

North Carolina Reading Comprehension Tests

Technical Report
(CITABLE DRAFT)

Grade 3 Reading Comprehension Pretest

End-of-Grade Reading Comprehension Tests

High School Comprehensive Test

English I

Initial draft prepared by:

Mildred Bazemore, Section Chief, Test Development, NCDPI
Pamela B. Van Dyk, Ph.D.
September 2004

In compliance with federal law, including the provisions of Title IX of the Education Amendments of 1972, NC Public Schools administers all state-operated educational programs, employment activities and admissions without discrimination because of race, national or ethnic origin, color, age, military service, disability, or gender, except where exemption is appropriate and allowed by law.

Inquiries or complaints should be directed to:

The Office of Curriculum and School Reform Services

6307 Mail Service Center

Raleigh, NC 27699-6307

919-807-3761 (phone); 919-807-3767 (fax)

Table of Contents

Chapter One: Introduction	9
1.1 Local Participation	9
1.2 The North Carolina Testing Program	10
1.3 The North Carolina Reading Tests	11
Chapter Two: Test Development	13
2.1 Test Development Process for the North Carolina Testing Program	13
2.2 Curriculum Connection	15
2.3 Item Writing	27
2.4 Test Specifications	28
2.5 Selecting and Training Item Writers	29
2.6 Reviewing Items for Field Testing	29
2.7 Assembling Field-Test Forms	30
2.8 Sampling Procedures.....	31
2.9 Item Analysis and Selection	31
2.10 Classical Measurement Theory	32
2.11 Item Response Theory	32
2.12 Three-Parameter Logistic Model (3PL)	33
2.13 Sensitivity Analysis	34
2.14 Criteria for Inclusion in Item Pools	35
2.15 Item Parameter Estimates	36
2.16 Bias Review Committee	36
2.17 Operational Test Construction	37
2.18 Setting the Target p-value for Operational Tests	37

2.19 Setting the Test Administration Time	38
2.20 Reviewing Assembled Operational Tests	38
Chapter Three Test Administration	40
3.1 Test Administration	40
3.2 Training for Administrators	40
3.3 Preparation for Test Administration	41
3.4 Test Security and Handling Materials	41
3.5 Student Participation	42
3.6 Alternate Assessments	43
3.7 Testing Accommodations	43
3.8 Students with Limited English Proficiency	43
3.9 Medical Exclusions	45
3.10 Reporting Student Scores	45
3.11 Confidentiality of Student Test Scores	45
Chapter Four: Scaling and Standard Setting for the EOG and EOC Tests of Reading Comprehension	47
4.1 Conversion of Test Scores	47
4.2 Constructing a Developmental Scale	47
4.3 Setting the Standards for the North Carolina Reading Comprehension Tests	49
4.4 Achievement Level Descriptors	50
4.5 Achievement Level Cut Scores	51
4.6 Achievement Level Trends	51
4.7 Percentile Ranking	53
Chapter Five: Reports	54
5.1 Use of Test Score Reports Provided by the North Carolina Testing Program	54

5.2 Reporting by Student	54
5.3 Reporting by School	54
5.4 Reporting by State	55
Chapter Six: Descriptive Statistics and Reliability	56
6.1 Means and Standard Deviations for the First Operational Administration	56
6.2 Population Demographics for the First Operational Administration	56
6.3 Scale Score Frequency Distributions	57
6.4 Reliability of the North Carolina Reading Tests	62
6.5 Internal Consistency of the North Carolina Reading Tests	62
6.6 Standard Error of Measurement	63
Chapter Seven: Evidence of Validity	64
7.1 Evidence of Validity	64
7.2 Content Validity	64
7.3 Criterion-Related Validity	66
Chapter Eight: Quality Control	68
8.1 Quality Control Prior to Test Administration	68
8.2 Quality Control in Data Preparation and Test Administration	68
8.3 Quality Control in Data Input	69
8.4 Quality Control of Test Scores	69
8.5 Quality Control in Reporting	69
Glossary of Key Terms	70
References	74
Appendix A: Item-Writing Guidelines	76
Appendix B: Scale Score Frequency Distribution Tables	78

Appendix C: Developmental Scale Report 96
Appendix D: Testing Code of Ethics 105

List of Tables

Table 1: English Language Arts goals and objectives linked to the four categories for Grade 3

Table 2: English Language Arts goals and objectives linked to the four categories for Grade 4

Table 3: English Language Arts goals and objectives linked to the four categories for Grade 5

Table 4: English Language Arts goals and objectives linked to the four categories for Grade 6

Table 5: English Language Arts goals and objectives linked to the four categories for Grade 7

Table 6: English Language Arts goals and objectives linked to the four categories for Grade 8

Table 7: Breakdown of the number of forms, number of items per form, and number of pages reading averaged across forms for the Spring 2002 field test

Table 8: Field test sample characteristics

Table 9: Average item pool parameter estimates for the North Carolina Reading Tests

Table 10: Comparison of p-values of item pools with p-values of assembled forms

Table 11: Number of items per test and time allotted by grade

Table 12: Population means and standard deviations from the Spring 2002 item calibration for the North Carolina Reading Tests (n=1400 per form per grade)

Table 13: Comparison of population means and standard deviations for the first and second editions of the North Carolina Reading Tests

Table 14: Hypothetical percentages of contrasting-groups classifications

Table 15: Administrative Procedures Act 16 NCAC 6D .0501 (Definitions related to Student Accountability Standards)

Table 16: North Carolina Reading Tests achievement levels and corresponding scale scores

