests provided no item parameters, but in each case items were ordered on the test in terms of difficulty (e.g., PIAT). For these two tests, the empirical difficulties were approximated by the difficulty rank order of the items. In those cases where multiple questions were asked about a single passage, empirical item difficulties were averaged to yield a single observed difficulty for the passage.

Once theory-specified calibrations and empirical item difficulties were computed, the two arrays were correlated and plotted separately for each test. The plots were checked for unusual residual distributions and curvature, and it was discovered that the Lexile equation did not fit poetry items or noncontinuous prose items (e.g., recipes, menus, or shopping lists). This indicated that the universe to which the Lexile equation could be generalized was limited to continuous prose. The poetry and noncontinuous prose items were removed and correlations were recalculated. *Table 3* contains the results of this analysis.

Table 3. Correlations between theory-based calibrations produced by the Lexile equation and empirical item difficulties.

Test	Number of uestions	Number of Passages	Mean	SD	Range	Min	Max	r_{OT}	R_{OT}	R_{OT}
SRA	235	46	644	353	1303	33	1336	.95	.97	1.00
CAT-E	418	74	789	258	1339	212	1551	.91	.95	.98
Lexile	262	262	771	463	1910	304	1606	.93	.95	.97
PIAT	66	66	939	451	1515	242	1757	.93	.94	.97
CAT-C	253	43	744	238	810	314	1124	.83	.93	.96
CTBS	246	50	703	271	1133	173	1306	.74	.92	.95
NAEP	189	70	833	263	1162	169	1331	.65	.92	.94
Battery	26	26	491	560	2186	702	1484	.88	.84	.87
Mastery	85	85	593	488	2135	586	1549	.74	.75	.77
Total/ Mean	1780	722	767	343	1441	50	1491	.84	.91	.93

r_{OT} = raw correlation between observed difficulties (O) and theory-based calibrations (*T*).

R_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction.

R'_{OT} = correlation between observed difficulties (O) and theory-based calibrations (*T*) corrected for range restriction and measurement error.

*Means are computed on Fisher Ƶ transformed correlations.

The last three columns in *Table 3* show the raw correlation between observed (O) item difficulties and theoretical (T) item calibrations, with the correlations corrected for restriction in range and measurement error. The Fisher Ƶ mean of the raw correlations (r_{OT}) is 0.84. When corrections are made for range restriction and measurement error, the Fisher Ƶ mean disattenuated correlation between theory-based calibration and empirical difficulty in an unrestricted group of reading comprehension items (R'_{OT}) is 0.93. These results show that most attempts to measure reading comprehension, no matter what the item form, type of skill objectives assessed, or response requirement used, measure a common comprehension factor specified by the Lexile theory.

Text Measure Error Associated with the Lexile Framework

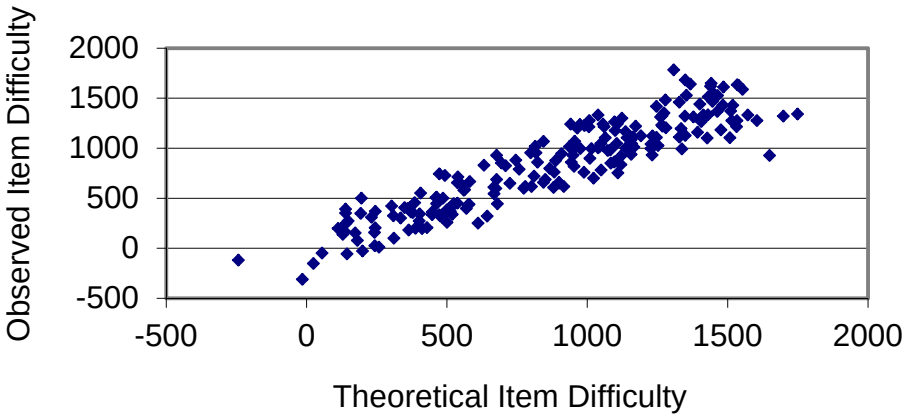
To determine a Lexile measure for a text, the standard procedure is to process the entire text. All pages in the work are concatenated into an electronic file that is processed by a software package called the Lexile Analyzer (developed by MetaMetrics, Inc.). The analyzer "slices" the text file into as many 125-word passages as possible, analyzes the set of slices, and then calibrates each slice in terms of the logit metric. That set of calibrations is then processed to determine the Lexile measure corresponding to a 75% comprehension rate. The analyzer uses the slice calibrations as test item calibrations and

then solves for the measure corresponding to a raw score of 75% (e.g., 30 out of 40 correct, as if the slices were test items). The Lexile Analyzer automates this process, but what "certainty" can be attached to each text measure?

Using the bootstrap procedure to examine error due to the text samples, the above analysis could be repeated (Efron, 1981; Sitter, 1992). The result would be an identical text measure to the first because there is no sampling error when a complete text is calibrated.

There is, however, another source of error that increases the uncertainty about where a text is located on the Lexile Map. The Lexile Theory is imperfect in its calibration of the difficulty of individual text slices. To examine this source of error, 200 items that had been previously calibrated and shown to fit the model were administered to 3,026 students in Grades 2 through 12 in a large urban school district. For each item the observed item difficulty calibrated from the Rasch model was compared with the theoretical item difficulty calibrated from the regression equation used to calibrate texts. A scatter plot of the data is presented in *Figure 1*.

Figure 1. Scatter plot between observed item difficulty and theoretical item difficulty.



The correlation between the observed and the theoretical calibrations for the 200 items was 0.92 and the root mean square error was 178L. Therefore, for an individual slice of text the measurement error is 178L.

The standard error of measurement associated with a text is a function of the error associated with one slice of text (178L) and the number of slices that are calibrated from a text. Very short books have larger uncertainties than longer books. A book with only four slices would have an uncertainty of 89L whereas a longer book such as *War and Peace* (4,082 slices of text) would only have an uncertainty of 3L (*Table 4*).

Table 4. Standard errors for selected values of the length of the text.

Title	Number of Slices	Text Measure	Standard Error of Text
The Stories Julian Tells	46	520	26
Bunnicula	102	710	18
The Pizza Mystery	137	620	15
Meditations of First Philosophy	206	1720	12
Metaphysics of Morals	209	1620	12
Adventures of Pinocchio	294	780	10
Red Badge of Courage	348	900	10
Scarlet Letter	597	1420	7
Pride and Prejudice	904	1100	6
Decameron	2431	1510	4
War and Peace	4082	1200	3

A typical Grade 3 reading test has approximately 2,000 words in the passages. To calibrate this text, it would be sliced into 16 125-word passages. The error associated with this text measure would be 45L. A typical Grade 7 reading test has approximately 3,000 words in the passages and the error associated with the text measure would be 36L. A typical Grade 10 reading test has approximately 4,000 words in the passages and the error associated with the text measure would be 30L.

The Find A Book (www.Lexile.com) contains information about each book analyzed: author, Lexile measure and Lexile Code, awards, ISBN, and developmental level as determined by the publisher. Information concerning the length of a book and the extent of illustrations—factors that affect a reader's perception of the difficultly of a book—can be obtained from MetaMetrics.

Lexile Item Bank

The Lexile Item Bank contains over 10,000 items that have been developed since 1986 for research purposes with the Lexile Framework.

Passage Selection. Passages selected for use are selected from "real world" reading materials that students may encounter both in and out of the classroom. Sources include textbooks, literature, and periodicals from a variety of interest areas and material written by authors of different backgrounds. The following criteria are used to select passages:

- the passage must develop one main idea or contain one complete piece of information;
- understanding of the passage is independent of the information that comes before or after the passage in the source text; and

- understanding of the passage is independent of prior knowledge not contained in the passage.

With the aid of a computer program, item writers examine blocks of text (minimum of three sentences) that are calibrated to be within 100L of the source text. From these blocks of text item writers are asked to select four to five that could be developed as items. If it is necessary to shorten or lengthen the passage in order to meet the criteria for passage selection, the item writer can immediately recalibrate the text to ensure that it is still targeted within 100L of the complete text (source targeting).

Item Format. The native Lexile item format is embedded completion. The embedded completion format is similar to the fill-in-the-blank format. When properly written, this format directly assesses the reader's ability to draw inferences and establish logical connections between the ideas in the passage (Haladyna, 1994). The reader is presented with a passage of approximately 30 to 150 words in length. The passages are shorter for beginning readers and longer for more advanced readers. The passage is then response illustrated (a statement is added at the end of the passage with a missing word or phrase followed by four options). From the four presented options, the reader is asked to select the "best" option that completes the statement. With this format, all options are semantically and syntactically appropriate completions of the sentence, but one option is unambiguously the "best" option when considered in the context of the passage.

The statement portion of the embedded completion item can assess a variety of skills related to reading comprehension: paraphrase information in the passage, draw a logical conclusion based on the information in the passage, make an inference, identify a supporting detail, or make a generalization based on the information in the passage. The statement is written to ensure that by reading and comprehending the passage the reader is able to select the correct option. When the embedded completion statement is read by itself, each of the four options is plausible.

Item Writer Training. Item writers are classroom teachers and other educators who have had experience with the everyday reading ability of students at various levels. The use of individuals with these types of experiences helps to ensure that the items are valid measures of reading comprehension. Item writers are provided with training materials concerning the embedded completion item format and guidelines for selecting passages, developing statements, and selecting options. The item writing materials also contain incorrect items that illustrate the criteria used to evaluate items and corrections based on those criteria. The final phase of item writer training is a short practice session with three items.

Item writers are provided vocabulary lists to use during statement and option development. The vocabulary lists were compiled from spelling books one grade level below the level the item would typically be used with. The rationale was that these

words should be part of a reader's "working" vocabulary since they had been learned the previous year.

Item writers are also given extensive training related to "sensitivity" issues. Part of the item writing materials address these issues and identify areas to avoid when selecting passages and developing items. The following areas are covered: violence and crime, depressing situations/death, offensive language, drugs/alcohol/tobacco, sex/attraction, race/ethnicity, class, gender, religion, supernatural/magic, parent/family, politics, animals/environment, and brand names/junk food. These materials were developed based on material published by McGraw-Hill (*Guidelines for Bias-Free Publishing*, 1983). This publication discusses the equal treatment of the sexes, fair representation of minority groups, and the fair representation of disabled individuals.

Item Review. All items are subjected to a two-stage review process. First, items are reviewed and edited by an editor according to the 19 criteria identified in the item writing materials and for sensitivity issues. Approximately 25% of the items developed are deleted for various reasons. Where possible items are edited and maintained in the item bank.

Items are then reviewed and edited by a group of specialists that represent various perspectives—test developers, editors, and curriculum specialists. These individuals examine each item for sensitivity issues and for the quality of the response options. During the second stage of the item review process, items are either "approved as presented," "approved with edits," or "deleted." Approximately 10% of the items written are "approved with edits" or "deleted" at this stage. When necessary, item writers receive additional on-going feedback and training.

Item Analyses. As part of the linking studies and research studies conducted by MetaMetrics, items in the Lexile Item Bank are evaluated in terms of difficulty (relationship between logit [observed Lexile measure] and theoretical Lexile measure), internal consistency (point-biserial correlation), and bias (ethnicity and gender where possible). Where necessary, items are deleted from the item bank or revised and recalibrated.

During the spring of 1999, 8 levels of a Lexile assessment were administered in a large urban school district to students in grades 1 through 12. The 8 test levels were administered in grades 1, 2, 3, 4, 5, 6, 7-8, and 9-12 and ranged from 40 to 70 items depending on the grade level. A total of 427 items were administered across the 8 test levels. Each item was answered by at least 9,000 students (the number of students per level ranged from 9,286 in grade 2 to 19,056 in grades 9-12). The item responses were submitted to a Winsteps IRT analysis. The resulting item difficulties (in logits) were

assigned Lexile measures by multiplying by 180 and anchoring each set of items to the mean theoretical difficulty of the items on the form.

The NC READY EOG Reading/EOC English II—Lexile Framework Linking Process

Description of the Assessments

North Carolina READY End-of-Grade Language Arts/Reading Assessments and End-of-Course English II Assessment. The 2013 North Carolina READY End-of-Grade Language Arts/Reading Assessments and End-of-Course English II Assessment are designed to measure students' proficiency on the Common Core State Standards (CCSS) for English Language Arts, adopted by the North Carolina State Board of Education in June 2010 (NCDPI, 2013d, 2013e). The Common Core State Standards are divided into strands which address a specific set of College and Career Readiness Anchor Standards. These strands are reading, writing, speaking, listening, and language.

The EOG assessments are administered annually to students in Grades 3 through 8 and the English II assessment is administered to students enrolled in English II (generally Grade 10) at the end of the course. 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 EOG English Language Arts/Reading assessments at Grades 3 through 8 are multiple-choice tests. These assessments are available only in paper-and pencil format for the 2012–13 school year. Students read authentic selections and then answer questions related to the selections. The reading selections are comprised of literary and informational text based on the *Common Core State Standards*. Knowledge of vocabulary is assessed indirectly through application and understanding of terms within the context of the selection and questions. The EOG assessments of English Language Arts/Reading at Grades 3 through 5 contain 52 total test items. The assessments at Grades 6 through 8 contain 56 total test items (NCDPI, 2013e).

The NC READY EOG Reading assessments were vertically scaled across grades. Each test has scale scores that range from 400 to 500. These scale scores can be compared directly from grade-to-grade.

The NC READY EOC English II assessment addresses a common set of standards for the second-year high school course of English language arts (NCDPI, 2013c). The English II assessment consists of reading passages and associated items addressing three strands of the CCSS: Reading, Language and Writing. The reading strand is further divided into two sub-strands of Reading Literature and Reading Information. The NC READY tests are approximately 30-35% Reading Literature, 35-40% Reading Information, 15-20% Language, and 15-20% Writing. The Speaking and Listening strands of the CCSS are not included in the assessment (NCDPI, 2013c).

The English II assessment is a criterion-referenced test (CRT) consisting of 50 operational four-response-option multiple-choice items and 3 operational constructed-response items. The constructed-response items appear throughout the test, integrated with multiple choice items related to text passages. The EOC English II scale scores range from 100 and 200, and these scale scores are on a separate scale.

The Lexile Framework for Reading. The Lexile Framework is a tool that can help teachers, parents, and students locate appropriate reading materials. Text complexity (difficulty) and reader ability are measured in the same unit—the Lexile. Text complexity is determined by examining such characteristics as word frequency and sentence length. Items and text are calibrated using the Rasch model. The typical range of the Lexile Scale is from 200L to 1600L, although actual Lexile measures can range from below zero (BR) to above 1600L (see the discussion on pages 5-6 for more information).

Using multiple-choice items, the Lexile Framework measures reading ability by focusing on skills readers use when studying written materials sampled from various content areas. Each test item consists of a passage that is response-illustrated (a statement is added at the end of the passage with a missing word or phrase followed by four options, or distractors). The skills measured by these items include referring to details in the passage, drawing conclusions, and making comparisons and generalizations. Lexile items do not require prior knowledge of ideas outside of the passage, vocabulary taken out of context, or formal logic.

The Lexile Linking Tests were developed for administration to students in Grades 3, 5, 7, 8, and English II. Characteristics of the Lexile Linking Tests were as similar as possible to the NC READY EOG Reading/EOC English II assessments, including the number of operational items per test and difficulty of the items. For each grade/course, two equivalent forms were developed and administered.

The Lexile Linking Tests contained 44 items on each test form for Grades 3 and 5, and 48 items on each test form for Grades 7 and 8. The number of items on the test for each grade was determined by the number of items on the NC READY EOG Reading/EOC English II assessments. Approximately 80% (35 for Grades 3 and 5, and 38 for Grades 7 and 8) of the items were common across the two grade-level test forms.

The English II Lexile Linking Test contained 56 items. The NC READY EOC English II assessment contains 50 operational multiple-choice items with 3 operational polytomous items and 15 experimental items. Because the Lexile Linking Test includes only dichotomous items, the total possible score for items on the NC READY EOC English II assessment was computed by summing the number of one-point multiple-choice items and the number of score points for the open-ended items. This process yielded a total of 56 score points.

The items for the Lexile Linking Tests were chosen to optimize the match to the target test. The IRT difficulty values associated with the NC READY EOG Reading/EOC English II items were converted to Lexile measures using a computer program developed by MetaMetrics, Inc. (no date). Each Lexile Linking Test had a mean Lexile measure established through analysis of the difficulties of the passages on the target test, normative grade-level means, and the item difficulties for the NC READY EOG Reading/EOC English II assessments for 2013. The following mean targets were set: Grade 3, 722L; Grade 5, 963L; Grade 7, 1129L; Grade 8, 1205L; and English II, 1273L.

Evaluation of T-parallel Lexile Linking Tests. After administration, the Lexile Linking Test items were reviewed. Based on the item examination, four items were removed from further analyses, one item from Grade 3 Form 1, one item from Grade 5 Form 1, one item from Grade 5 Form 2, and one item from English II Form 1. These items indicated an alternate answer choice was more attractive than the correct answer choice. While a few 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 the Winsteps analyses). The raw score descriptive statistics for the Lexile Linking Tests are presented in *Table 5*.