- Table 17: Achievement level trends for Grade 3
- Table 18: Achievement level trends for Grade 4
- Table 19: Achievement level trends for Grade 5
- Table 20: Achievement level trends for Grade 6
- Table 21: Achievement level trends for Grade 7
- Table 22: Achievement level trends for Grade 8
- Table 23: Descriptive statistics by grade for the first operational administration of the North Carolina Reading Tests
- Table 24: Population demographics for the first operational administration of the North Carolina Reading Tests
- Table 25: Reliability indices for the North Carolina Reading Tests
- Table 26: Ranges of standard error of measurement for scale scores by grade
- Table 27: Grade 3 Reading Test specifications
- Table 28: Grade 4 Reading Test specifications
- Table 29: Grade 5 Reading Test specifications
- Table 30: Grade 6 Reading Test specifications
- Table 31: Grade 7 Reading Test specifications
- Table 32: Grade 8 Reading Test specifications
- Table 33: Pearson Correlation Coefficients for the North Carolina Reading Tests (Grade 3 Pretest and Grades 3-8)
- Table 34: Pearson Correlation Coefficients for English I and the High School Comprehensive Reading Test

List of Figures

- Figure 1: Flow chart of the test development process used in the development of North Carolina tests
- Figure 2: Thinking skills framework used to develop items in the North Carolina Testing Program
- Figure 3: Typical item characteristic curve (ICC) for a typical 4-option multiple-choice item
- Figure 4: Scale score frequency distribution for the 2003 Grade 3 Reading Pretest
- Figure 5: Scale score frequency distribution for the 2003 Grade 3 Reading Test
- Figure 6: Scale score frequency distribution for the 2003 Grade 4 Reading Test
- Figure 7: Scale score frequency distribution for the 2003 Grade 5 Reading Test
- Figure 8: Scale score frequency distribution for the 2003 Grade 6 Reading Test
- Figure 9: Scale score frequency distribution for the 2003 Grade 7 Reading Test
- Figure 10: Scale score frequency distribution for the 2003 Grade 8 Reading Test
- Figure 11: Scale score frequency distribution for the 1998 High School Comprehensive Reading Test
- Figure 12: Scale score frequency distribution for the 2003 English I Reading Test

Chapter One: Introduction

The General Assembly believes that all children can learn. It is the intent of the General Assembly that the mission of the public school community is to challenge with high expectations each child to learn, to achieve, and to fulfill his or her potential(G.S. 115C-105.20a).

With that mission as its guide, the State Board of Education implemented the ABCs Accountability Program at grades K–8 effective with the 1996–1997 school year and grades 9–12 effective during the 1997–1998 school year to test students’ mastery of basic skills (reading, writing, and mathematics). The ABCs Accountability Program was developed under the *Public School Laws* mandating local participation in the program, the design of annual academic achievement standards, and the development of student academic achievement standards.

1.1 Local Participation

The School-Based Management and Accountability Program shall be based upon an accountability, recognition, assistance, and intervention process in order to hold each school and the school’s personnel accountable for improved student performance in the school (G.S. 115C-105.21c).

Schools are held accountable for student’s learning by reporting student performance results on North Carolina tests. Student’s scores are compiled each year and released in a report card. Schools are then recognized for the performance of their students. Schools that consistently do not make adequate progress may receive intervention from the state.

In April 1999, the State Board of Education unanimously approved Statewide Student Accountability Standards. These standards provide four Gateway Standards for student performance at grades 3, 5, 8, and 11. Students in the 3rd, 5th, and 8th grades are required to demonstrate grade-level performance in reading, writing (5th and 8th grades only), and mathematics in order to be promoted to the next grade. The law regarding student academic performance states:

The State Board of Education shall develop a plan to create rigorous student academic performance standards for kindergarten through eighth grade and student academic standards for courses in grades 9-12. The performance standards shall align, whenever possible, with the student academic performance standards developed for the National Assessment of Educational Progress (NAEP). The plan also shall include clear and understandable methods of reporting individual student academic performance to parents (G.S 115C-105.40).

1.2 The North Carolina Testing Program

The North Carolina Testing Program was designed to measure the extent to which students satisfy academic performance requirements. Tests developed by the North Carolina Department of Public Instruction Test Development Section, when properly administered and interpreted, provide reliable and valid information that 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 our students compare with students from other states; and
- Citizens to assess the performance of the public schools (North Carolina *Testing Code of Ethics*, 1997, revised 2000).

The North Carolina Testing Program was initiated in response to legislation passed by the North Carolina General Assembly. The following selection from *Public School Laws* (1994) describes the legislation. *Public School Law 115C-174.10* states the following purposes of the North Carolina Testing Program:

(1) to assure that all high school graduates possess the...skills and knowledge thought necessary to function as a member of society; (2) to provide a means of identifying strengths and weaknesses in the education process; and (3) to establish additional means for making the education system accountable to the public for results.

Tests included in the North Carolina Testing Program are designed for use as federal, state, and local indicators of student performance. Interpretation of test scores in the North Carolina Testing Program provides information about a student's performance on the test in percentiles, scale scores, and achievement levels. Percentiles provide an indicator of how a child performs relative to other children who took the test in the norming year, or the first year the test was administered. Percentiles range from 1 to 99. A percentile rank of 69 indicates that a child performed equal to or better than 69 percent of the children who took the test during the norming year.

Scale scores are derived from a raw score or "number right" score for the test. Each test has a translation table that provides a scale score for each raw test score. Scale scores are reported by achievement levels, which are predetermined academic achievement standards. The four achievement levels for the North Carolina Testing Program are:

Level I: Students performing at this level do not have sufficient mastery of knowledge and skills in a particular subject area to be successful at the next grade level.

Level II: Students performing at this level demonstrate inconsistent mastery of knowledge and skills in the subject area and are minimally prepared to be successful at the next grade level.

Level III: Students performing at this level consistently demonstrate mastery of the grade level subject matter and skills and are well prepared for the next grade.

Level IV: Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level.

The North Carolina End-of-Grade (EOG) Tests include multiple-choice assessments of reading comprehension and mathematics in grades 3 through 8 and 10. The North Carolina End-of-Course (EOC) Tests include multiple-choice assessments of reading comprehension and mathematics in English I, Algebra I, Geometry, and Algebra II. In addition to the reading comprehension and mathematics tests, the North Carolina Testing Program includes science EOC tests (Biology, Chemistry, Physical Science, Physics), social studies EOC tests which are currently under revision (Civics and Economics, U.S. History), writing assessments in grades 4, 7, and 10, the North Carolina Tests of Computer Skills, the North Carolina Competency Test, and two alternate assessments (North Carolina Alternate Assessment Academic Inventory and the North Carolina Alternate Assessment Portfolio).

The EOG reading comprehension and mathematics tests are used to monitor growth and student performance against absolute standards (performance composite) for school accountability. A student's EOG scores from the prior grade are used to determine his or her entering level of knowledge and skills and to determine the amount of growth during one school year. Beginning in 1996, a student's growth at grade 3 was determined by comparing the grade 3 EOG score with a grade 3 pretest administered during the first three weeks of the school year. The Grade Level Proficiency Guidelines, approved by the State Board of Education (February, 1995), established Level III (of those achievement levels listed above) as the standard for each grade level. The EOC tests measure a student's mastery of course-level material.

1.3 The North Carolina Reading Tests

In 1999, the State Board of Education adopted a new curriculum for English Language Arts. In response to that curriculum shift, a revised measure of accountability for students' mastery of English Language Arts was designed. These tests include the Grade 3 Reading Pretest, the End-of-Grade (EOG) Reading Comprehension Tests in grades

three through eight, the North Carolina High School Comprehensive Reading Test in grade 10, and the English I End-of-Course Reading Test.

The purpose of this document is to provide an overview and technical documentation specifically for the North Carolina Reading Tests. Part One provides an overview of the test administration. Part Two describes the test development process. Part Three describes the process for scaling the tests and setting the standards. Part Four provides the descriptive statistics for the field-test population, reliability coefficients, and evidence of validity.

Chapter Two: Test Development

2.1 Test Development Process for the North Carolina Testing Program

In June of 2003, the State Board of Education codified the process used in develop all tests in the North Carolina Testing Program. The development of tests for the North Carolina Testing Program follows a prescribed sequence of events. A flow chart of those events is found in figure 1.

Figure 1: Flow Chart of the Test Development Process used in Development of North Carolina Tests

Curriculum Adoption	Step 7 Review Item Tryout Statistics	Step 14^b Conduct Bias Reviews
Step 1^a Develop Test Specifications (Blueprint)	Step 8^b Develop New Items	Step 15 Assemble Equivalent and Parallel Forms
Step 2^b Develop Test Items	Step 9^b Review Items for Field Test	Step 16^b Review Assembled Test
Step 3^b Review Items for Tryouts	Step 10 Assemble Field Test Forms	Step 17 Final Review of Test
Step 4 Assemble Item Tryout Forms	Step 11^b Review Field Test Forms	Step 18^{ab} Administer Test as Pilot
Step 5^b Review Item Tryout Forms	Step 12^b Administer Field Test	Step 19 Score Test
Step 6^b Administer Item Tryouts	Step 13 Review Field Test Statistics	Step 20^{ab} Establish Standards
		Step 21^b Administer Test as Fully Operational
		Step 22 Report Test Results

^aActivities done only at implementation of new curriculum

^bActivities involving NC teachers

Phase 1 (step 1) requires 4 months
 Phase 2 (steps 2-7) requires 12 months
 Phase 3 (steps 8-14) requires 20 months
 Phase 4 (steps 15-20) requires 4 months for EOC and 9 months for EOG
 Phase 5 (step 21) requires 4 months
 Phase 6 (step 22) requires 1 month
 TOTAL 44-49 months

NOTES: Whenever possible, item tryouts should precede field-testing items. Professional development opportunities are integral and ongoing to the curriculum and test development process.

2.2 The Curriculum Connection

Testing of North Carolina students' reading comprehension skills relative to the English Language Arts competency goals and objectives in the *Standard Course of Study (SCS)* is one component of the North Carolina Testing Program. Students are tested in English Language Arts at the end of grades three through eight and at the end of the English I course. In grades three through eight, English Language Arts concepts are measured in four cognitive constructs: cognition, interpretation, critical stance and connections. The four categories are operationally defined below.

- Cognition

Cognition is the initial stage of a reader understanding a reading selection. It is focused on the purpose and organization of the selection. It considers the text as a whole or in broad perspective and includes strategies like using context clues to determine meaning or summarizing to include main points.

- Interpretation

Interpretation requires the student to develop a more complete understanding. It may ask students to clarify, to explain the significance of, to extend, and/or to adapt ideas/concepts.

- Critical Stance

Critical stance refers to tasks that ask a student to consider the selection objectively. It involves processes like comparing/contrasting and understanding the impact of literary elements.

- Connections

Connections refer to connecting knowledge obtained from reading the selection with other information and experiences. It involves the student being able to relate the selection to events outside of the selection. In addition, the student will make associations outside the selection and between selections.

In addition to measuring a particular category, each item on the North Carolina Grade 3 Reading Pretest and End-of-Grade Reading Tests is aligned to an objective from the NC *SCS* for English Language Arts. While some objectives can be measured readily by multiple-choice questions and are assessed by the tests, other objectives address the skills and background knowledge that are needed to do well on the tests, but are not easily measured in a multiple-choice format. To facilitate an understanding of the link between the objectives in the NC *SCS* for English Language Arts – Grades 3-8 and the individual categories, each objective in grades 3-8 is listed below. Beside each objective, the categories are indicated as follows. An illustration of the link between the four categories around which test items are developed and the NC *SCS* is provided below in Tables 1-6.