Table 5. Descriptive statistics from the development of the Lexile Linking Tests raw scores.

Grade	Test Form	*N*	Raw Score Mean (SD)	Minimum Score		Maximum Score	
				Observed	Possible	Observed	Possible
3	1	1,197	27.72 (9.3)	4	0	43	43
3	2	1,144	28.97 (9.7)	5	0	44	44
5	1	1,151	31.18 (7.8)	1	0	43	43
5	2	1,134	31.18 (7.9)	8	0	43	43
7	1	1,142	33.15 (9.5)	2	0	48	48
7	2	1,110	32.79 (9.5)	0	0	48	48
8	1	1,485	31.27 (9.8)	5	0	48	48
8	2	1,473	31.11 (9.4)	2	0	48	48
Eng II	1	1,334	38.67 (11.9)	0	0	55	55
Eng II	2	1,320	38.92 (11.9)	4	0	56	56
Total		12,490					

Selected item statistics for the Lexile Linking Tests are presented in *Table 6*.

Table 6. Item statistics from the administration of the Lexile Linking Tests.

Grade		*N* (Persons)	*N* (Items)	Percent Correct Mean (Range)	Point-Biserial Range	Coefficient Alpha
3	1	1,197	43	64 (22 - 94)	0.24 - 0.60	0.920
3	2	1,144	44	66 (25 - 89)	0.29 - 0.61	0.926
5	1	1,151	43	73 (28 - 97)	0.08 - 0.57	0.902
5	2	1,134	43	73 (34 - 98)	0.23 - 0.57	0.903
7	1	1,142	48	69 (31 - 92)	0.13 - 0.59	0.918
7	2	1,110	48	68 (21 - 93)	0.12 - 0.61	0.918
8	1	1,485	48	65 (28 - 89)	0.11 - 0.56	0.919
8	2	1,473	48	65 (33 - 90)	0.11 - 0.54	0.910
Eng II	1	1,334	55	70 (31 - 91)	0.26 - 0.64	0.944
Eng II	2	1,320	56	70 (26 - 93)	0.20 - 0.64	0.941
Total		12,490				

The Coefficient Alpha correlations for each of the ten Lexile Linking Tests, two for each grade/course, ranged from 0.902 to 0.944. This indicates strong internal consistency reliability for each of the ten tests and high consistency across these ten 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 sets of items to examinees is operationally possible, and (2) differential order effects are not expected to occur" (pp. 16–17). The NC READY EOG Reading assessments were administered between April 8, 2013 and April 26, 2013. The Lexile Linking Tests were administered within two weeks of the administration of the NC READY EOG Reading assessments. The NC READY EOC English II assessment was administered between April 29, 2013 and May 15, 2013. The Lexile Linking Test was administered within two weeks of the administration of the NC READY EOC English II assessment.

Description of the Sample

The sample of students for the study was selected by the North Carolina Department of Public Instruction. The participating schools were located from across North Carolina with a total of 121 schools from 75 districts participating in the linking study.

Table 7 presents the number of students tested in the linking study and the percentage of students with complete data (both a NC READY EOG Reading/EOC English II score and a Lexile Linking Test Lexile measure). A total of 12,356 students (Grades 3, 5, 7, 8, and English II), or 98.9%, had both test scores. This sample will be referred to as the matched sample.

Table 7. Number of student tests received and number of students in the matched sample.

Grade	NC READY EOG Reading/EOC English II Received *N*	Lexile Linking Test *N*	Matched *N*	Matched Percent
3	103,173	2,341	2,318	99.0
5	109,836	2,285	2,260	98.9
7	110,944	2,252	2,224	98.8
8	108,983	2,958	2,939	99.4
Eng II	108,188	2,654	2,615	98.5
Total	541,124	12,490	12,356	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. 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,
- 100% correct on the Lexile Linking Test,
- Missing total score on the NC READY EOG Reading/EOC English II assessment,
- Misfit to the Rasch model, or
- Showed greater than a 25-percentile-rank difference between the NC READY EOG Reading/EOC English II assessment scale scores and Lexile Linking Test Lexile measures within grade.

Table 8 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 12,356 students in the matched sample, 9,777 (79.1%) remained in the final sample. The table also summarizes the number of student test scores (by grade) removed from analysis, and the reason for their removal.

Table 8. Comparison of matched sample and final sample and the reason for student removal.

Matched Sample		*N* Removed by Reason				Final Sample	
Grade	*N*	Accommodated Students	Misfit to Rasch	Scores Removed	Percentile Rank Difference	*N*	Percent of Matched Sample
3	2,318	3	91	40	281	1,903	82.1
5	2,260	2	130	24	377	1,727	76.4
7	2,224	1	59	15	379	1,770	79.6
8	2,939	9	74	23	524	2,309	78.6
Eng II	2,615	0	47	49	451	2,068	79.1
Total	12,356	15	401	151	2,012	9,777	79.1

* Note: Students with a 100% correct on the linking test or with an invalid NC READY EOG Reading/EOC English II assessment score.

Table 9 presents the demographic characteristics of all students in the NC READY EOG Reading/EOC English II sample, the matched sample, and the final sample of students included in this study. Across the samples, the final sample is similar to the other two samples.

Table 9. Percentage of students in the NC READY EOG Reading/EOC English II sample, matched sample, and final sample for selected demographic characteristics.

Student Characteristic	Category	State Sample *N*=541,124	Matched Sample *N*=12,356	Final Sample *N*=9,777
Grade or Course	3	19.1	18.8	19.5
	5	20.3	18.3	17.7
	7	20.5	18.0	18.1
	8	20.1	23.8	23.6
	English II	20.0	21.2	21.2
Gender	Female	49.6	49.6	50.4
	Male	50.4	50.4	49.6
	Unknown/not avail	0.1	0.0	0.0
Race/Ethnicity	American Indian	1.4	0.9	1.0
	Asian	2.6	2.4	2.4
	Black	25.7	24.7	24.5
	Hispanic	13.4	12.8	13.2
	Pacific Islander	0.1	0.2	0.2
	White	53.1	55.6	55.3
	Two or more	3.7	3.4	3.5
	N/A	0.1	0.0	0.0
LEP Status	Currently identified	5.4	5.1	5.4
	Exit by committee	0.0	0.0	0.0
	Exits LEP	5.6	5.7	5.7
	Never identified	88.8	89.1	88.7
	No Status	0.1	0.0	0.0
	Parental refusal of IPT testing	0.1	0.1	0.1
Student/Disability	Exited within 2 years	1.7	1.6	1.5
	Yes	8.9	8.5	8.8
	No	89.4	90.0	89.7

Student Characteristic	Category	State Sample *N*=541,124	Matched Sample *N*=12,356	Final Sample *N*=9,777
EC Code	Autism	0.5	0.6	0.6
	Deaf-Blindness	0.0	0.0	0.0
	Deafness	0.0	0.0	0.0
	Developmental Delay	0.1	0.0	0.0
	Hearing Impairment	0.1	0.1	0.1
	Intell. Disability - Mild	0.2	0.2	0.2
	Intell. Disability - Moderate	0.0	0.0	0.0
	Multiple Disabilities	0.0	0.0	0.0
	Not Provided	89.4	90.0	89.7
	Orthopedic Impairment	0.0	0.1	0.1
	Other Health Impairment	2.3	2.1	2.1
	Serious Emotional Disability	0.4	0.2	0.2
	Specific Learning Disability	5.2	4.7	4.9
	Speech or Language Impairment	1.9	2.1	2.1
	Traumatic Brain Injury	0.0	0.0	0.0
	VI	0.0	0.0	0.0
Plan-504	Yes	1.1	1.4	1.4
	No	98.9	98.6	98.6
Word To Word Bilingual	Yes	0.2	0.1	0.0
	No	99.8	99.9	100.0
Acad/Intell Gifted - Reading	Yes	10.8	10.1	10.0
	No	89.2	89.9	90.0

Table 10 presents the descriptive statistics for the NC READY EOG Reading/EOC English II scale scores and the Lexile Linking Test Lexile measures for the matched sample. The correlations between the NC READY EOG Reading/EOC English II scale scores and the Lexile Linking Test measures range from 0.769 to 0.824. Based upon the correlations between the NC READY EOG Reading/EOC English II scale scores and the

Lexile Linking Test Lexile measures presented in *Table 10,* it can be concluded that the two tests are measuring similar reading comprehension constructs.

Table 10. Descriptive statistics for the NC READY EOG Reading/EOC English II scale scores and Lexile measures and the Lexile Linking Test Lexile measures, matched sample (*N* = 12,356).

Grade	*N*	Matched Sample NC READY EOG Reading/EOC English II Scale Score Mean (SD)	Matched Sample Lexile Linking Test Lexile Measure Mean (SD)	*r*
3	2,318	440.18 (10.4)	697.98 (253.4)	0.824
5	2,260	449.18 (9.5)	1019.58 (226.5)	0.795
7	2,224	455.81 (10.2)	1138.34 (237.4)	0.769
8	2,939	458.55 (10.7)	1168.69 (226.8)	0.770
Eng II	2,615	150.68 (9.0)	1295.86 (259.2)	0.769
Total	12,356			

Table 11 presents the descriptive statistics of the NC READY EOG Reading/EOC English II test scale scores as well as the Lexile Linking Test Lexile measures for the final sample. The correlations between the final sample NC READY EOG Reading/EOC English II scale scores and the final sample Lexile Linking Test measures range from 0.877 to 0.893. These correlations between the two scores are strong and higher than the matched sample.

Table 11. Descriptive statistics for the NC READY EOG Reading/EOC English II scale scores and the Lexile Linking Test Lexile measures, final sample (*N* = 9,777).

Grade	*N*	Final Sample NC READY EOG Reading/EOC English II Scale Score Mean (SD)	Final Sample Lexile Linking Test Lexile Measure Mean (SD)	*r*
3	1,903	439.69 (10.1)	686.13 (233.3)	0.893
5	1,727	449.12 (9.3)	1016.02 (209.8)	0.883
7	1,770	455.65 (10.3)	1135.65 (229.9)	0.877
8	2,309	458.41 (10.7)	1169.21 (217.5)	0.888
Eng II	2,068	150.30 (9.1)	1285.82 (239.1)	0.887
Total	9,777			

Figures 2 through *11* shows the relationship between the NC READY EOG Reading/EOC English II scale scores and the Lexile Linking Test Lexile measures for the matched and final samples for each grade/course. In each grade/course, it can be seen that there is a linear relationship between the NC READY EOG Reading/EOC English II scale score and the final sample Lexile measure reinforcing the use of linear equating.

Figure 2. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 3 matched sample (*N* = 2,318).

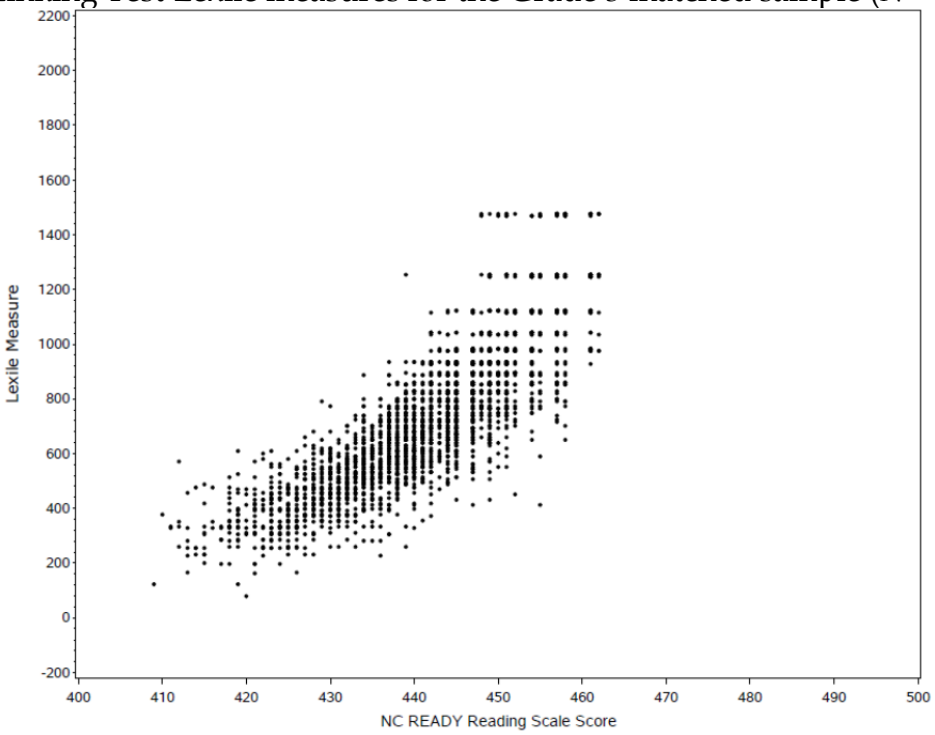


Figure 3. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 3 final sample (*N* = 1,903).

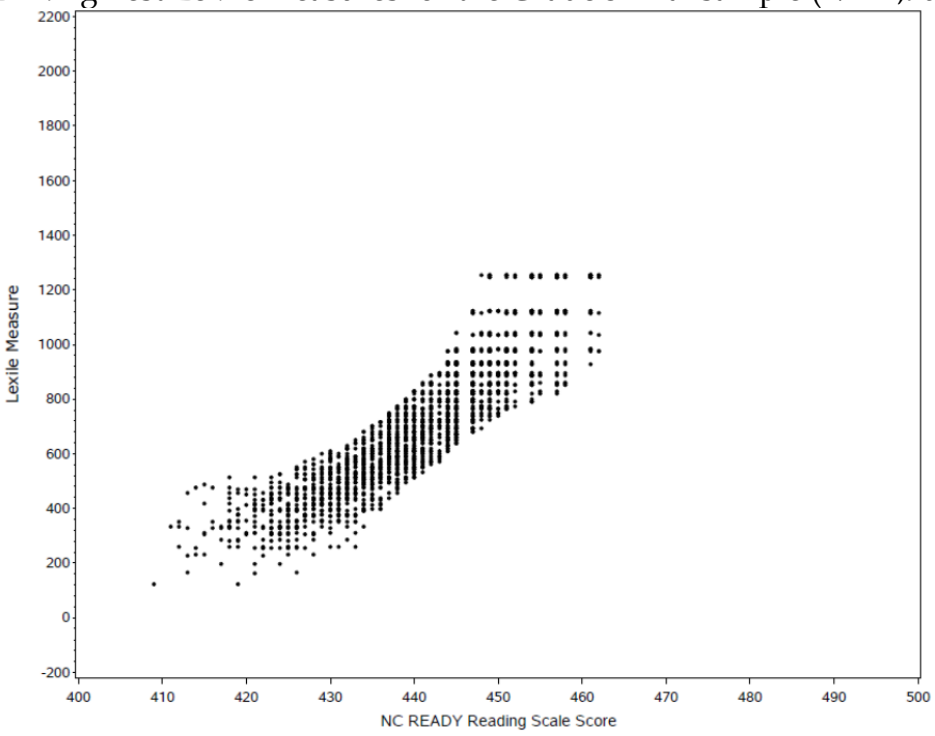


Figure 4. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 5 matched sample (*N* = 2,260).

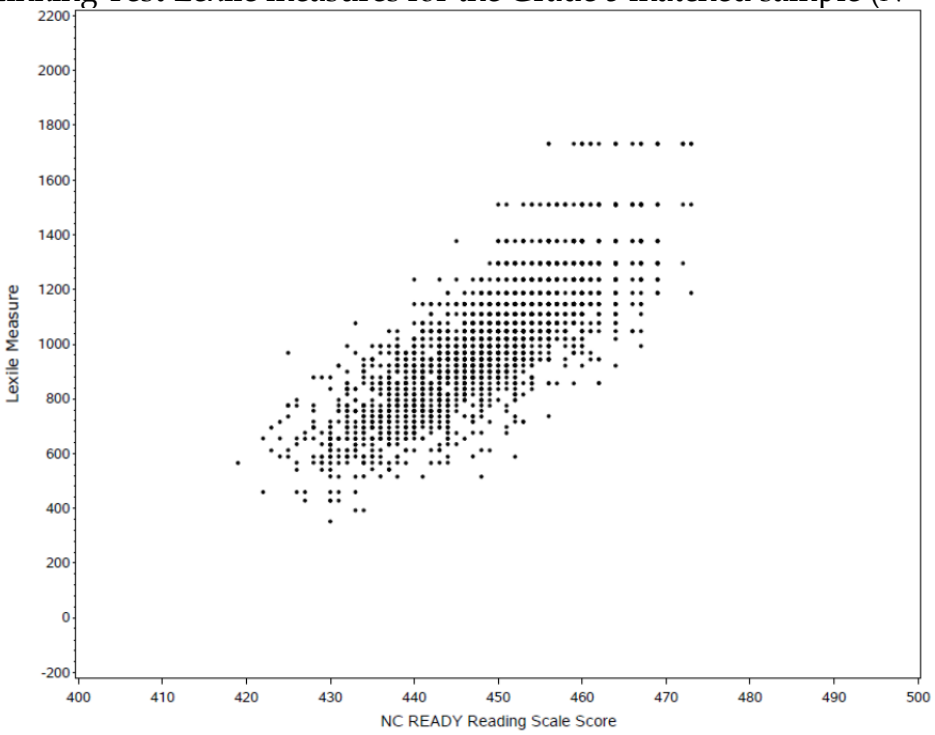


Figure 5. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 5 final sample (*N* = 1,727).

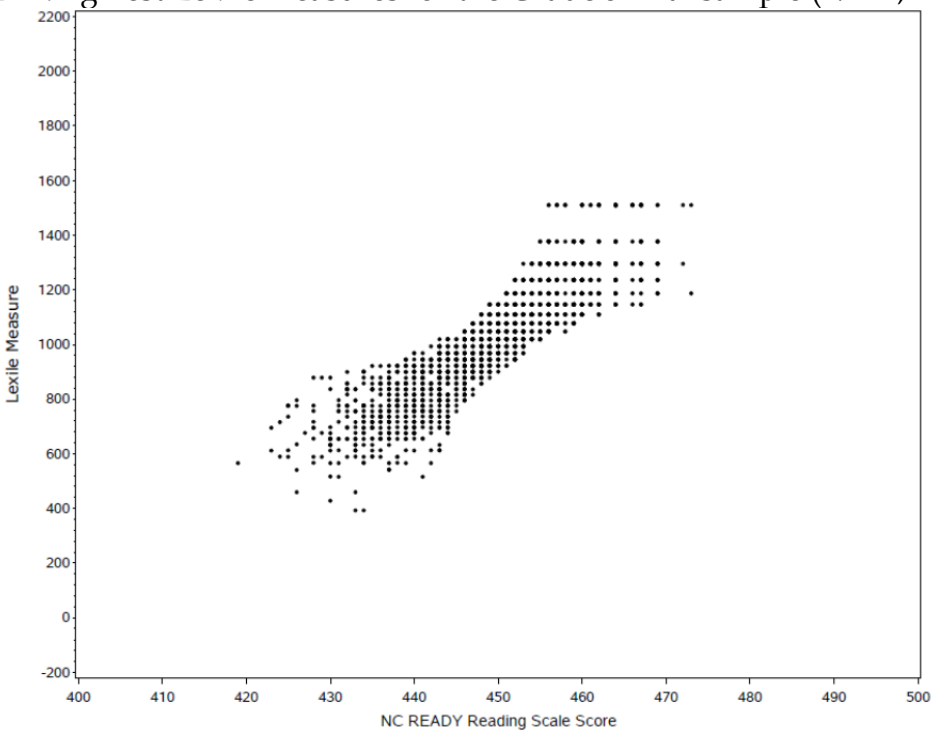


Figure 6. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 7 matched sample (*N* =2,224).

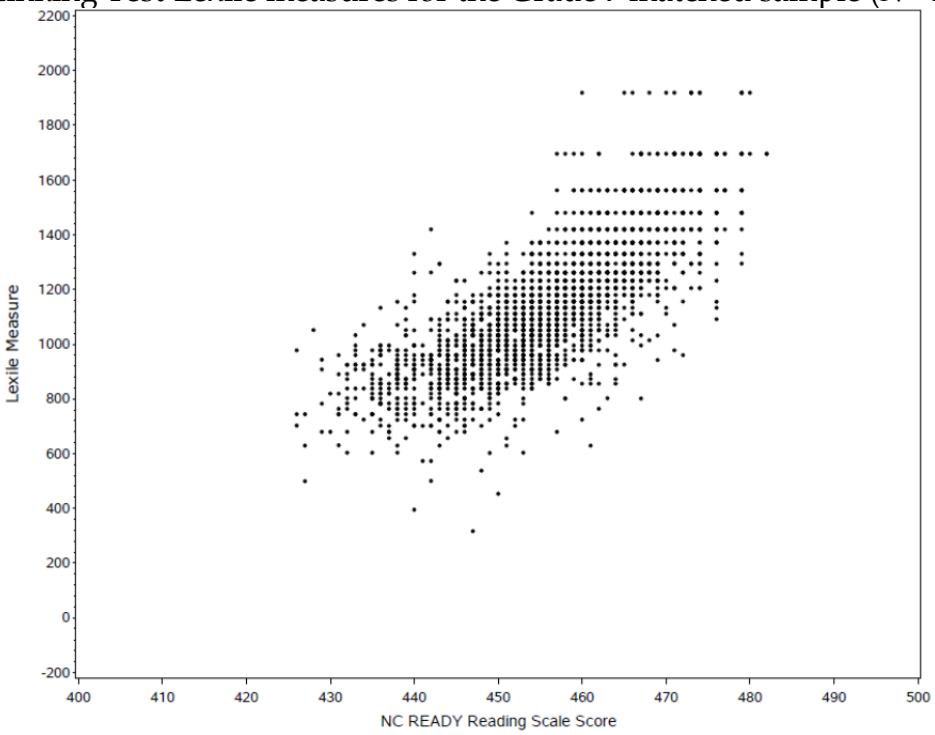


Figure 7. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 7 final sample (*N* = 1,770).

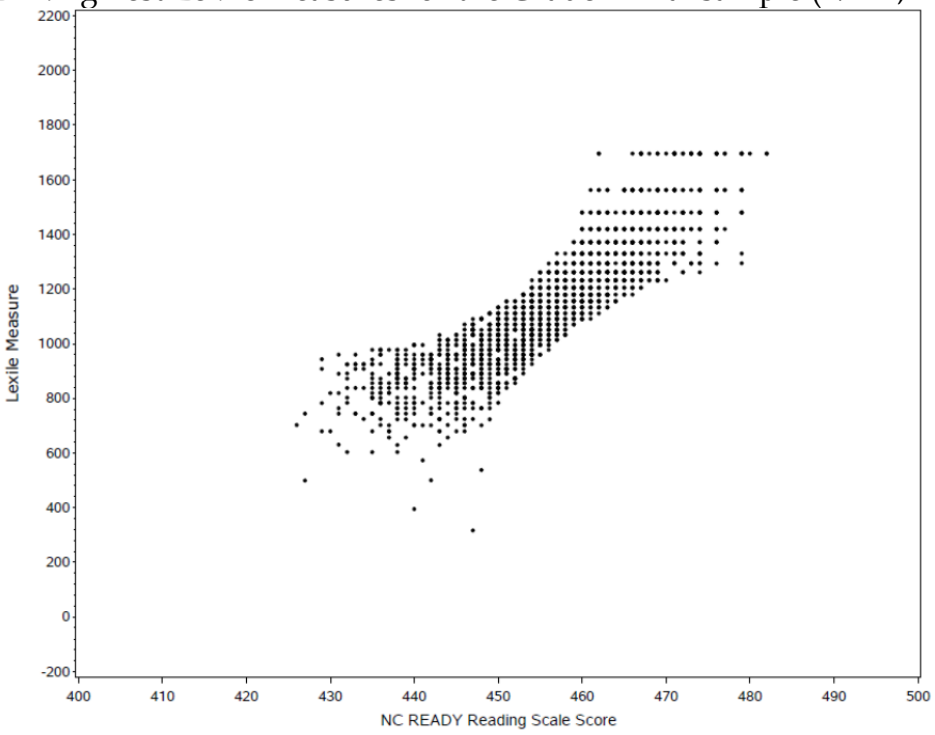


Figure 8. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 8 matched sample (*N* = 2,939).

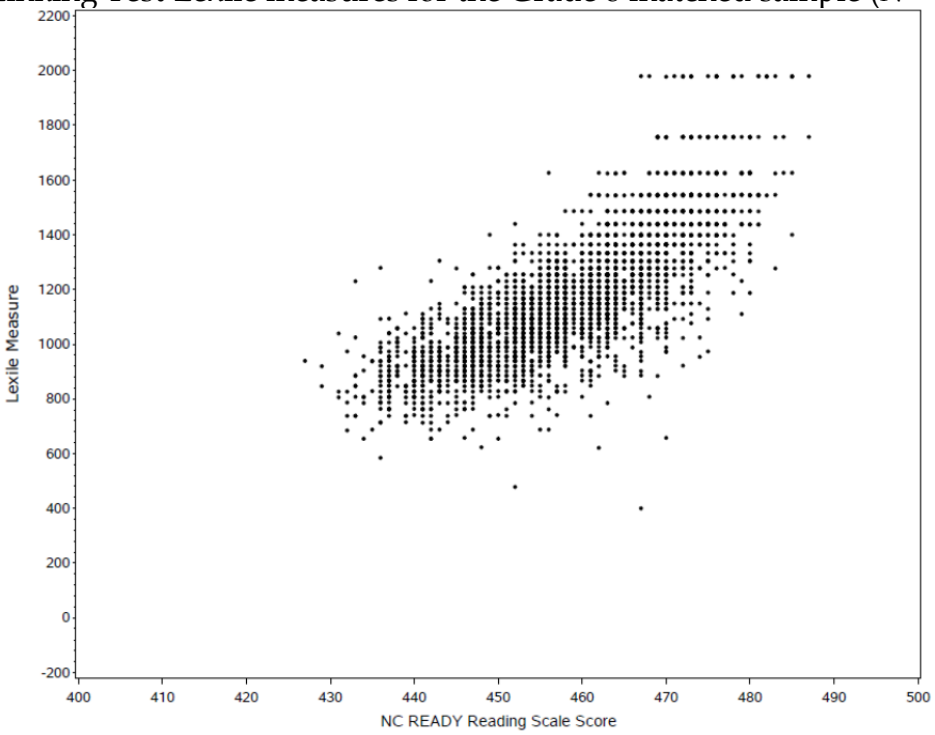


Figure 9. Scatter plot of the NC READY EOG Reading scale scores and the Lexile Linking Test Lexile measures for the Grade 8 final sample (*N* = 2,309).

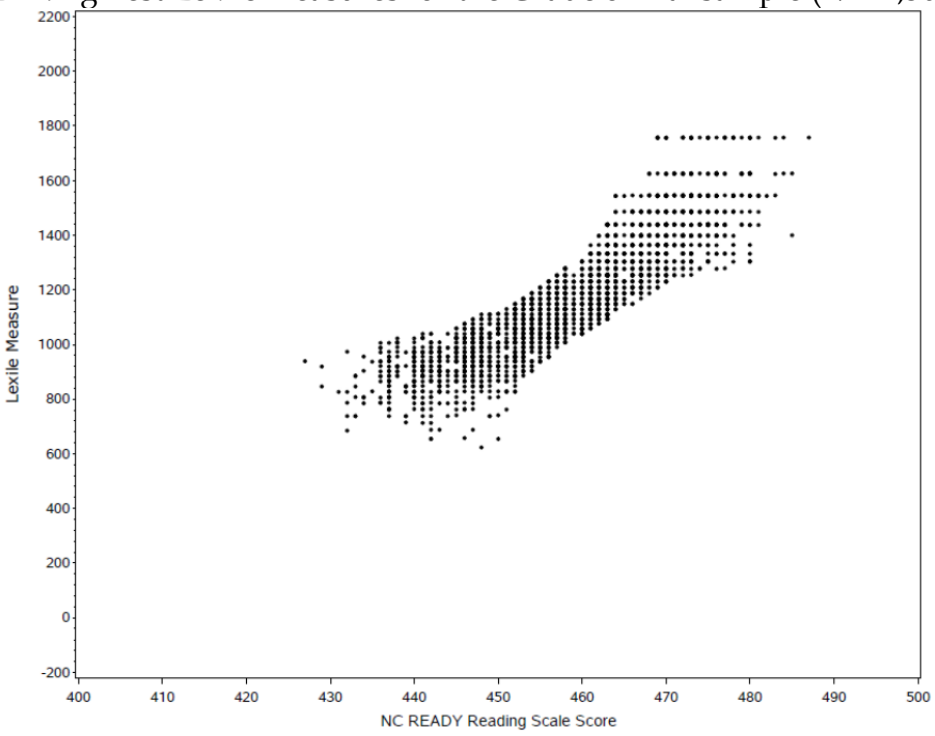


Figure 10. Scatter plot of the NC READY EOC English II scale scores and the Lexile Linking Test Lexile measures for the English II matched sample (*N* = 2,615).

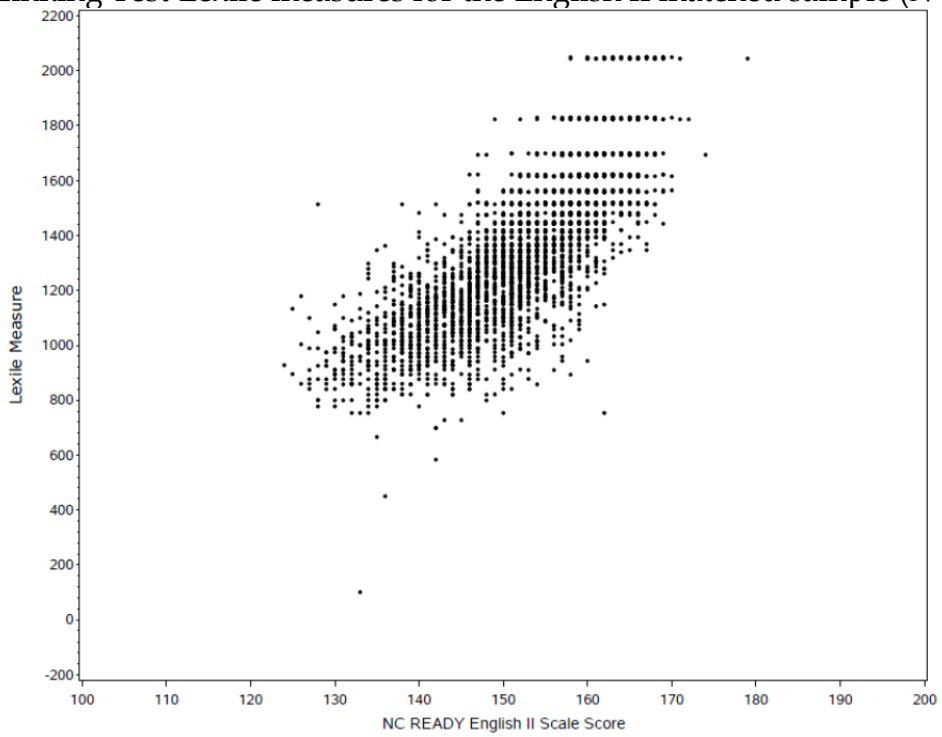
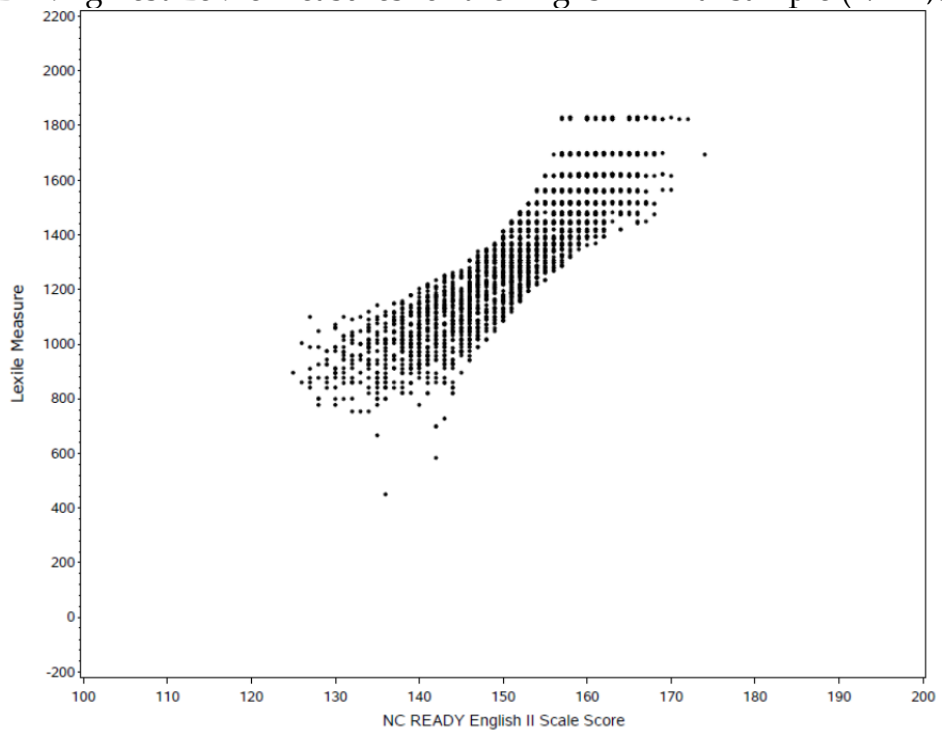


Figure 11. Scatter plot of the NC READY EOC English II scale scores and the Lexile Linking Test Lexile measures for the English II final sample (*N* = 2,068).