Table 1: English Language Arts goals and objectives linked to the four categories for Grade 3

English Language Arts, Grade 3 Objectives from Goals 1, 2, and 3				
<p>α The objective addresses skills or concepts related to a particular category and can be directly tested by a multiple-choice question.</p> <p>☆ The objective addresses skills or concepts related to a particular category that students may apply when answering a multiple-choice question, but the objective is not directly tested on the competency test.</p> <p>□ (empty box) The objective addresses skills and concepts that are not directly related to the particular category and is not directly tested on the competency test.</p>	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will apply enabling strategies and skills to read and write.</i>				
1.01 Apply phonics and structural analysis to decode words	☆			
1.02 Apply meanings of common prefixes and suffixes to decode words in text to assist comprehension.	α			
1.03 Integrate prior experiences and all sources of information in the text when reading orally and silently.	α	☆	☆	☆
1.04 Increase sight vocabulary, reading vocabulary, and writing vocabulary...	☆			
1.05 Use word reference materials to confirm decoding skill, verify spelling, and extend meanings of words.	α			
1.06 Read independently daily from self-selected materials.	☆	☆	☆	☆
<i>Competency Goal 2: The learner will apply strategies and skills to comprehend text that is read, heard, and viewed.</i>				
2.01 Use metacognitive strategies to comprehend text.	α	☆	α	☆
2.02 Interact with the text before, during, and after reading, listening, or viewing by setting a purpose, previewing the text, making predictions, asking questions, locating information for specific purposes, making connections, and using story structure and text organization to comprehend.	α	α	☆	α
2.03 Read a variety of texts including fiction, nonfiction, poetry, and drama.	☆	☆	☆	☆
2.04 Identify and interpret elements of fiction and nonfiction and support by referencing the text to determine the author's purpose, plot, conflict, sequence, resolution, lesson and/or message, main idea and supporting details, cause and effect, fact and opinion, point of view, and author's use of	α	α	α	

figurative language.				
2.05 Draw conclusions, make generalizations, and gather support by referencing the text.		α	α	α
2.06 Summarize the main idea (s) from texts using succinct language.	α			
2.07 Explain choice of reading materials congruent with purposes.	☆		☆	
2.08 Listen actively by facing the speaker, making eye contact, and asking questions.	☆	☆	☆	☆
<i>Competency Goal 3: The learner will make connections through the use of oral language, written language, and media and technology.</i>				
3.01 Respond to fiction, nonfiction, poetry, and drama using interpretive, critical, and evaluative processes.	α	α	α	α
3.02 Identify and discuss similarities and differences in events and characters within and across selections and support them by referencing the text.			α	α
3.03 Use text and own experiences to verify facts, concepts, and ideas.	α	α	α	α
3.04 Make informed judgments about television productions.		☆	☆	
3.05 Compare and contrast printed and visual information (graphs, charts, maps).		☆	α	α
3.06 Conduct research for assigned and self-selected projects.	☆	☆	☆	☆

Table 2: English Language Arts goals and objectives linked to the four categories for Grade 4.

English Language Arts, Grade 4 Objectives from Goals 1, 2, and 3	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will apply enabling strategies and skills to read and write.</i>				
1.01 Use word identification strategies appropriately and automatically when encountering unknown words.	★			
1.02 Infer word meanings from taught roots, prefixes, and suffixes to decode words in text to assist comprehension.	α			
1.03 Identify key words and discover their meanings and relationships through a variety of strategies.	α			
1.04 Increase reading and writing vocabulary.	★			
1.05 Use word reference materials to identify and comprehend unknown words.	α			
1.06 Read independently daily from self-selected materials.	★	★	★	★
<i>Competency Goal 2: The learner will apply strategies and skills to comprehend text that is read, heard, and viewed.</i>				
2.01 Use metacognitive strategies to comprehend text and to clarify meaning of vocabulary.	α	★	α	★
2.02 Interact with the text before, during, and after reading, listening, or viewing by setting a purpose using prior knowledge and text information, making predictions, formulating questions, locating relevant information, and making connections with previous experiences, information, and ideas.	α	α	α	α
2.03 Read a variety of texts including fiction, nonfiction, poetry, and drama.	★	★	★	★
2.04 Identify and interpret elements of fiction and nonfiction and support by referencing the text to determine the plot, theme, main idea and author’s choice of words.	α		α	
2.05 Make inferences, draw conclusions, make generalizations, and support by referencing the text.		α		
2.06 Summarize major points fiction and nonfiction text(s) to clarify and retain information and ideas.	α			
2.07 Determine usefulness of information and ideas consistent with purpose.	α		α	
2.08 Verify the meaning or accuracy of the author’s statement(s) by referencing the text or other resources.	★	α	α	★
2.09 Listen actively.	★	★	★	★

<i>Competency Goal 3: The learner will make connections through the use of oral language, written language, and media and technology.</i>				
3.01 Respond to fiction, nonfiction, poetry, and drama using interpretive, critical, and evaluative processes.	α	α	α	α
3.02 Analyze characters, events, and plots from different selections and cite supporting evidence.		☆	α	α
3.03 Consider the ways language and visuals bring characters to life, enhance plot development, and produce a response.		α	α	☆
3.04 Make informed judgments about television and film/video productions.		☆	☆	
3.05 Integrate information from two or more sources to expand understanding of text.		☆	☆	α
3.06 Conduct research for assigned or self-selected projects (with assistance) from a variety of sources through the use of technological and informal tools.	☆	☆	☆	☆

Table 3: English Language Arts goals and objectives linked to the four categories for Grade 5

English Language Arts, Grade 5 Objectives from Goals 1, 2, and 3	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will apply enabling strategies and skills to read and write.</i>				
1.01 Expand and refine vocabulary through knowledge of prefixes, suffixes, roots, derivatives, and etymologies to assist comprehension.	α			
1.02 Select key vocabulary critical to the text and apply appropriate meanings as necessary for comprehension.	α			
1.03 Increase reading and writing vocabulary.	★			
1.04 Use word reference materials to identify and comprehend unknown words.	α			
1.05 Read independently daily from self-selected materials.	★	★	★	★
<i>Competency Goal 2: The learner will apply strategies and skills to comprehend text that is read, heard, and viewed.</i>				
2.01 Use metacognitive strategies independently and flexibly to monitor comprehension and extend vocabulary.	α	★	α	★
2.02 Interact with the text before, during, and after reading, listening, and viewing by making predictions, formulating questions, supporting answers from textual information, previous experience, and/or other sources, drawing on personal, literary, and cultural understandings, seeking additional information.	α	α	α	α
2.03 Read a variety of texts including fiction, nonfiction, poetry, and drama.	★	★	★	★
2.04 Identify elements of fiction and nonfiction and support by referencing the text to determine the plot development, author’s choice of words, and effectiveness of figurative language.	α	α	α	
2.05 Evaluate inferences, conclusions, and generalizations and provide evidence by referencing the text(s).		α	α	
2.06 Analyze choice of reading materials congruent with purposes.	α		α	
2.07 Evaluate the usefulness and quality of information and ideas based on purpose, experiences, text(s) and graphics.	α		α	
2.08 Explain and evaluate relationships that are causal, hierarchical, temporal, problem-solution.	★	α	α	★

2.09 Listen actively and critically.	☆	☆	☆	☆
2.10 Identify strategies used by a speaker or a writer to inform, entertain, or influence an audience.			α	
<i>Competency Goal 3: The learner will make connections through the use of oral language, written language, and media and technology.</i>				
3.01 Respond to fiction, nonfiction, poetry, and drama using interpretive, critical, and evaluative processes.	α	α	α	α
3.02 Make connections between texts by recognizing similarities and differences based on a common lesson, theme, or message.		☆	α	α
3.03 Justify evaluation of characters and events from different selections by citing supporting evidence in the text(s).		☆	☆	α
3.04 Make informed judgments about television, radio, video/film productions, and other electronic mediums and/or formats.	☆	☆	☆	☆
3.05 Integrate main idea and supporting details from multiple sources to expand understanding of texts.	☆	☆	α	α
3.06 Conduct research (with assistance) from a variety of sources for assigned or self-selected projects.	☆	☆	☆	☆
3.07 Make informed judgments about bias, propaganda, stereotyping, and media techniques.	☆	α	α	☆

Table 4: English Language Arts goals and objectives linked to the four categories for Grade 6.

English Language Arts, Grade 6 Goals 1, 2, 3, 4, 5, 6	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will use language to express individual perspectives drawn from personal or related experience.</i>				
1.01 Narrate a fictional or autobiographical account.	★	★	★	★
1.02 Explore expressive materials that are read, heard, and viewed.	★	★	★	★
1.03 Interact appropriately in group settings.	★	★	★	★
1.04 Reflect on learning experiences.		★	★	★
<i>Competency Goal 2: The learner will explore and analyze information from a variety of sources.</i>				
2.01 Explore informational materials that are read, heard, and/or viewed.	α	α	α	α
2.02 Use multiple sources of print and non-print information in developing informational materials such as brochures, newsletters, and infomercials.	★	★	★	★
<i>Competency Goal 3: The learner will examine the foundations and the use of argument.</i>				
3.01 Respond to public documents such as editorials and school or community policies that establish a position.	α	α	α	★
3.02 Explore the problem solution process.	★	★	★	★
3.03 Study arguments that evaluate.		★	★	★
<i>Competency Goal 4: The learner will use critical thinking skills and create criteria to evaluate text and multimedia.</i>				
4.01 Determine the purpose of the author or creator.	α	α	α	★
4.02 Develop (with teacher assistance) and apply appropriate criteria to evaluate the quality of the communication.	★	α	α	α
4.03 Recognize and develop a stance of a critic.		★	α	★
<i>Competency Goal 5: The learner will respond to various literary genres using interpretive and evaluative processes.</i>				
5.01 Increase fluency, comprehension, and insight through a meaningful and comprehensive reading program.	α	α	α	α

5.02 Study the characteristics of literary genres (fiction, nonfiction, drama, and poetry).	☆	☆	α	☆
<i>Competency Goal 6: The learner will apply conventions of grammar and language usage.</i>				
6.01 Demonstrate an understanding of conventional written and spoken expression.	α	☆	α	
6.02 Identify and edit errors in spoken and written English.	☆			

Table 5: English Language Arts goals and objectives linked to the four categories for Grade 7

English Language Arts, Grade 7 Goals 1, 2, 3, 4, 5, 6	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will use language to express individual perspectives in response to personal, social, cultural, and historical issues.</i>				
1.01 Narrate an account such as a news story or historical episode.	★	★	★	★
1.02 Explore expressive materials that are read, heard, and viewed.	★	★	★	★
1.03 Interact in group settings.	★	★	★	★
1.04 Reflect on learning experiences.		★	★	★
<i>Competency Goal 2: The learner will synthesize and use information from a variety of sources.</i>				
2.01 Explore informational materials that are read, heard, and/or viewed.	α	α	α	α
2.02 Develop informational products and/or presentations that use and cite at least three print or non-print sources.	★	★	★	★
<i>Competency Goal 3: The learner will refine the understanding and use of argument.</i>				
3.01 Analyze a variety of public documents that establish a position or point of view.	α	★	α	★
3.02 Use the problem solution process.	★	★	★	★
3.03 Create arguments that evaluate.	★	★	★	★
<i>Competency Goal 4: The learner will refine critical thinking skills and create criteria to evaluate text and multimedia.</i>				
4.01 Analyze the purpose of the author or creator.	α	α	α	★
4.02 Develop (with assistance) and apply appropriate criteria to evaluate the quality of the communication.	★	α	α	α
4.03 Develop the stance of a critic.		★	α	★
<i>Competency Goal 5: The learner will respond to various literary genres using interpretive and evaluative processes.</i>				
5.01 Increase fluency, comprehension, and insight through a meaningful and comprehensive reading program.	α	α	α	α
5.02 Study the characteristics of literary genres (fiction, nonfiction, drama, and poetry).	★	★	α	★

<i>Competency Goal 6: The learner will apply conventions of grammar and language usage.</i>				
6.01 Model an understanding of conventional written and spoken expression.	α	☆	α	
6.02 Continue to identify and edit errors in spoken and written English.	☆			

Table 6: English Language Arts goals and objectives linked to the four categories for Grade 8