Linking the NC READY EOG Reading/EOC English II Scale Scores with the Lexile Scale

Linking in general means "putting the scores from two or more tests on the same scale" (National Research Council, 1999, p.15). MetaMetrics and the North Carolina Department of Public Instruction conducted this linking study for the purpose of matching students with books and texts—to predict the books and texts a student should be matched with for successful reading experiences, given their performance on the NC READY EOG Reading/EOC English II assessment.

Evaluation of linkage assumptions. Factors that affect the linkage between two assessments include the domain to be assessed, the definition of the framework for assessment, the test specifications, and the items sampled.

Based upon the correlations between the NC READY EOG Reading/EOC English II scale scores and the Lexile Linking Tests Lexile measures presented in *Table 11,* it can be concluded that the two assessments measure similar constructs. The correlations between the two assessments are above or within the typical range of alternate-form reliability coefficients; therefore, the Lexile Linking Tests can be considered a T-parallel form of the NC READY EOG Reading/EOC English II test (see Note 1). By using alternate-form reliability coefficients as a comparison, similar sources of variation are accounted for (differences in testing occasions and items). In addition, the linking tests were constructed to have a similar number of items and the same level of difficulty as the NC READY EOG Reading/EOC English II assessments.

Linking Analyses. Two score scales (e.g., the NC READY EOG Reading/EOC English II scale and the Lexile Scale) can be linked using linear equating when (1) test forms have similar difficulties; and (2) 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 sets of items 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, Kolen, and Hoover, 1989; Kolen and Brennan, 2004). Given scores x and y on tests X and Y, the linear relationship is

$$\frac{(x-\mu_X)}{\sigma_X} = \frac{(y-\mu_y)}{\sigma_y} \qquad \text{(Equation 2)}$$

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

$$x = l_x(y) = \left(\frac{\sigma_X}{\sigma_y}\right)y + \left(\mu_x - \frac{\mu_y\sigma_X}{\sigma_y}\right) \qquad \text{(Equation 3)}$$

Linear equating by definition has the same mean and standard deviation for the overall equation when the scale is vertically aligned. The means and standard deviations are the same for the Linking test and the Target test when calculated across grades. The values are somewhat different when the formula is developed by grade. 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 NC READY EOG Reading/EOC English II scale scores and Lexile measures can be written as:

$$\text{Lexile measure} = \text{Slope}_g(\text{NC READY EOG Reading/EOC English II scale score}) + \text{constant}_g \qquad \text{(Equation 4)}$$

where the slope is the ratio of the standard deviations of the NC READY EOG Reading/EOC English II scale scores and Lexile Linking Test Lexile measures. These values for each grade range/course can be found in *Table 11*.

Using the final sample data described in *Table 11*, the linear linking functions relating the NC READY EOG Reading/EOC English II scale scores and Lexile measures for students in the final sample are presented in *Table 12*. One linking function was developed for each of the following groups (*g*): (1) Grades 3 through 8 of the NC READY EOG Reading assessment and (2) EOC English II assessment.

Table 12. Linear linking equation coefficients used to predict Lexile measures from the NC READY EOG Reading and the EOC English II scale scores.

Group (*g*)	Slope	Intercept
3 - 8	23.488825	-9587.222
English II	26.264583	-2661.751

Conversion tables were developed for all grade levels in order to express the NC READY EOG Reading/EOC English II scale scores in the Lexile metric and were delivered to the North Carolina Department of Public Instruction in electronic format.

Table 13 contains the maximum reported Lexile measures by grade. The measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose of the test is accountability (at the student, school, or district level), then uncapped Lexile measures should be reported. 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 of the national Lexile norms). In an instructional environment where the purpose of the Lexile measure is to appropriately match readers with texts, all scores below 0L should be reported as "BRxxxL." No student should receive a negative Lexile measure on a score report. The lowest reported value below 0L is BR400L.

Table 13. Capped values of the Lexile measure by grade/course.

Grade/Course	Capped Lexile Measure
3	1200L
4	1300L
5	1400L
6	1500L
7	1600L
8	1700L
Eng II	1750L

Validity of the NC READY EOG Reading/EOC English II—Lexile Link

Table 14 presents the descriptive statistics and effect size statistics of the NC READY EOG Reading/EOC English II Lexile measures as well as the Lexile Linking Test Lexile measures for the final sample.

Table 14. Descriptive statistics and effect size statistics for the final sample NC READY EOG Reading/EOC English II Lexile measures and the Lexile Linking Test Lexile measures.

Grade	*N*	Final Sample NC READY EOG Reading/EOC English II Lexile Measure Mean (SD)	Final Sample Lexile Linking Test Lexile Measure Mean (SD)	Effect Size
3	1,903	740.42 (237.1)	686.13 (233.3)	0.230793
5	1,727	961.98 (218.7)	1016.02 (209.8)	-0.252219
7	1,770	1115.5 (240.9)	1135.66 (229.9)	-0.085595
8	2,309	1180.38 (252.7)	1169.21 (217.5)	0.047384
Eng II	2,068	1285.82 (239.2)	1285.82 (239.1)	0.000003
Total	9,777			

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

Table 15. Interpretation chart for effect size.

Small	0.20
Medium	0.50
Large	0.80

In *Table 14*, for the 5 comparisons, effect sizes were minimal for three comparisons indicating no significant difference between the NC READY EOG Reading/EOC English II Lexile measures and the Lexile Linking Test Lexile measures. Two comparisons, Grades 3 and 5, were slightly larger by at most only .05 within the medium range which was not a concern.

Table 16 contains the percentile ranks of the Lexile Linking Test Lexile measures and the NC READY EOG Reading/EOC English II assessment Lexile measures based on the final sample. The criterion of a half standard deviation (100L) on the Lexile 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 Lexile measures on the NC READY EOG Reading/EOC English II assessments.

Table 16. Comparison of the Lexile measures for selected percentile ranks for the final sample Lexile Linking Test and the NC READY EOG Reading/EOC English II assessment.

Grade 3		
Percentile Rank	Linking Test Lexile Measure	NC READY EOG Reading Sample Lexile Measure
1	255	184
5	333	349
10	398	419
25	507	583
50	659	748
75	852	912
90	983	1030
95	1115	1100
99	1254	1241

Grade 5		
Percentile Rank	Linking Test Lexile Measure	NC READY EOG Reading Sample Lexile Measure
1	567	466
5	675	583
10	736	677
25	878	818
50	1019	959
75	1187	1124
90	1296	1241
95	1377	1312
99	1510	1429

Grade 7		
Percentile Rank	Linking Test Lexile Measure	NC READY EOG Reading Sample Lexile Measure
1	679	560
5	783	701
10	855	795
25	960	959
50	1133	1124
75	1294	1288
90	1420	1429
95	1562	1500
99	1696	1617

Grade 8		
Percentile Rank	Linking Test Lexile Measure	NC READY EOG Reading Sample Lexile Measure
1	741	654
5	848	748
10	902	818
25	1007	1006
50	1149	1171
75	1305	1359
90	1485	1500
95	1546	1570
99	1756	1687

Table 16. (continued). Comparison of the Lexile measures for selected percentile ranks for the final sample Lexile Linking Test and the NC READY EOG Reading/EOC English II assessment.

English II		
Percentile Rank	Linking Test Lexile Measure	NC READY EOC English II Sample Lexile Measure
1	800	726
5	912	858
10	974	963
25	1104	1120
50	1279	1304
75	1449	1462
90	1616	1593
95	1694	1646
99	1829	1751

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 Instruction established four achievement levels: Level 1, Level 2, Level 3, and Level 4 (NCDPI, 2013b). As an example, the four achievement levels for the Grade 3 NC READY EOG Reading Assessment are:

Level 1: Students performing at this level have **limited command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) Reading Standards for Literature as assessed by referring to the text when asking and answering questions; recounting stories and determining a central message, explaining how the message is conveyed through key details in the text; describing characters and explaining how their actions contribute to the plot; and determining the meaning of words and phrases as they are used in a text, especially literal and nonliteral language. They will need academic support to engage successfully 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*) Reading Standards for Literature as assessed by referring to the text when asking and answering questions; recounting stories and determining a central message, explaining how the message is conveyed through key details in the text; describing

characters and explaining how their actions contribute to the plot; and determining the meaning of words and phrases as they are used in a text, especially literal and nonliteral language. They will likely need academic support to engage successfully 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*) Reading Standards for Literature as assessed by referring to the text when asking and answering questions; recounting stories and determining a central message, explaining how the message is conveyed through key details in the text; describing characters and explaining how their actions contribute to the plot; and determining the meaning of words and phrases as they are used in a text, especially literal and nonliteral language. They are academically prepared to engage successfully 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*) Reading Standards for Literature as assessed by referring to the text when asking and answering questions; recounting stories and determining a central message, explaining how the message is conveyed through key details in the text; describing characters and explaining how their actions contribute to the plot; and determining the meaning of words and phrases as they are used in a text, especially literal and nonliteral language. They are academically well-prepared to engage successfully in this content area.

The four achievement levels for NC READY EOC English II Assessment (NCDPI, 2013a) are:

Level 1: Students performing at this level have **limited command** of the knowledge and skills contained in the *Common Core State Standards* (*CCSS*) Reading Standards for Literature as assessed by supporting analysis of the text with textual evidence; determining and analyzing the development and refinement of a theme or idea throughout a text; summarizing a text objectively; analyzing the development, interaction, and contribution of characters in a text; determining meanings of words or phrases in a text; analyzing the impact of word choice on meaning and tone; analyzing how authors' choices create literary effects, such as tension; analyzing point of view and cultural experiences in literature from outside the U.S., drawing on world literature. They will need academic support to engage successfully 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*) Reading Standards for Literature as assessed by supporting analysis of the text with textual evidence; determining and analyzing the development and refinement of a theme or idea throughout a text; summarizing a text objectively; analyzing the development, interaction, and contribution of characters in a text; determining

meanings of words or phrases in a text; analyzing the impact of word choice on meaning and tone; analyzing how authors' choices create literary effects, such as tension; analyzing point of view and cultural experiences in literature from outside the U.S., drawing on world literature. They will likely need academic support to engage successfully 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*) Reading Standards for Literature as assessed by supporting analysis of the text with textual evidence; determining and analyzing the development and refinement of a theme or idea throughout a text; summarizing a text objectively; analyzing the development, interaction, and contribution of characters in a text; determining meanings of words or phrases in a text; analyzing the impact of word choice on meaning and tone; analyzing how authors' choices create literary effects, such as tension; analyzing point of view and cultural experiences in literature from outside the U.S., drawing on world literature. They are academically prepared to engage successfully 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*) Reading Standards for Literature as assessed by supporting analysis of the text with textual evidence; determining and analyzing the development and refinement of a theme or idea throughout a text; summarizing a text objectively; analyzing the development, interaction, and contribution of characters in a text; determining meanings of words or phrases in a text; analyzing the impact of word choice on meaning and tone; analyzing how authors' choices create literary effects, such as tension; analyzing point of view and cultural experiences in literature from outside the U.S., drawing on world literature. They are academically well-prepared to engage successfully in this content area.

Table 17 presents the achievement level cut scores on the NC READY EOG Reading/EOC English II assessments and the associated Lexile measures. There are four achievement levels: Level 1, Level 2, Level 3, and Level 4 (NCDPI, 2013a, 2013b). The values in the table are the cut scores associated with the bottom score for each category.

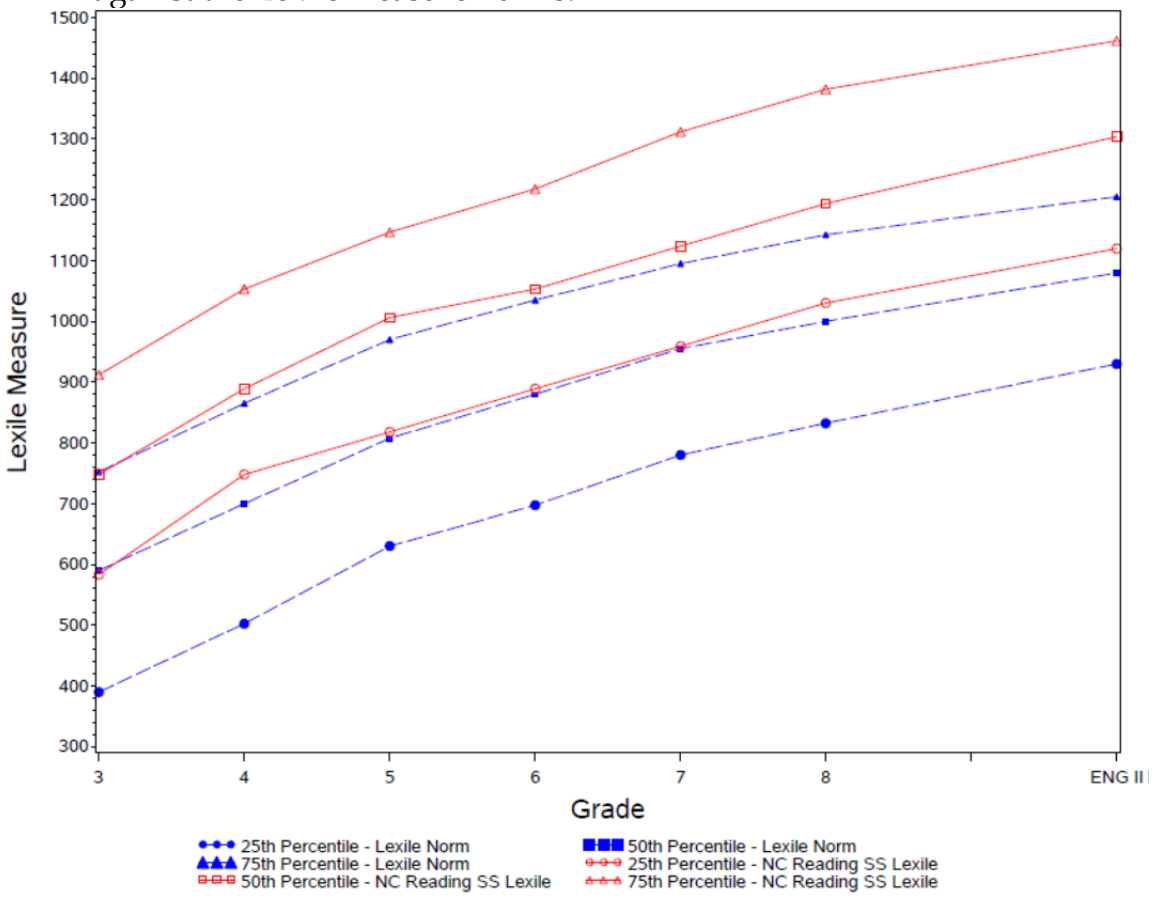
Table 17. NC READY EOG Reading/EOC English II performance level cut scores and the associated Lexile measures.

Grade	Level 2		Level 3		Level 4	
	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure	NC READY EOG Reading/EOC English II Scale Score	Lexile Measure
3	432	560L	442	795L	452	1030L
4	439	725L	448	935L	460	1220L
5	443	820L	453	1055L	464	1310L
6	442	795L	454	1075L	465	1335L
7	445	865L	457	1145L	469	1430L
8	449	960L	462	1265L	473	1525L
E II	141	1040L	151	1305L	165	1670L

Figure 12 shows the Lexile measures for the NC READY EOG Reading/EOC English II assessment as compared to the norms that have been developed for use with The Lexile Framework for Reading. These norms were created based on linking studies conducted with the Lexile Framework.

Overall, it can be seen that the NC READY EOG Reading/EOC English II Lexile measures are higher across the grades at each percentile. The 25th percentile for the NC READY EOG Reading/EOC English II Lexile measures is closer to the 50th percentile Lexile measures. The 50th percentile for the NC READY EOG Reading/EOC English II Lexile measures is closer to the 75th percentile Lexile measures. Therefore, the NC READY EOG Reading/EOC English II scores were higher than the Lexile norms. This translates to the statement that the students in North Carolina were more able than the Lexile norms for a national population.

Figure 12. Selected Percentiles (25th, 50th, and 75th) plotted for the NC READY EOG Reading/EOC English II Lexile measure for the final sample (*N* = 9,777) against the Lexile measure norms.



The following box and whisker plots (*Figures 13, 14,* and *15*) show the progression of scores (the *y*-axis) from grade to grade (the *x*-axis) (note, that English II is placed as Grade 10 which is the typical grade for students taking the course). For each grade, the box refers to the interquartile range. The line within the box indicates the median and the • represents the mean. The end of each whisker represents the minimum and maximum values of the scores (the *y*-axis).

The Lexile measures are on a vertical scale and *Figures 13, 14,* and *15* demonstrate this by showing that as the grade increases so do the NC READY EOG Reading/EOC English II Lexile measures. All three plots show a similar profile.

Figure 13. Box and whisker plot of the Lexile Linking Tests Lexile measures by grade, final sample (*N* =9,777).