English Language Arts, Grade 8 Goals 1, 2, 3, 4, 5, 6	Cognition	Interpretation	Critical Stance	Connections
<i>Competency Goal 1: The learner will use language to express individual perspectives in response to personal, social, cultural, and historical issues.</i>				
1.01 Narrate an account.	☆	☆	☆	☆
1.02 Explore expressive materials that are read, heard, and viewed.	☆	☆	☆	☆
1.03 Interact in group and/or seminars.	☆	☆	☆	☆
1.04 Reflect on learning experiences.		☆	☆	☆
<i>Competency Goal 2: The learner will use and evaluate information from a variety of sources.</i>				
2.01 Explore informational materials that are read, heard, and/or viewed.	α	α	α	α
2.02 Create a research product in both written and presentational form.	☆	☆	☆	☆
<i>Competency Goal 3: The learner will continue to refine the understanding and use of argument.</i>				
3.01 Evaluate a variety of public documents.	α	☆	α	☆
3.02 Refine the use of the problem solution process.	☆	☆	☆	☆
3.03 Create arguments that persuade.		☆	☆	☆
<i>Competency Goal 4: The learner will continue to refine critical thinking skills and apply criteria to evaluate text and multimedia.</i>				
4.01 Analyze the purpose of the author or creator and the impact of that purpose.	α	α	α	☆
4.02 Develop (with limited assistance) and apply appropriate criteria to evaluate the quality of the communication.	☆	α	α	α
4.03 Use the stance of a critic.		☆	α	☆
<i>Competency Goal 5: The learner will respond to various literary genres using interpretive and evaluative processes.</i>				
5.01 Increase fluency, comprehension, and insight through a meaningful and comprehensive reading program.	α	α	α	α

5.02 Study the characteristics of literary genres (fiction, nonfiction, drama, and poetry).	☆	☆	α	☆
Competency Goal 6: The learner will apply conventions of grammar and language usage.				
6.01 Model an understanding of conventional written and spoken expression.	α	☆	α	
6.02 Continue to identify and edit errors in spoken and written English.	☆			

For the NC EOC Test of Reading Comprehension (English I), the tests are developed directly around the objectives found in the NCSCS rather than the four categories described above.

2.3 Item Writing

Using the NCSCS as the foundation, a test blueprint was developed to outline the average number of passages and items per passage for each strand. From these test blueprints, test specifications were generally designed to include the following:

- (1) Percentage of questions from higher or lower thinking skills and classification of each test question by level of difficulty;
- (2) Percentage of item types such as graphs, charts, diagrams, political cartoons, analogies, and other specialized constraints;
- (3) Percentage of test questions that measure a specific goal, objective, domain, or category;
- (4) For tests that contain reading selections, the percentage or number of types of reading selections (e.g. literary vs. nonliterary passages, percentage of composition vs. literary analysis, etc.)

Items on the North Carolina EOG and EOC Tests of Reading Comprehension were developed using categories of both “level of difficulty” and “thinking skill level.” Item writers used these frameworks when developing items. The purpose of the categories in the development of items was to ensure a balance of items across difficulty as well as a balance of items across the different cognitive levels of learning in the North Carolina EOG and EOC Tests of Mathematics.

Items were classified into three levels of difficulty: easy, medium, and hard. Easy items are those items that can be answered correctly by approximately 70 percent of the examinees. Medium items are those items that can be answered correctly by 50 to 60 percent of the examinees. Difficult items are those items that can be answered correctly by approximately 20 to 30 percent of the examinees. These targets were used for item pool development to ensure an adequate range of difficulty.

A more recent consideration for item development is the classification of items by “thinking skill level” or the cognitive skills that an examinee must use to solve a problem or answer a test question. Thinking skill levels are based on Marzano’s *Dimensions of*

Thinking (1988). In addition to its use in framing achievement tests, it is also a practical framework for curriculum development, instruction, assessment, and staff development. Thinking skills begin with the basic skill of “information-gathering” and move to more complex thinking skills such as integration and evaluation. A visual representation of the framework is provided in Figure 1.

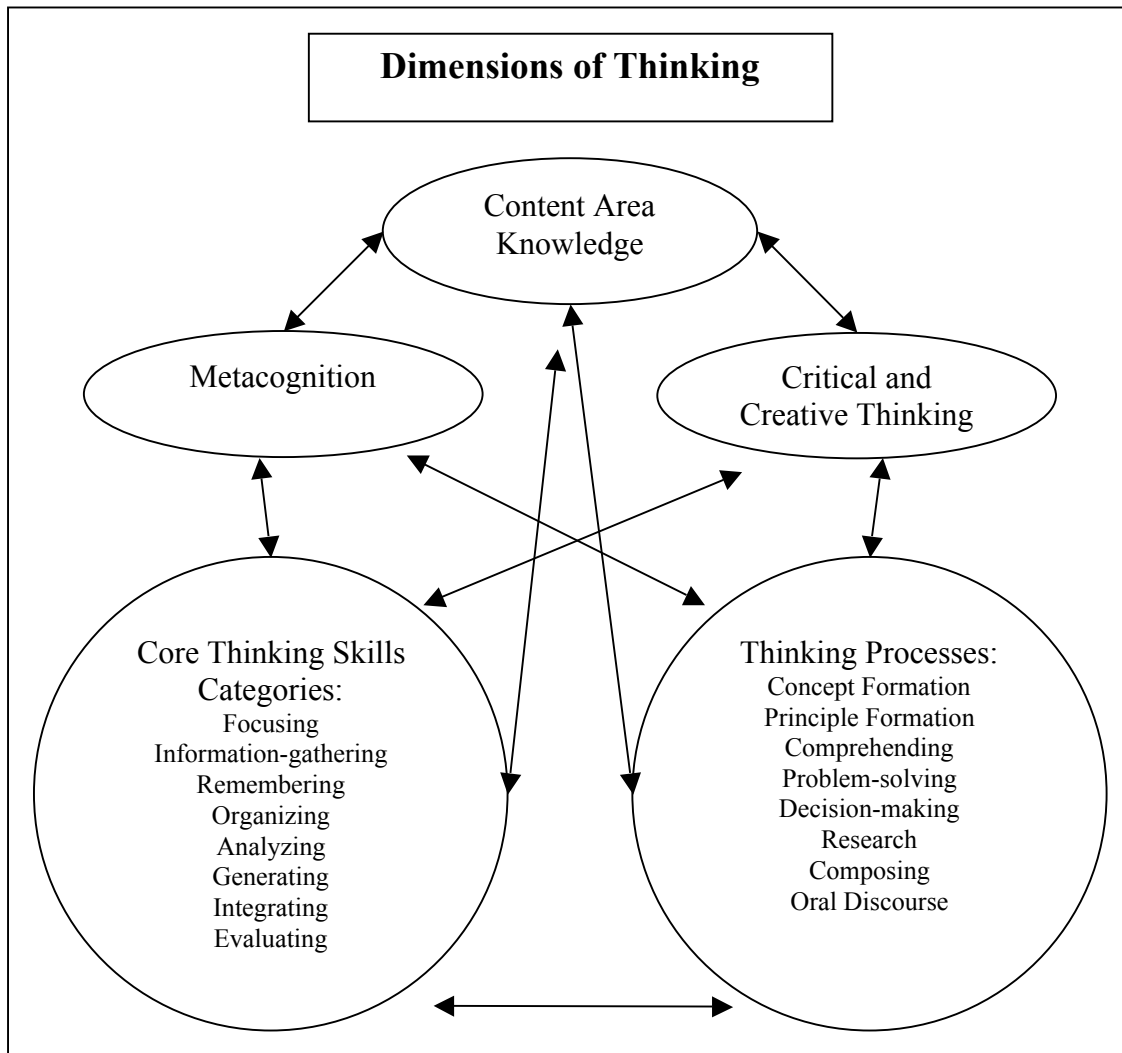


Figure 2: Thinking skills framework used to develop items in the North Carolina Testing Program adapted from Robert Marzano (1988)

2.3 Test Specifications

Delineating the purpose of a test must come before the test design. A clear statement of purpose provides the overall framework for test specifications, test blueprint, item development, tryout, and review. A clear statement of test purpose also contributes significantly to appropriate test use in practical contexts (Millman & Greene, 1993). The tests in the North Carolina Testing Program are designed in alignment with the NCSCS.

Test specifications for the North Carolina reading tests are developed to cover a wide range of styles and that provide students with authentic reading selections. Test specifications are generally designed to include the following:

- (1) Percentage of questions from higher or lower thinking skills and classification of each test question into level of difficulty;
- (2) Percentage of reading selections by type.

2.4 Selecting and Training Item Writers

Once the test specifications were outlined for the North Carolina EOG and EOC Tests of Reading Comprehension, North Carolina educators were recruited and trained to write new items for the state tests. Diversity among the item writers and their knowledge of the current NCSCS was addressed during recruitment. The purpose of using North Carolina educators to develop items was to ensure instructional validity of the items. For some items, item development was contracted to an external vendor. The vendor was encouraged to use North Carolina educators in addition to professional item writers to generate items that would align with the NCSCS for mathematics.

Training for item writers occurred over a three-day period. Item writers received a packet of materials designed from the mathematics curriculum, which included information on content and procedural guidelines as well as information on stem, foil, and distractor development. The item writing guidelines are included in Appendix A. The items developed during the training were evaluated by content specialists, who then provided feedback to the item writers on the quality of their items.

2.5 Reviewing Items for Field Testing

Each item was reviewed by North Carolina educators prior to being placed on a field test. Once items were reviewed by educators, test development staff members, with input from curriculum specialists, reviewed each item. Items were also reviewed by staff members who are familiar with the needs of students with disabilities and students with limited English proficiency.

The criteria used by the review team to evaluate each test item included the following:

- 1) Conceptual criteria:
 - objective match (curricular appropriateness)
 - thinking skill match
 - fair representation
 - lack of bias
 - clear statement
 - single problem
 - one best answer
 - common context in foils
 - each foil credible

- meets all technical criteria for item parameters
- 2) Language criteria:
- appropriate for age
 - correct punctuation
 - spelling and grammar
 - lack of excess words
 - no stem/foil clues
 - no negative in foils
- 3) Format criteria:
- logical order of foils
 - familiar presentation style, print size, and type
 - correct mechanics and appearance
 - equal length foils
- 4) Diagram criteria:
- necessary
 - clean
 - relevant
 - unbiased

The detailed review of items prior to field testing helped to prevent the loss of items due to quality issues.

2.6 Assembling Field Test Forms

When developing tests for the North Carolina Testing Program, items written for each subject/course area were assembled into forms for field testing. The forms were organized according to specifications set forth for the operational tests. Additional teachers reviewed the assembled forms for correctness, potential bias, and curricular appropriateness. Similar to the operational test review, North Carolina educators reviewed the assembled field test forms for clarity, correctness, potential bias, and curricular appropriateness. The following table provides a breakdown of the number of forms, number of items per form, and number of total items per grade or subject.

Table 7: Breakdown of the number of forms, number of items per form, and number of pages of reading averaged across forms for the field test forms

Grade/Course	Number of Forms	Average Number of Items per Form
3 Pre	14	34
3	13	66
4	13	68
5	13	69
6	13	74

7	13	75
8	17	75
10	11	66
English I	13	76

2.7 Sampling Procedures

Reading selections and items for the test were field tested using a randomly selected sample of students at each grade. The resulting sample was checked to determine its representative nature relative to the target population of students. The following table provides a breakdown of the field test sample.

Table 8: Field-test sample characteristics

Grade/Course	Year	Number of students tested	% Male	% Female	% American Indian	% Black	% White	% Other	% LEP
3 Pre	2003	16,135	51.35	48.64	1.2	30.18	58.83	9.79	2.84
3	2003	29,313	54.12	45.88	2.04	29.26	59.01	9.68	2.5
4	2003	30,862	50.32	49.67	1.71	28.65	60.37	9.27	2.4
5	2003	42,521	49.97	50.02	1.78	29.64	59.75	8.82	1.8
6	2003	40,404	50.62	49.37	1.30	29.18	60.54	8.95	1.3
7	2003	31,286	49.43	50.56	1.64	29.81	59.72	8.82	0.2
8	2003	27,613	49.66	50.34	1.33	28.06	62.93	7.69	1.2
10	1997	13,090	48.8	51.1	1.4	26.2	65.2	7.3	0.9
English I	2003	13,310	49.88	50.11	1.9	29.1	60.3	8.7	1.4

2.8 Item Analysis and Selection

Field testing provides important data for determining whether an item will be retained for use on an operational North Carolina EOG or EOC Test of Reading Comprehension. The North Carolina Testing Program uses both classical measurement analysis and item response theory analysis to determine if an item has sound psychometric properties. These analyses provide information that assists North Carolina Testing Program staff and consultants in determining the extent to which an item can accurately measure a student's level of achievement.