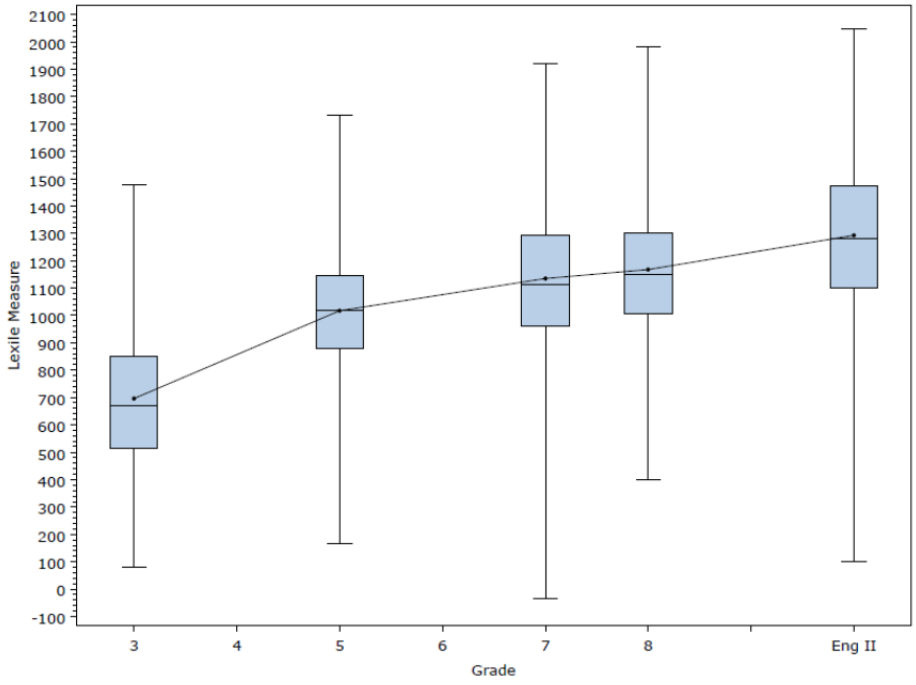


Figure 14. Box and whisker plot of the NC READY EOG Reading/EOC English II Lexile measures by grade, matched sample (*N* = 12,356).

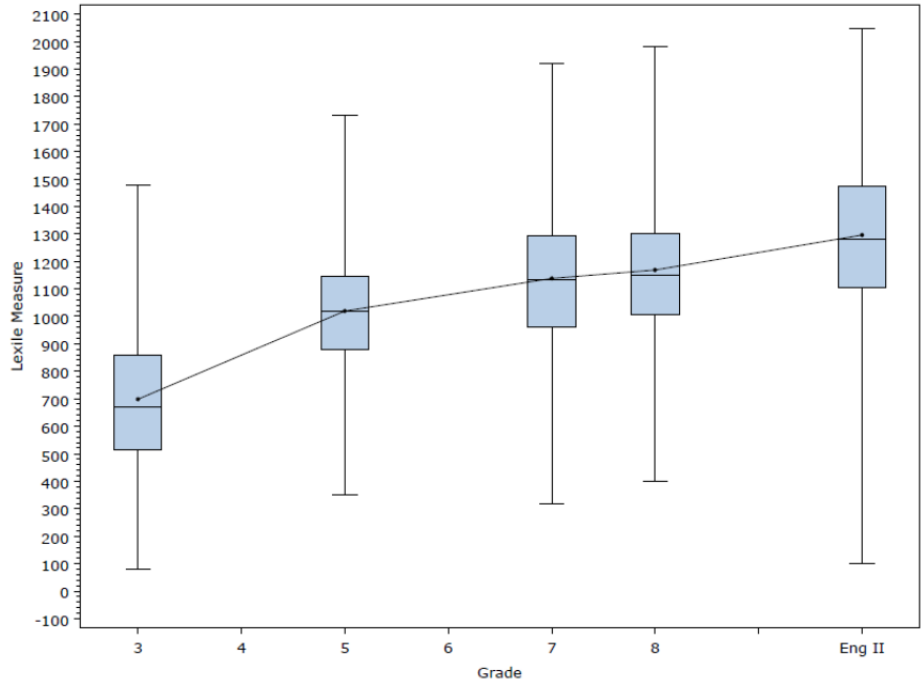
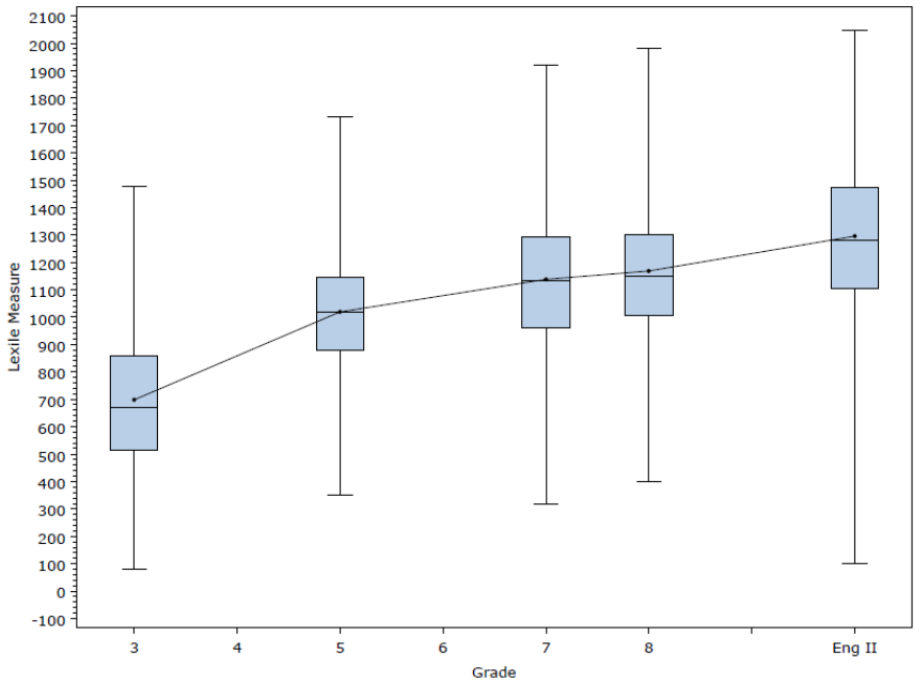


Figure 15. Box and whisker plot of the NC READY EOG Reading/EOC English II Lexile measures by grade, final sample (*N* = 9,777).



The Lexile Framework and Forecasted Comprehension Rates

A reader with a measure of 600L who is given a text measured at 600L is expected to have a 75-percent comprehension rate. This 75-percent comprehension rate is the basis for selecting text that is targeted to a reader's reading ability, but what exactly does it mean? And what would the comprehension rate be if this same reader were given a text measured at 350L or one at 850L?

The 75-percent comprehension rate for a reader-text pairing can be given an operational meaning by imagining the text is carved into item-sized slices of approximately 125-140 words with a question embedded in each slice. A reader who answers three-fourths of the questions correctly has a 75-percent comprehension rate.

Suppose instead that the text and reader measures are not the same. It is the difference in Lexile measures between reader and text that governs comprehension. If the text measure is less than the reader measure, the comprehension rate will exceed 75 percent. If not, it will be less. The question is "By how much?" What is the expected comprehension rate when a 600L reader reads a 350L text?

If all the item-sized slices in the 350L text had the same calibration, the 250L difference between the 600L reader and the 350L text could be determined using the Rasch model equation. This equation describes the relationship between the measure of a student's

level of reading comprehension and the calibration of the items. Unfortunately, comprehension rates calculated by this procedure would be biased because the calibrations of the slices in ordinary prose are not all the same. The average difficulty level of the slices *and* their variability both affect the comprehension rate. Although the exact relationship between comprehension rate and the pattern of slice calibrations is complicated, Equation 5 is an unbiased approximation:

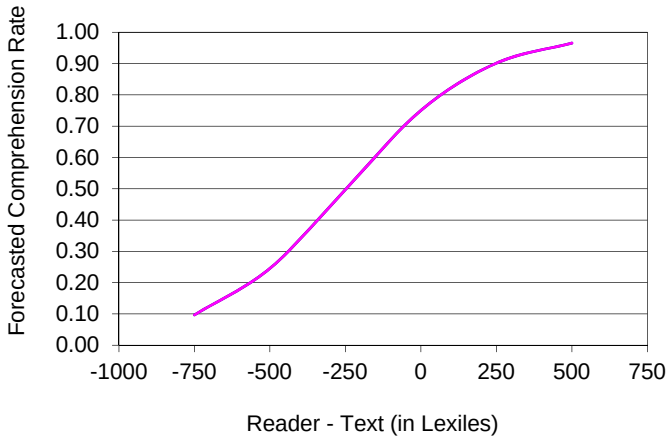
$$Rate = \frac{e^{ELD+1.1}}{1+e^{ELD+1.1}} \qquad \text{(Equation 5)}$$

where ELD is the "effective logit difference" given by

$$\text{ELD} = (\text{Reader Lexile measure} - \text{Text Lexile measure}) \div 225. \qquad \text{(Equation 6)}$$

Figure 16 shows the general relationship between reader-text discrepancy and forecasted comprehension rate. When the reader measure and the text calibration are the same (difference of 0L) then the forecasted comprehension rate is 75 percent. In the example in the preceding paragraph, the difference between the reader measure of 600L and the text calibration of 350L is 250L. Referring to *Figure 16* and using +250L (reader minus text), the forecasted comprehension rate for this reader-text combination would be 90 percent.

Figure 16. Relationship between reader-text discrepancy and forecasted comprehension rate.



Tables 18 and *19* show comprehension rates calculated for various combinations of reader measures and text calibrations.

Table 18. Comprehension rates for the same individual with materials of varying comprehension difficulty.

Person Measure	Text Calibration	Sample Titles	Forecast Comprehension
1000	500	*Tornado* (Byars)	96
1000	750	*The Martian Chronicles* (Bradbury)	90
1000	1000	*Reader's Digest*	75
1000	1250	*The Call of the Wild* (London)	50
1000	1500	*On the Equality Among Mankind* (Rousseau)	25

Table 19. Comprehension rates of different person abilities with the same material.

Person Measure	Calibration for a Grade 10 Biology Textbook	Forecast Comprehension Rate
500	1000	25
750	1000	50
1000	1000	75
1250	1000	90
1500	1000	96

The subjective experience of 50-percent, 75-percent, and 90-percent comprehension as reported by readers varies greatly. A 1000L reader reading 1000L text (75-percent comprehension) reports confidence and competence. Teachers listening to such a reader report that the reader can sustain the meaning thread of the text and can read with motivation and appropriate emotion and emphasis. In short, such readers appear to comprehend what they are reading. A 1000L reader reading 1250L text (50-percent comprehension) encounters so much unfamiliar vocabulary and difficult syntactic structures that the meaning thread is frequently lost. Such readers report frustration and seldom choose to read independently at this level of comprehension. Finally, a 1000L reader reading 750L text (90-percent comprehension) reports total control of the text, reads with speed, and experiences automaticity during the reading process.

The primary utility of the Lexile Framework is its ability to forecast what happens when readers confront text. With every application by teacher, student, librarian, or parent there is a test of the Framework's accuracy. The Framework makes a point prediction every time a text is chosen for a reader. Anecdotal evidence suggests that the Lexile Framework predicts as intended. That is not to say that there is an absence of error in forecasted comprehension. There is error in text measures, reader measures, and their difference modeled as forecasted comprehension. However, the error is sufficiently small that the judgments about readers, texts, and comprehension rates are useful.

Relationship between Linking Error and Forecasted Comprehension Rate. Using Equation 5 with different combinations of reader measure and text difficulty, the effect of linking error on forecasted comprehension rate can be examined. *Table 20* shows the changes in the forecasted comprehension rate for different combinations of reader and text interactions. When the linking error is small, 5–10L, then the effect on forecasted comprehension rate is a minimal difference (1 to 2 percent) increase or decrease in comprehension.

Table 20. Effect of reader-text discrepancy on forecasted comprehension rate.

Reader Lexile Measure	Text Lexile Measure	Difference	Forecasted Comprehension Rate
1000L	970L	30L	77.4
1000L	975L	25L	77.0
1000L	980L	20L	76.7
1000L	985L	15L	76.3
1000L	990L	10L	75.8
1000L	995L	5L	75.4
1000L	1000L	0L	75.0
1000L	1005L	5L	74.6
1000L	1010L	10L	74.2
1000L	1015L	15L	73.8
1000L	1020L	20L	73.3
1000L	1025L	25L	72.9
1000L	1030L	30L	72.4

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 can my child actually read?"),
- 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 Reading/EOC English II scale scores and the Lexile measures permits readers to be matched with books and texts that provide an appropriate level of challenge while avoiding frustration. The result of this purposeful match may be that students will read more, and, thereby read better. The real power of the Lexile Framework is in examining the growth of readers—wherever the reader may be in the development of his or her reading skills. Readers can be matched with texts that they are forecasted to read with 75-percent comprehension. As a reader grows, he or she can be matched with more demanding texts. And, as the texts become more demanding, then the reader grows.

Recommendations about reporting Lexile measures for readers. Lexile measures are reported as a number followed by a capital "L" for "Lexile." There is no space between the measure and the "L," and measures of 1,000 or greater are reported without a comma (e.g., 1050L). All Lexile measures should be rounded to the nearest 5L to avoid over interpretation of the measures. As with any test score, uncertainty in the form of measurement error is present.

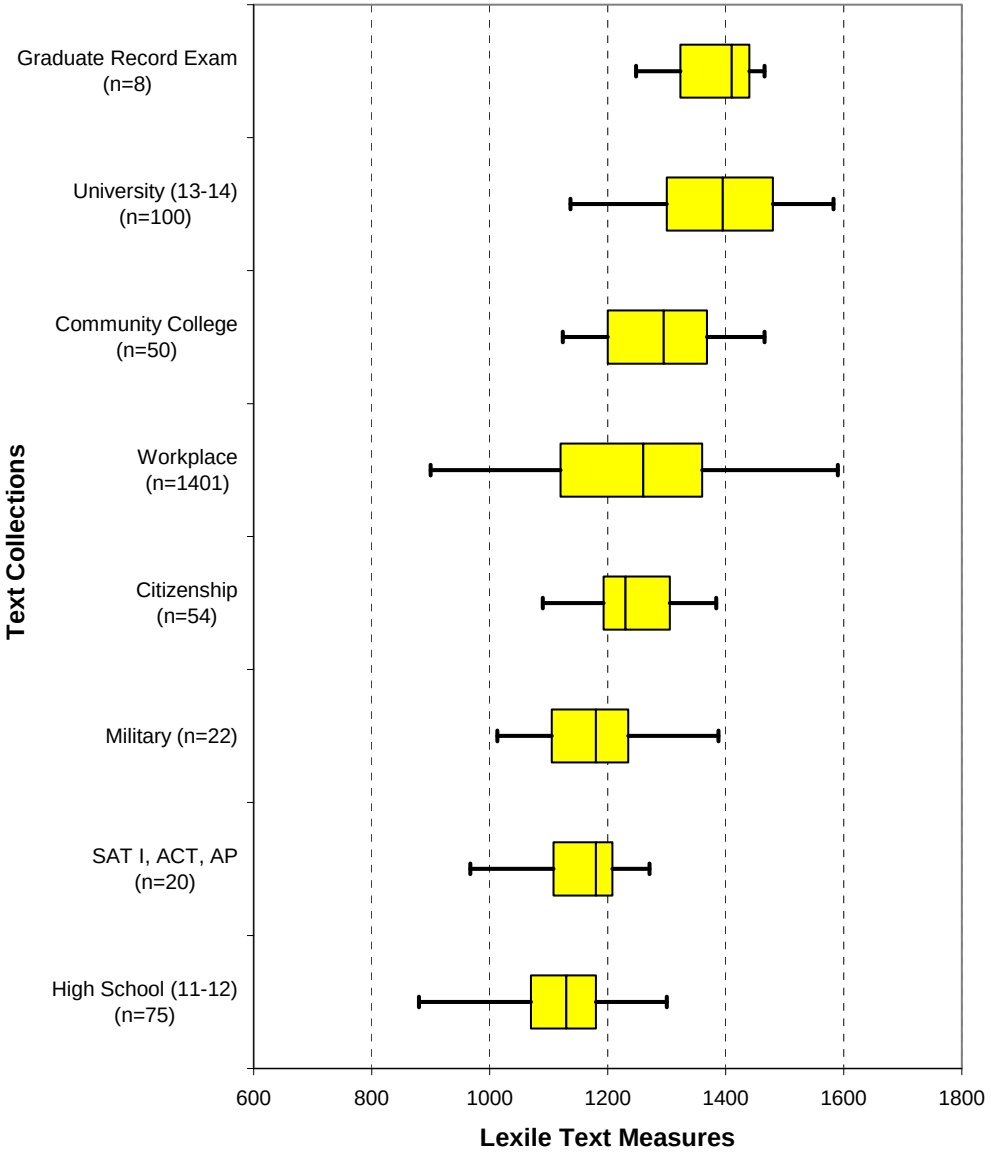
Lexile measures that are reported for an individual student should reflect the purpose for which they will be used. If the purpose is research (e.g., to measure growth at the student, grade, school, district, or state level), then actual measures should be used at all score points, rounded to the nearest integer. A computed Lexile measure of 772.51 would be reported as 773L. If the purpose is instructional, then the Lexile measures should be capped at the upper bound of measurement error (e.g., at the 95th percentile of the national Lexile norms) to ensure developmental appropriateness of the material. MetaMetrics expresses these as "Reported Lexile Measures" and recommends that these measures be reported on individual score reports. In instructional environments where the purpose of the Lexile measure is to appropriately match readers with texts, all scores below 0L should be reported as "BRxxxL." No student should receive a negative Lexile measure on a score report. The lowest reported value below 0L is BR400L.

Some assessments report a Lexile range for each student, which is 50L above and 100L below the student's actual Lexile measure. This range represents the boundaries between the easiest kind of reading material for the student and the level at which the student will be more challenged, yet can still read successfully.

Text Complexity. 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.

In the *Journal of Advanced Academics* (Summer 2008), Williamson investigated the gap between high school textbooks and various reading materials across several postsecondary domains. As can be seen in *Figure 17*, the resources Williamson used were organized into four domains that correspond to the three major postsecondary endeavors that students can choose—further education, the workplace, or the military—and the broad area of citizenship, which cuts across all postsecondary endeavors. Williamson discovered a substantial increase in reading expectations and text complexity from high school to postsecondary domains— a gap large enough to help account for high remediation rates and disheartening graduation statistics (Smith, 2011).

Figure 17. A continuum of text difficulty for the transition from high school to postsecondary experiences (box plot percentiles: 5th, 25th, 50th, 75th, and 95th).[1]



Expanding on Williamson's work, Stenner, Sanford-Moore, and Williamson (2012) aggregated the readability information across the various postsecondary options available to a high school graduate to arrive at a standard of reading needed by individuals to be considered "college and career ready." In their study, they included additional citizenship materials beyond those examined by Williamson (e.g., national and international newspapers and other adult reading materials such as Wikipedia articles). Using a weighted mean of the medians for each of the postsecondary options

[1] Reprinted from Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19*(4), 602-632.

(education, military, work place, and citizenship), a measure of 1300L was defined as the general reading demand for postsecondary options and could be used to judge a student's "college and career readiness."

In Texas, two studies were conducted to examine the reading demands in various postsecondary options – technical college, community college, and 4-year university programs. Under Commissioner Raymond Paredes, THECB conducted a research study in 2007 (and extended in 2008) which addressed the focal question of "how well does a student need to read to be successful in community colleges, technical colleges, and universities in Texas?" THECB staff collected a sample of books that first year students in Texas would be required to read in each setting. These books were measured in terms of their text complexity using The Lexile Framework for Reading. Since the TAKS had already been linked with Lexile measures for several years, the THECB study was able to overlay the TAKS cut scores onto the post high school reading requirements. (For a complete description of this report, please visit www.thecb.state.tx.us/index.cfm?objectid=31BFFF6B-BB41-8A43-C76A99EDA0F38B7D.)

Since the THECB study was completed, other states have followed the Texas example and used the same approach in examining the gap from high school to the postsecondary world. In 2009, a similar study was conducted for the Georgia Department of Education; and in 2010, a study was conducted for the Tennessee Department of Education. In terms of mean text demand, the results across the three states produced similar estimates of the reading ability needed in higher-education institutions: Texas, 1230L; Georgia, 1220L; and Tennessee, 1260L. When these results are incorporated with the reading demands of other postsecondary endeavors (military, citizenship, workplace, and adult reading materials [national and international newspapers] and Wikipedia articles) used by Stenner, Koons, and Swartz (2010), the college and career readiness standard for reading is 1293L. These results are based on more than 105,000,000 words from approximately 3,100 sources from the adult text space.

The question for educators becomes how to determine if a student is "on track" for college and career as previously defined in the Common Core State Standards and described above. "As state departments of education, and the districts and schools within those respective states, transition from *adopting* the new Common Core State Standards to the more difficult task of *implementing* them, the challenge now becomes how to translate these higher standards into tangible, practical and cost-effective curricula" (Smith, 2012). Implementing the Common Core will require districts and schools to develop new instructional strategies and complementary resources that are not only aligned with these national college- and career-readiness standards, but also utilize and incorporate proven and cost-effective tools that are universally accessible to all stakeholders.

The Standards for English Language Arts focus on the importance of text complexity. As stated in Standard 10, students must be able to "read and comprehend complex literary and informational texts independently and proficiently" (Common Core State Standards for English Language Arts, College and Career Readiness Anchor Standards for Reading, NGA Center and CCSSO, 2010, p.10).

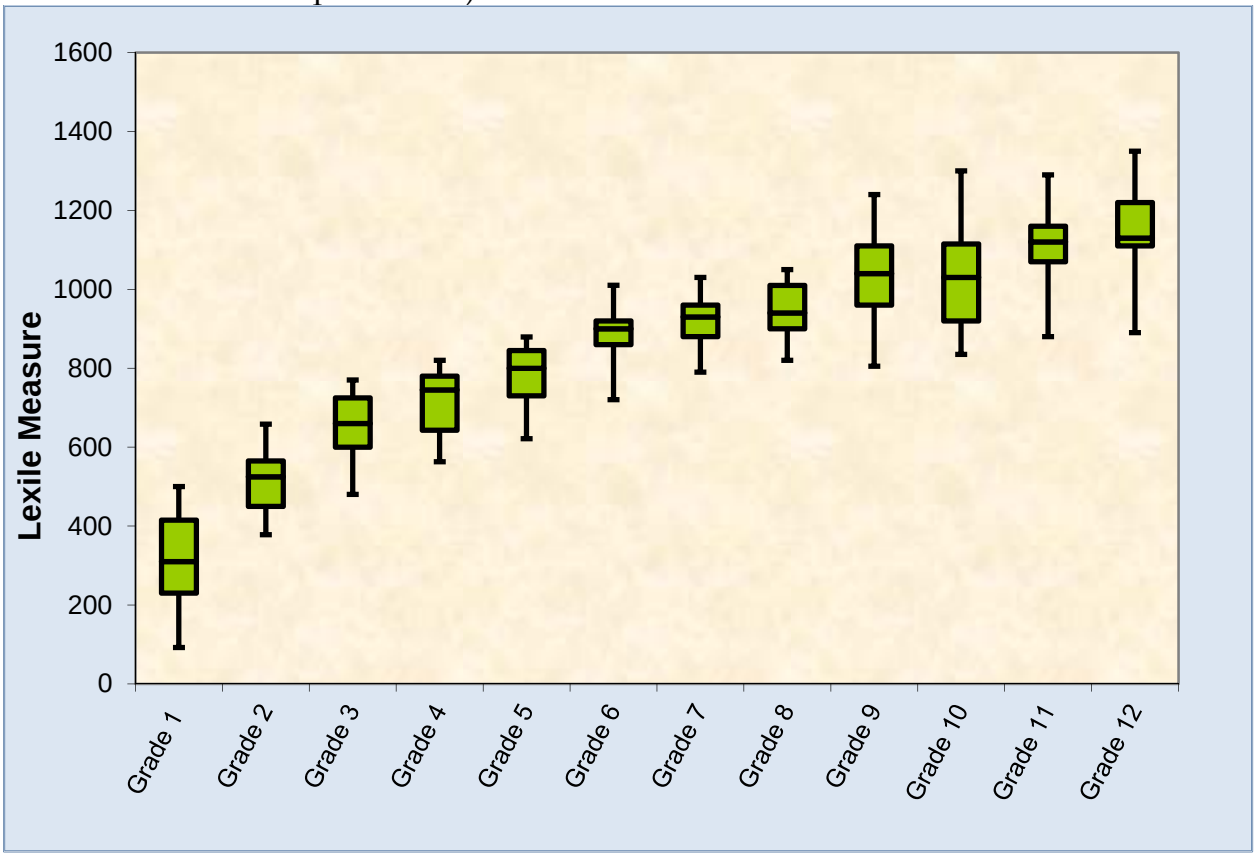
The Common Core State Standards recommends a three-part model for evaluating the complexity of a text that takes into account its qualitative dimensions, quantitative measure, and reader and task considerations. It describes text complexity as "the inherent difficulty of reading and comprehending a text combined with consideration of reader and task variables…a three-part assessment of text [complexity] that pairs qualitative and quantitative measures with reader-task considerations" (NGA Center and CCSSO, 2010, p. 43). In simpler terms, *text complexity is a transaction between text, reader, and task*. The quantitative aspect of defining text complexity consists of a stair-step progression of increasingly difficult text by grade levels (Common Core State Standards for English Language Arts, Appendix A, NGA Center and CCSSO, 2010, p. 8).

Table 21. Lexile ranges aligned to college- and career-readiness expectations, by grade.

Grade	2012 Stretch Text Measure
1	190L to 530L
2	420L to 650L
3	520L to 820L
4	740L to 940L
5	830L to 1010L
6	925L to 1070L
7	970L to 1120L
8	1010L to 1185L
9	1050L to 1260L
10	1080L to 1335L
11-12	1185L to 1385L

Between 2004 and 2008, MetaMetrics (Williamson, Koons, Sandvik, and Sanford-Moore, 2012) collected and measured textbooks across the K-12 educational continuum. The box-and-whisker plot in *Figure 4* shows the Lexile measures (*y*-axis) across grades as defined in the US. For each grade, the box refers to the interquartile range. The line within the box indicates the median. The end of each whisker shows the 5th and 95th percentile text complexity measures in the Lexile metric for each grade. This information can provide a basis for defining at what level students need to be able to read to be ready for various postsecondary endeavors such as further education beyond high school and entering the work force.

Figure 18. Text complexity distributions, in Lexile units, by grade (whiskers represent 5th and 95th percentiles).



This continuum can be "stretched" to describe the reading demands expected of students in Grades 1-12 who are "on track" for college and career (Sanford-Moore and Williamson, 2012). The quantitative aspect of defining text complexity consists of a stair-step progression of increasingly difficult text by grade levels (Common Core State Standards for English Language Arts, Appendix A, NGA Center and CCSSO, 2010, p. 8).

MetaMetrics' research on the typical reading demands of college and careers contributed to the Common Core State Standards as a whole and, more specifically, to the Lexile-based grade bands in *Figure 19*. *Figure 19* shows the relationship between the "Level 3" performance standard for each grade level established on the NC READY EOG Reading/EOC English II Assessment and the "stretch" reading demands. This shows that the NC READY EOG Reading/EOC English II performance standards for "Level 3" at each grade level is set at a level that is consistent with being "on track" for college and career readiness at the end of Grade 12.

Figure 19. Comparison of NC READY EOG Reading/EOC English II "Level 3" standards with college and career reading levels described by the CCSS.

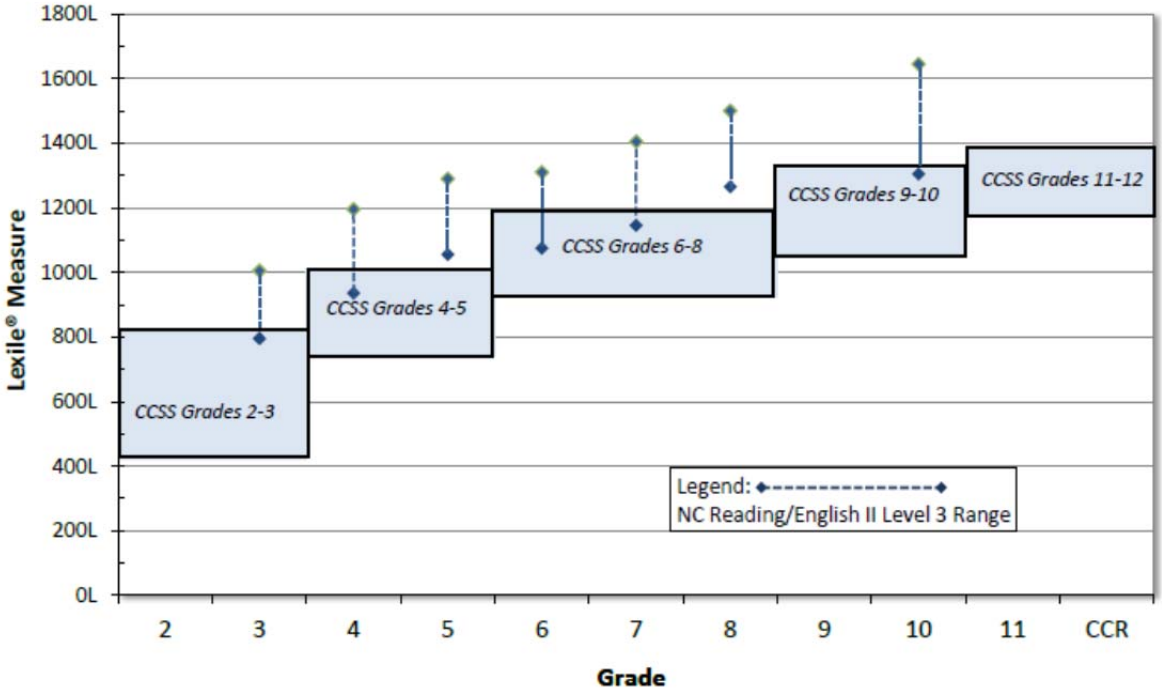
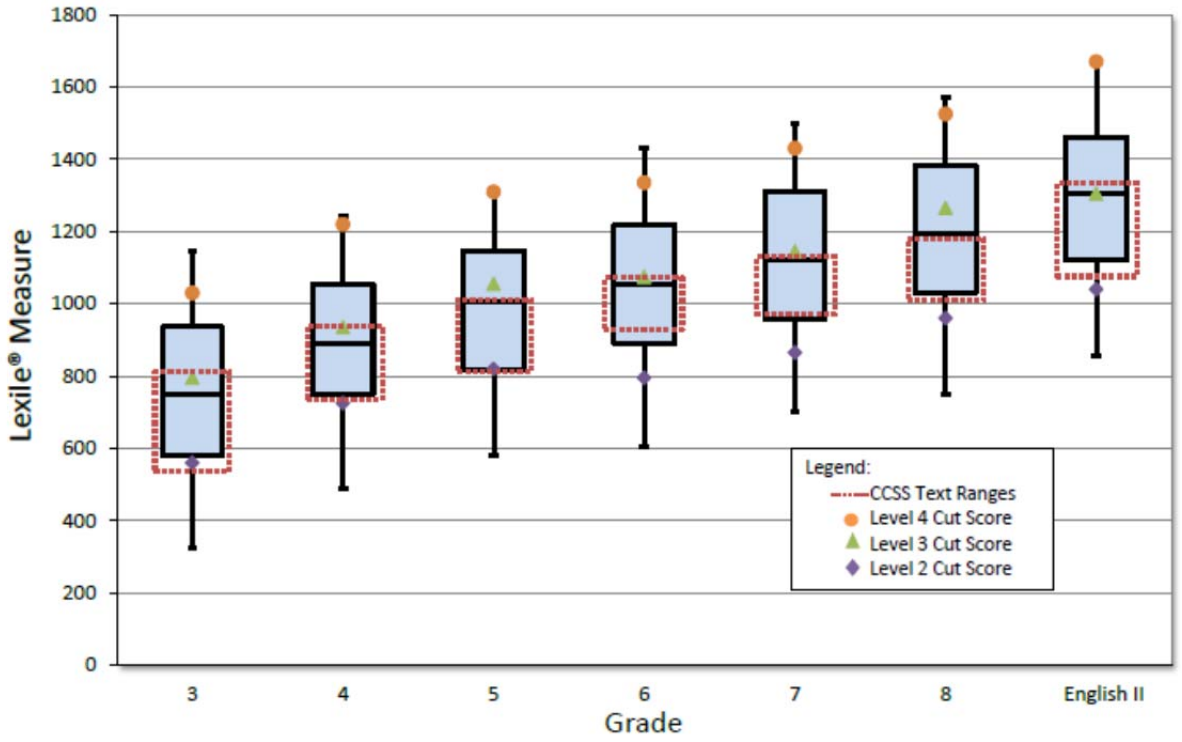


Figure 20 shows that the spring 2013 student performance on the NC READY EOG Reading/EOC English II assessments at each grade level is "on track" for college and career readiness. Students can be matched with reading materials that are at or above the recommendations in Appendix A of the CCSS for ELA for each grade level.

Figure 10. NC READY EOG Reading/EOC English II 2012-2013 student performance expressed as Lexile measures.



In 2008, MetaMetrics and the North Carolina Department of Public Instruction conducted a study to link the NCEOG Reading Test with the Lexile scale (MetaMetrics, 2008). The minimum score considered "proficient" (Level 3) at each grade level on the NCEOG Reading is presented in *Table 22*. In 2013, NCDPI transitioned their assessment program to the NC READY EOG Reading Assessment to align with the Common Core State Standards in English/Language Arts and to describe student reading performance in relation to college and career readiness. One outcome of this change was to set the performance standards for NC READY EOG Reading at a higher level. For comparison purposes, the minimum "proficient" score for the NC READY EOG Reading assessment is also repeated from *Table 17*. The Lexile scale can be used as an external "yardstick" to evaluate this change in reading demand on the North Carolina reading assessment. The information in *Table 22* shows that the NC READY EOG Reading standards are demanding more of students in terms of reading ability in 2013.

Table 22. Minimum "Level 3" Lexile measure on NCEOG Reading (2008) and NC READY EOG Reading (2013).

Grade	Proficient Level 3 Cut Score (2008)	Proficient Level 3 Cut Score (2013)
3	665L	795L
4	790L	935L
5	940L	1055L
6	990L	1075L
7	1115L	1145L
8	1165L	1265L

Next Steps. To utilize the results from this study, Lexile measures need to be incorporated into the NC READY EOG Reading/EOC English II results processing and interpretation frameworks. This information can then be used in a variety of areas within the educational system—instruction, assessment, communication to name a few.

Within the *instructional area*, suggested book lists can be developed for ranges of readers. Care must be taken to ensure that the books on the lists are also developmentally appropriate for the readers. The Lexile measure is one factor related to comprehension and is a good starting point in the selection process of a book for a specific reader. Other factors such as student developmental level, motivation, and interest; amount of background knowledge possessed by the reader; and characteristics of the text such as illustrations and formatting also need to be considered when matching a book with a reader.

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 texts and supplementary materials.

The Lexile Framework is an objective tool that can be used to determine a student's readiness for a reading experience; the Lexile Framework "targets" text (books, newspapers, periodicals) for readers at a 75-percent comprehension level—a level that is challenging, but not frustrating (Schnick and Knickelbine, 2000).

Within the *communication* area, Lexile measures can be used to communicate with students, parents, teachers, educators, and the community by providing a common language to use to talk about reading growth and development. By aligning all areas of the educational system, parents can be included in the instructional process. With a variety of data related to a student's reading level a more complete picture can be formed and more informed decisions can be made concerning reading-group placement, amount of extra instruction needed, and promotion/retention decisions.

It is much easier to understand what a national percentile rank of 50 means when it is tied to the reading demands of book titles that are familiar to adults. Parents are encouraged to help their children achieve high standards by expecting their children to succeed at school, communicating with their children's teachers and the school, and helping their children keep pace and do homework.

Through the customized reading lists and electronic database of titles, parents can assist their children in the selection of reading materials that are at the appropriate level of challenge and monitor the reading process at home. A link can be provided to the "Find a Book" website. This site provides a quick, free resource to battle "summer slide" – the learning losses that students often experience during the summer months when they are not in school. Lexile measures make it easy to help students read and learn all summer long and during the school year. This website can help build a reading list of books at a young person's reading level that are about subjects that interest him or her. This website can be viewed at http://www.lexile.com/findabook/.

In one large school district, the end-of-year testing results are sent home to parents in a folder. The folder consists of a Lexile Map on one side and a letter from the superintendent on the other side. The school district considers this type of material as "refrigerator-friendly." They encourage parents to put the Lexile Map on the refrigerator and use it to monitor and track the reading progress of their child throughout the school year.

The community-at-large (business leaders, citizens, politicians, and visitors) sees the educational system as a reflection of the community. Through the reporting of assessment results (after all, that is what the community is most interested in—results), people can understand what the community values and see the return for its investment in the schools and its children.

One way to involve the community is to work with the public libraries and local bookstores when developing reading lists. The organizations should be contacted early enough so that they can be sure that the books will be available. Often books can be displayed with their Lexile measures for easy access.

Many school districts make presentations to civic groups to educate the community as to their reading initiatives and how the Lexile Framework is being utilized in the school. Conversely, many civic groups are looking for an activity to sponsor, and it could be as simple as "donate-a-book" or "sponsor-a-reader" campaigns.

Notes

1. A T-parallel test is a test that is designed to be "theoretically parallel" to another test in that it has the same number of items/points, the same overall level of difficulty in terms of raw score means and standard deviations, and assesses the same construct domain (MetaMetrics, Inc. 1998).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bormuth, J.R. (1966). Readability: New approach. *Reading Research Quarterly*, *7*, 79-132.

Carroll, J.B., Davies, P., & Richman, B. (1971). *Word frequency book*. Boston: Houghton Mifflin.

Carver, R.P. (1974). Measuring the primary effect of reading: Reading storage technique, understanding judgments and cloze. *Journal of Reading Behavior*, *6*, 249-274.

Chall, J.S. (1988). "The beginning years." In B.L. Zakaluk and S.J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Crain, S. & Shankweiler, D. (1988). "Syntactic complexity and reading acquisition." In A. Davidson and G.M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered*. Hillsdale, NJ: Erlbaum Associates.

Davidson, A. & Kantor, R.N. (1982). On the failure of readability formulas to define readable text: A case study from adaptations. *Reading Research Quarterly*, *17*, 187-209.

Dunn, L.M. & Dunn, L.M. (1981). *Manual for Forms L and M of the Peabody Picture Vocabulary Test – Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L.M. & Markwardt, F.C. (1970). *Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service.

Efron, B. (1981). Nonparametric estimates of the standard error: The Jackknife, the Bootstrap, and other resampling techniques. *Biometrika*. *68*, 589-599.

Gorsuch, R.L. (1983). *Factor analysis*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Grolier, Inc. (1986). *The electronic encyclopedia*. Danbury, CT: Author.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Bulletin*. *9*, 139-164.

Klare, G.R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices*. 2nd edition. New York: Springer Science + Business Media, LLC.

Liberman, I.Y., Mann, V.A., Shankweiler, D., & Westelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to reading ability. *Cortex*, *18*, 367-375.

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.

McGraw-Hill Book Company. (1983). *Guidelines for bias-free publishing*. Monterey, CA: Author.

MetaMetrics, Inc. (2008). *Linking the North Carolina EOG Reading and EOC English I Tests with the Lexile Framework*. Durham, NC: Author.

Miller, G.A. & Gildea, P.M. (1987). How children learn words. *Scientific American*, *257*, 94-99.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects, Appendix A. Washington, DC: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2012). Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects, Revised Appendix A. Washington, DC: Author.

National Research Council. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, D.C.: 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 Instruction. (2013c). *Common Core State Standards (CCSS) for English Language Arts: North Carolina Assessment Specifications Summary, READY EOG Assessments, Grades 3-8 READY EOC English II Assessments.* Retrieved on October 31, 2013 from http://www.ncpublicschools.org/docs/acre/assessment/ela.pdf

North Carolina Department of Instruction. (2013d). *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 Instruction. (2013e). *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* (Third Edition) (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.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.

Sanford-Moore, E., & Williamson, G. L. (2012). *Bending the text complexity curve to close the gap* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Schnick, T. & Knickelbine, M. (2000). *The Lexile Framework: An introduction for educators*. Durham, NC: MetaMetrics, Inc.

Shankweiler, D. & Crain, S. (1986). Language mechanisms and reading disorder: A modular approach. *Cognition*, *14*, 139-168.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics*, *20*(2), 135-154.

Smith, M. (2011, March 30). *Bending the reading growth trajectory: Instructional strategies to promote reading skills and close the readiness gap*. MetaMetrics Policy Brief. Durham, NC: MetaMetrics, Inc.

Smith, M. (2012, February). *Not so common: Comparing Lexile® measures with the standards' other text complexity tools*. MetaMetrics White Paper. Durham, NC: MetaMetrics, Inc.

Stenner, A.J. (1990). Objectivity: Specific and general. *Rasch Measurement Transactions*, *4*, 111.

Stenner, A. J., Sanford-Moore, E., & Williamson, G. L. (2012). *The Lexile® Framework for Reading quantifies the reading ability needed for "College & Career Readiness."* MetaMetrics Research Brief. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement*, *20*(4), 305-315.

Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987a). Fit of the Lexile Theory to item difficulties on fourteen standardized reading comprehension tests. Durham, NC: MetaMetrics, Inc.

Stenner, A.J., Smith, D.R., Horabin, I., & Smith, M. (1987b). Fit of the Lexile Theory to sequenced units from eleven basal series. Durham, NC: MetaMetrics, Inc.

Tomlinson, C.A. (1999). *The differentiated classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics, 19*(4), 602-632.

Williamson, G. L. (2011, March 16-17). *Growth and growth norms: Standards for academic growth.* Presentation at the first semi-annual meeting of MetaMetrics' Technical Advisory Committee, held at MetaMetrics, Inc., Durham, NC.

Williamson, G. L., Koons, H., Sandvik, T., & Sanford-Moore, E. (2012). *The text complexity continuum in grades 1-12* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Williamson, G. L., Thompson, C. L., and Baker, R. F. (2007, April). *North Carolina's growth in reading and mathematics.* AERA Distinguished Paper Presentation at the 2007 annual meeting of the American Educational Research Association (AERA), Chicago, IL.

Wright, B.D. & Linacre, J.M. (1994, August). *The Rasch model as a foundation for the Lexile Framework*. Unpublished manuscript.

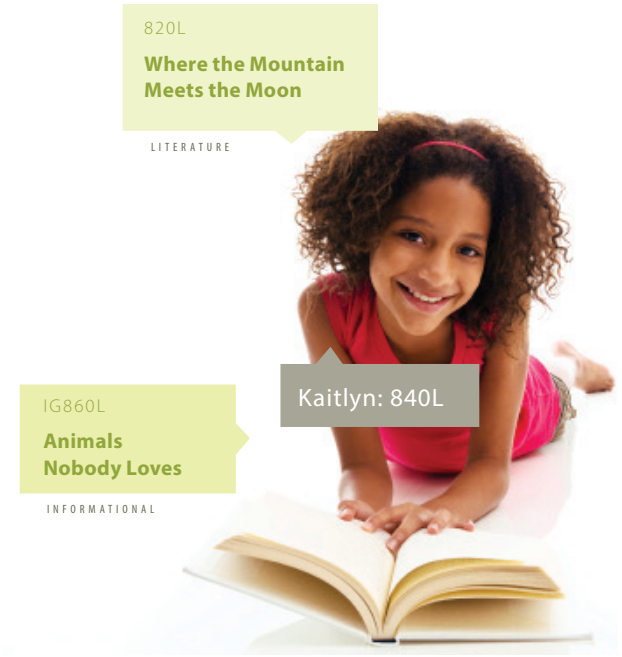
Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

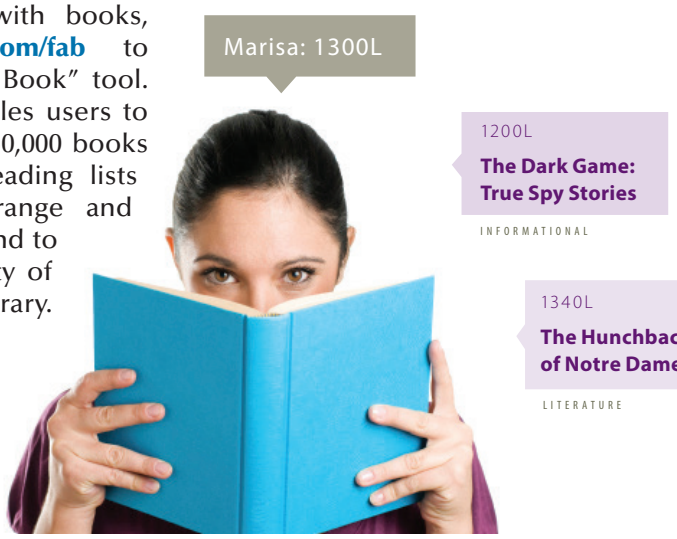
THE LEXILE® FRAMEWORK FOR READING MAP

Matching Readers with Text

Imagine getting students excited about reading while also improving their reading abilities. With the Lexile® Map, students have a chance to match books with their reading levels, and celebrate as they are able to read increasingly complex texts!

Let your students find books that fit them! Build custom book lists for your students by accessing our "Find a Book" tool at Lexile.com/fab.





HOW IT WORKS

The Lexile® Map provides examples of popular books and sample texts that are matched to various points on the Lexile® scale, from 200L for emergent reader text to 1600L for more advanced texts. The examples on the map help to define text complexity and help readers identify books of various levels of text complexity. Both literature and informational texts are presented on the Lexile Map.

HOW TO USE IT

Lexile reader and text measures can be used together to forecast how well a reader will likely comprehend a text at a specific Lexile level. A Lexile reader measure is usually obtained by having the reader take a reading comprehension test. Numerous tests report Lexile reader measures including many state end-of-year assessments, national norm-referenced assessments, and reading program assessments. A Lexile reader measure places students on the same Lexile scale as the texts. This scale ranges from below 200L to above 1600L. The Lexile website also provides a way to estimate a reader measure by using information about the reader's grade level and self-reported reading ability.

Individuals reading within their Lexile ranges (100L below to 50L above their Lexile reader measures) are likely to comprehend approximately 75 percent of the text when reading independently. This "targeted reading" rate is the point at which a reader will comprehend enough to understand the text but will also face some reading challenge. The result is growth in reading ability and a rewarding reading experience.

For more guidance concerning targeting readers with books, visit **www.Lexile.com/fab** to access the "Find a Book" tool. "Find a Book" enables users to search from over 150,000 books to build custom reading lists based on Lexile range and personal interests and to check the availability of books at the local library.

The Lexile Framework for Reading

1300 L ‣ 1500L+
LEXILE RANGE

1500L+

1500L ***Don Quixote**** CERVANTES

The Words were to me so many Pearls of Eloquence, and his Voice sweeter to my Ears than Sugar to the Taste. The Reflection on the Misfortune which these Verses brought on me, has often made me applaud Plato's Design of banishing all Poets from a good and well governed Commonwealth, especially those who write wantonly or lasciviously. For, instead of composing lamentable Verses, like those of the Marquiss of Mantua, that make Women and Children cry by the Fireside, they try their utmost Skill on such soft Strokes as enter the Soul, and wound it, like that Thunder which hurts and consumes all within, yet leaves the Garment sound. Another Time he entertained me with the following Song.

SAMPLE TITLES

LITERATURE	1640L	**The Plot Against America** (ROTH)
	1560L	**Rob Roy** (SCOTT)
	1530L	**The Good Earth** (BUCK)
	1520L	**A Fable** (FAULKNER)
	1500L	**The Decameron** (BOCCACCIO)
INFORMATIONAL	1600L	**Sustaining Life: How Human Health Depends on Biodiversity** (CHIVIAN & BERNSTEIN)
	1550L	**The Art of War** (SUN TZU)
	1560L	**The United States' Constitution**
	1520L	**Fair Play: The Ethics of Sport** (SIMON)
	1500L	**Critique of Pure Reason** (KANT)

1400L ‣ 1495L

1400L ***Nathaniel's Nutmeg*** MILTON

Setting sail once again they kept a sharp look-out for Busse Island, discovered thirty years previously by Martin Frobisher, but the rolling sea mists had grown too thick. Storms and gale—force winds plagued them for days on end and at one point grew so ferocious that the foremast cracked, splintered and was hurled into the sea. It was with considerable relief that the crew sighted through the mist the coast of Newfoundland—a vague geographical term in Hudson's day—at the beginning of July. They dropped anchor in Penobscot Bay, some one hundred miles west of Nova Scotia.

SAMPLE TITLES

LITERATURE	1460L	**The Legend of Sleepy Hollow** (IRVING)
	1450L	**Billy Budd**** (MELVILLE)
	1430L	**The Story of King Arthur and His Knights** (PYLE)
	1420L	**Life All Around Me by Ellen Foster** (GIBBONS)
	1420L	**The Scarlet Letter**** (HAWTHORNE)
INFORMATIONAL	1490L	**America's Constitution: A Biography**** (AMAR)
	1490L	**Gettysburg Address** (LINCOLN)
	1480L	**The Declaration of Independence**
	1410L	**Profiles in Courage** (KENNEDY)
	1400L	**The Life and Times of Frederick Douglass** (DOUGLASS)

1300L ‣ 1395L

1300L ***1776: America and Britain at War**** MCCULLOUGH

But from this point on, the citizen-soldiers of Washington's army were no longer to be fighting only for the defense of their country, or for their rightful liberties as freeborn Englishmen, as they had at Lexington and Concord, Bunker Hill and through the long siege at Boston. It was now a proudly proclaimed, all-out war for an independent America, a new America, and thus a new day of freedom and equality. At his home in Newport, Nathanael Greene's mentor, the Reverend Ezra Stiles, wrote in his diary almost in disbelief: Thus the Congress has tied a Gordian knot, which the Parl [iament] will find they can neither cut, nor untie. The thirteen united colonies now rise into an Independent Republic among the kingdoms, states, and empires on earth...And have I lived to see such an important and astonishing revolution?

SAMPLE TITLES

LITERATURE	1360L	**Robinson Crusoe** (DEFOE)
	1350L	**The Secret Sharer** (CONRAD)
	1340L	**The Hunchback of Notre Dame** (HUGO)
	1340L	**The Metamorphosis**** (KAFKA)
	1340L	**Fever Pitch** (HORNBY)
INFORMATIONAL	1390L	**In Defense of Food: An Eater's Manifesto** (POLLAN)
	1380L	**Politics and the English Language**** (ORWELL)
	1370L	**Jane Austen's Pride and Prejudice** (BLOOM)
	1340L	**Walden**** (THOREAU)
	1300L	**Arctic Dreams: Imagination and Desire in a Northern Landscape** (LOPEZ)

**Common Core State Standards Text Exemplar

The Lexile Framework for Reading

1000 L ‣ 1295L
LEXILE RANGE

1200L ‣ 1295L

1200L ***Why We Can't Wait*** KING

We sing the freedom songs today for the same reason the slaves sang them, because we too are in bondage and the songs add hope to our determination that "We shall overcome, Black and white together, We shall overcome someday." I have stood in a meeting with hundreds of youngsters and joined in while they sang "Ain't Gonna Let Nobody Turn Me 'Round." It is not just a song; it is a resolve. A few minutes later, I have seen those same youngsters refuse to turn around from the onrush of a police We sing the freedom songs today for the same reason the slaves sang them, because we too are in bondage and the songs add hope to our determination that "We shall overcome, Black and white together, We shall overcome someday."

SAMPLE TITLES

LITERATURE	1280L	**The House of the Spirits** (ALLENDE)
	1270L	**Tarzan of the Apes** (BURROUGHS)
	1270L	**Chronicle of a Death Foretold** (GARCÍA MÁRQUEZ)
	1220L	**Annie John** (KINCAID)
	1210L	**The Namesake**** (LAHIRI)
INFORMATIONAL	1290L	**A Brief History of Time** (HAWKING)
	1280L	**Black, Blue, and Gray: African Americans in the Civil War**** (HASKINS)
	1240L	**Blood Done Sign My Name** (TYSON)
	1230L	**Stiff: The Curious Lives of Human Cadavers** (ROACH)
	1200L	**The Dark Game: True Spy Stories** (JANECZKO)

1100L ‣ 1195L

1100L ***Pride and Prejudice**** AUSTEN

Lydia was a stout, well-grown girl of fifteen, with a fine complexion and good-humoured countenance; a favourite with her mother, whose affection had brought her into public at an early age. She had high animal spirits, and a sort of natural self-consequence, which the attentions of the officers, to whom her uncle's good dinners and her own easy manners recommended her, had increased into assurance. She was very equal therefore to address Mr. Bingley on the subject of the ball, and abruptly reminded him of his promise; adding, that it would be the most shameful thing in the world if he did not keep it. His answer to this sudden attack was delightful to their mother's ear.

SAMPLE TITLES

LITERATURE	1180L	**The Curious Incident of the Dog in the Night-time** (HADDON)
	1170L	**The Amazing Adventures of Kavalier & Clay** (CHABON)
	1150L	**A Wizard of Earthsea** (LE GUIN)
	1130L	**All the King's Men** (WARREN)
	1110L	**A Separate Peace** (KNOWLES)
INFORMATIONAL	1160L	**The Longitude Prize**** (DASH)
	1160L	**In Search of Our Mothers' Gardens** (WALKER)
	1140L	**Winterdance: The Fine Madness of Running the Iditarod** (PAULSEN)
	1130L	**The Great Fire**** (MURPHY)
	1100L	**Vincent Van Gogh: Portrait of an Artist**** (GREENBERG & JORDAN)

1000L ‣ 1095L

1000L ***Mythbusters Science Fair Book*** MARGLES

There may be less bacteria on the food that's picked up quickly, but playing it safe is the best idea. If it hits the floor, the next thing it should hit is the trash. If putting together petri dishes and dealing with incubation seems like a bigger project than you're ready to take on, there's a simpler way to observe bacterial growth. Practically all you need is some bread and your own two hands. Cut the edges off each slice of bread so that they'll fit into the plastic containers. Put one slice of bread into each container. Measure one tablespoon of water and splash it into the first piece of bread. Put the lid on the container and use your pen and tape to label this your control.

SAMPLE TITLES

LITERATURE	1080L	**I Heard the Owl Call My Name** (CRAVEN)
	1070L	**Savvy** (LAW)
	1070L	**Around the World in 80 Days** (VERNE)
	1010L	**The Pearl** (STEINBECK)
	1000L	**The Hobbit or There and Back Again** (TOLKIEN)
INFORMATIONAL	1070L	**Geeks: How Two Lost Boys Rode the Internet Out of Idaho**** (KATZ)
	1030L	**Phineas Gage** (FLEISCHMAN)
	1020L	**This Land Was Made for You and Me: The Life and Songs of Woody Guthrie** (PARTRIDGE)
	1010L	**Travels With Charley: In Search of America**** (STEINBECK)
	1000L	**Claudette Colvin: Twice Toward Justice** (HOOSE)

**Common Core State Standards Text Exemplar

The Lexile Framework for Reading

700L ▸ 995L
LEXILE RANGE

900L ▸ 995L

900L ***We are the Ship: The Story of Negro League Baseball*** NELSON

Rube ran his ball club like it was a major league team. Most Negro teams back then weren't very well organized. Didn't always have enough equipment or even matching uniforms. Most times they went from game to game scattered among different cars, or sometimes they'd even have to "hobo"—which means hitch a ride on the back of someone's truck to get to the next town for a game. But not Rube's team. They were always well equipped, with clean, new uniforms, bats, and balls. They rode to the games in fancy Pullman cars Rube rented and hitched to the back of the train. It was something to see that group of Negroes stepping out of the train, dressed in suits and hats. They were big-leaguers.

SAMPLE TITLES

LITERATURE	980L	**Dovey Coe** (DOWELL)
	950L	**Bud, Not Buddy** (CURTIS)
	940L	**Harry Potter and the Chamber of Secrets** (ROWLING)
	940L	**Heat** (LUPICA)
	900L	**City of Fire** (YEP)
INFORMATIONAL	990L	**Seabiscuit** (HILLENBRAND)
	970L	**The Kid's Guide to Money: Earning It, Saving It, Spending It, Growing It, Sharing It**** (OTFINOSKI)
	950L	**Jim Thorpe, Original All-American** (BRUCHAC)
	930L	**Colin Powell A & E Biography** (FINLAYSON)
	920L	**Talking with Artists** (CUMMINGS)

800L ▸ 895L

800L ***Moon Over Manifest*** VANDERPOOL

There wasn't much left in the tree fort from previous dwellers. Just an old hammer and a few rusted tin cans holding some even rustier nails. A couple of wood crates with the salt girl holding her umbrella painted on top. And a shabby plaque dangling sideways on one nail, FORT TREECONDEROGA. Probably named after the famous fort from Revolutionary War days. Anything else that might have been left behind had probably been weathered to bits and fallen through the cracks. No matter. I'd have this place whipped into shape lickety-split. First off, I picked out the straightest nail I could find and fixed that sign up right. Fort Treeconderoga was open for business.

SAMPLE TITLES

LITERATURE	GN840L*	**The Odyssey** (HINDS)
	830L	**Baseball in April and Other Stories** (SOTO)
	820L	**Maniac Magee** (SPINELLI)
	820L	**Where the Mountain Meets the Moon**** (LIN)
	800L	**Homeless Bird** (WHELEN)
INFORMATIONAL	880L	**The Circuit** (JIMENEZ)
	870L	**The 7 Habits of Highly Effective Teens** (COVEY)
	IG860L*	**Animals Nobody Loves** (SEYMOUR)
	860L	**Through My Eyes: Ruby Bridges** (BRIDGES)
	830L	**Quest for the Tree Kangaroo: An Expedition to the Cloud Forest of New Guinea**** (MONTGOMERY)

700L ▸ 795L

700L ***The Miraculous Journey of Edward Tulane*** DICAMILLO

Edward, for lack of anything better to do, began to think. He thought about the stars. He remembered what they looked like from his bedroom window. What made them shine so brightly, he wondered, and were they still shining somewhere even though he could not see them? Never in my life, he thought, have I been farther away from the stars than I am now. He considered, too, the fate of the beautiful princess who had become a warthog. Why had she become a warthog? Because the ugly witch turned her into one-that was why. And then the rabbit thought about Pellegrina. He felt, in some way that he could not explain to himself, that she was responsible for what had happened to him. It was almost as if it was she, and not the boys, who had thrown Edward overboard.

SAMPLE TITLES

LITERATURE	770L	**Walk Two Moons** (CREECH)
	760L	**Hoot** (HIAASEN)
	750L	**Esperanza Rising** (RYAN)
	720L	**Nancy's Mysterious Letter** (KEENE)
	GN720L*	**Sherlock Holmes and the Adventure at the Copper Beeches** (DOYLE)
INFORMATIONAL	790L	**Be Water, My Friend: The Early Years of Bruce Lee** (MOCHIZUKI)
	760L	**Stay: The True Story of Ten Dogs** (MUNTEAN)
	IG760L*	**Mapping Shipwrecks with Coordinate Planes** (WALL)
	720L	**Pretty in Print: Questioning Magazines** (BOTZAKIS)
	720L	**Spiders in the Hairdo: Modern Urban Legends** (HOLT & MOONEY)

*GN denotes Graphic Novel, IG denotes Illustrated Guide
**Common Core State Standards Text Exemplar

The Lexile Framework for Reading

400L ▸ 695L
LEXILE RANGE

600L ▸ 695L

600L ***You're on Your Way, Teddy Roosevelt*** ST. GEORGE & FAULKNER

But from his first workout in Wood's Gymnasium he had been determined to control his asthma and illnesses rather than letting his asthma and illnesses control him. And he had. On that hot summer day in August he had proved to himself—and everyone else—that he had taken charge of his own life. In 1876 Teedie—now known as Teddy—entered Harvard College. He was on his own ...without Papa. That was all right. "I am to do everything for myself," he wrote in his diary. Why not? He was stronger and in better health than he had ever been. And ready and eager for the adventures and opportunities that lay ahead.

SAMPLE TITLES

LITERATURE	680L	**Charlotte's Web** (WHITE)
	660L	**Holes** (SACHAR)
	620L	**M.C. Higgins, the Great**** (HAMILTON)
	610L	**Mountain Bike Mania** (CHRISTOPHER)
	610L	**A Year Down Yonder** (PECK)
INFORMATIONAL	690L	**Where Do Polar Bears Live?**** (THOMSON)
	680L	**An Eye for Color: The Story of Josef Albers** (WING)
	660L	**Remember: The Journey to School Integration** (MORRISON)
	660L	**From Seed to Plant**** (GIBBONS)
	630L	**Sadako and the Thousand Paper Cranes** (COERR)

500L ▸ 595L

500L ***A Germ's Journey*** ROOKE

Excuse me! Let's blow out of this place! In real life, germs are very small. They can't be seen without a microscope. Rudy forgot to use a tissue. His cold germs fly across the room at more than 100 miles an hour. Whee! I can fly! Best ride ever! A few germs land on Ernie. But skin acts like a suit of armor. It protects against harm. The germs won't find a new home there. Healthy skin keeps germs out. But germs can sneak into the body through cuts, scrapes, or cracks in the skin. Most germs enter through a person's mouth or nose. Rudy's germs continue to fall on nearly everything in the room—including Brenda's candy.

SAMPLE TITLES

LITERATURE	560L	**Sarah, Plain and Tall** (MACLACHLAN)
	530L	**It's All Greek to Me** (SCIESZKA)
	520L	**John Henry: An American Legend** (KEATS)
	500L	**Judy Moody Saves the World** (MCDONALD)
	500L	**The Curse of the Cheese Pyramid** (STILTON)
INFORMATIONAL	IG590L*	**Claude Monet** (CONNOLLY)
	560L	**Lemons and Lemonade: A Book about Supply and Demand** (LOEWEN)
	560L	**Molly the Pony** (KASTER)
	530L	**Langston Hughes: Great American Poet** (MCKISSACK)
	510L	**A Picture for Marc** (KIMMEL)

400L ▸ 495L

400L ***How Not to Babysit Your Brother*** HAPKA

I continued to search. I checked under Steve's bed. Then I checked under my bed. I searched the basement, the garage, and my closet. There was no sign of Steve. This was going to be harder than I thought. Where was Steve hiding? CRASH! Uh-oh, I thought. I heard Buster barking in the kitchen. I ran to see what was going on. When I got there, the dog food bin was tipped over. Steve's head and shoulders were sticking out of the top. Dog food was stuck in his hair, on his clothes, and up his nose. He looked like an alien from the planet Yuck. He giggled as Buster licked some crumbs off his ear.

SAMPLE TITLES

LITERATURE	460L	**Chrysanthemum** (HENKES)
	410L	**The Enormous Crocodile** (DAHL)
	GN400L*	**Pilot And Huxley** (MCGUINESS)
	400L	**The Fire Cat**** (AVERILL)
	400L	**Cowgirl Kate and Cocoa**** (SILVERMAN)
INFORMATIONAL	480L	**Martin Luther King, Jr. and the March on Washington**** (RUFFIN)
	460L	**True Life Treasure Hunts** (DONNELLY)
	460L	**Half You Heard of Fractions?** (ADAMSON)
	420L	**Rally for Recycling** (BULLARD)
	400L	**Animals in Winter** (RUSTAD)

*GN denotes Graphic Novel, IG denotes Illustrated Guide
**Common Core State Standards Text Exemplar

The Lexile Framework for Reading

200L ▸ 395L LEXILE RANGE

300L ▸ 395L

300L ***Princess Posey and the Next-Door Dog*** GREENE

"We have to stop now," said Miss Lee. "It's time for reading." "Ohhh..." A disappointed sound went up around the circle. "Here's what we'll do." Miss Lee stood up. "You are all very interested in dogs. So this week, you can write a story about your own dog or pet. Then you can read it to the class." Everyone got excited again. Except Posey. She didn't have a pet. Not a dog. Not a cat. Not a hamster. "Those of you who don't have a pet," Miss Lee said, "can write about the pet you hope to own someday." Miss Lee had saved the day! Now Posey had something to write about, too. Posey told her mom about Luca's puppy on the way home.

SAMPLE TITLES

LITERATURE	380L	**Martha Bakes a Cake** (BARSS)
	380L	**Junie B. Jones is (Almost) a Flower Girl** (PARK)
	360L	**Poppleton in Winter**** (RYLANT)
	340L	**Never Swipe a Bully's Bear** (APPLEGATE)
	330L	**Frog and Toad Together**** (LOBEL)
INFORMATIONAL	GN380L*	**BMX Blitz** (CIENCIN)
	380L	**Lemonade for Sale** (MURPHY)
	350L	**A Snowy Day** (SCHAEFER)
	330L	**Freedom River** (RAPPAPORT)
	300L	**From Tree to Paper** (MARSHALL)

200L ▸ 295L

200L ***Ronald Morgan Goes to Bat*** GIFF

He smacked the ball with the bat. The ball flew across the field. "Good;' said Mr. Spano. "Great, Slugger!" I yelled. ''We'll win every game. It was my turn next. I put on the helmet, and stood at home plate. "Ronald Morgan," said Rosemary. "You're holding the wrong end of the bat." Quickly I turned it around. I clutched it close to the end. Whoosh went the first ball. Whoosh went the second one. Wham went the third. It hit me in the knee. "Are you all right?" asked Michael. But I heard Tom say, "I knew it. Ronald Morgan's the worst." At snack time, we told Miss Tyler about the team.

SAMPLE TITLES

LITERATURE	280L	**Hi! Fly Guy**** (ARNOLD)
	260L	**The Cat in the Hat** (SEUSS)
	GN240L*	**Lunch Lady and the Cyborg Substitute** (KROSOCZKA)
	200L	**Dixie** (GILMAN)
	200L	**The Best Bug Parade** (MURPHY)
INFORMATIONAL	290L	**The Story of Pocahontas** (JENNER)
	250L	**Math in the Kitchen** (AMATO)
	230L	**What makes Day and Night** (BRANLEY)
	220L	**I Love Trains!** (STURGES)
	210L	**Sharks!** (CLARKE)

*GN denotes Graphic Novel

**Common Core State Standards Text Exemplar

Please note:

The Lexile measure (text complexity) of a book is an excellent starting point for a student's book selection. It's important to understand that the book's Lexile measure should not be the only factor in a student's book selection process. Lexile measures do not consider factors such as age-appropriateness, interest, and prior knowledge. These are also key factors when matching children and adolescents with books they might like and are able to read.

MetaMetrics
LINKING ASSESSMENT WITH INSTRUCTION

Lexile codes provide more information about developmental appropriateness, reading difficulty, and common or intended usage of books. For more information on Lexile codes, please visit Lexile.com.

LEXILE TEXT RANGES TO GUIDE READING FOR COLLEGE AND CAREER READINESS

GRADES	CCSS LEXILE TEXT RANGE
11–12	1185L–1385L
9–10	1050L–1335L
6–8	925L–1185L
4–5	740L–1010L
2–3	420L–820L
1	190L–530L

COMMON CORE STATE STANDARDS FOR ENGLISH LANGUAGE ARTS, APPENDIX A (ADDITIONAL INFORMATION), NGA AND CCSSO, 2012