Field-test data for the North Carolina Reading Tests were analyzed by the NCDPI psychometric staff. Item statistics and description information were then printed on labels and attached to the item record for each item. Item records contained: the statistical, descriptive, and historical information for an item; a copy of the item as it was field-tested; any comments by reviewers; and the psychometric notations.

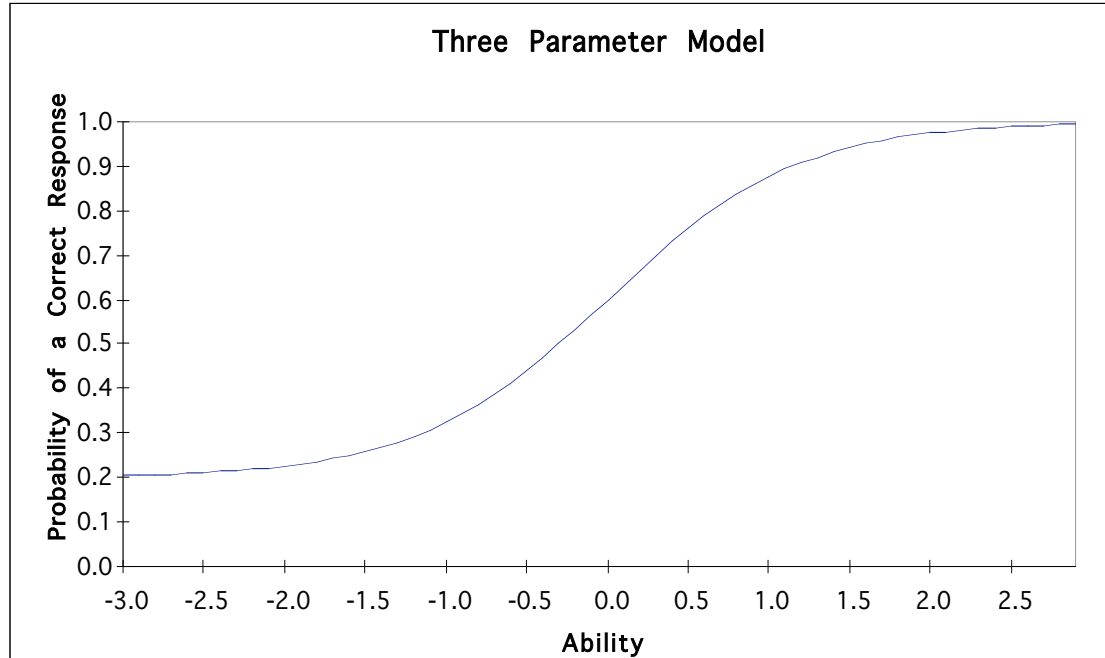
2.9 Classical Measurement Analysis

For each item the p-value (proportion of examinees answering an item correctly) and the point-biserial correlation between the item score and the total test score were computed using SAS. In addition, frequency distributions of the response choices were tabulated. While the p-value is an important statistic and is one component used in determining the selection of an item, the North Carolina Testing Program used item response theory (IRT) parameters to assess the psychometric appropriateness of the North Carolina EOG and EOC Tests of Reading Comprehension.

2.10 Item Response Theory Analysis

Many factors determine the appropriateness of IRT to a specific set of data including the content of the test, the nature of the population taking the test, and conditions under which the test is taken (for example, whether it is given by computer or paper and pencil). Item response theory is, with increasing frequency, being used with achievement level testing. “The reason for this may be the desire for item statistics to be independent of a particular group and for scores describing examinee proficiency to be independent of test difficulty, and for the need to assess reliability of tests without the tests being strictly parallel (Hambleton, 1993, in Linn, R. ed. *Educational Measurement*, p. 148).” Item response theory meets these needs and provides two additional advantages. The *invariance of item parameters* and the *invariance of ability parameters* make IRT analysis ideal for achievement testing. Regardless of the distribution of the sample, the parameter estimates will be linearly related to the parameters estimated with some other sample drawn from the same population. IRT allows the comparison of two students’ ability estimates even though they may have taken different items. An important characteristic of item response theory is the item-level orientation. IRT makes a statement about the relationship between the probability of answering an item correctly and the student’s ability or level of achievement. The relationship between an examinee’s item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases. The following figure shows the item characteristic curves for a typical 4-option multiple-choice item.

Figure 3: Typical item characteristic curve (ICC) for a typical 4-option multiple-choice item



2.11 Three-Parameter Logistic Model (3PL)

The three-parameter logistic model (3PL) of item response theory, the model used in generating EOG statistics, takes into account the difficulty of the item and the ability of the examinee. An examinee's probability of answering a given item correctly depends on the examinee's ability and the characteristics of the item. The 3PL model has three assumptions:

- (1) unidimensionality—only one ability is assessed by the set of items (for example, a spelling test only assesses a student's ability to spell);
- (2) local independence—when abilities influencing test performance are held constant, an examinee's responses to any pair of items are statistically independent (conditional independence, i.e., the only reason an examinee scores similarly on several items is because of his or her ability, not because the items are correlated); and
- (3) the item characteristic curve (ICC) specified below reflects the true relationship among the unobservable variable (ability) and the observable variable (item response).

The formula for the three-parameter logistic model is:

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

$P_i(\theta)$ -- is the probability that a randomly chosen examinee with ability θ answers item i correctly (this is an S-shaped curve with values between 0 and 1 over the ability scale)

a -- the slope or the discrimination power of the item (the slope of a typical item is 1.00)

b -- the threshold or the point on the ability scale where the probability of a correct response is 50% (the threshold of a typical item is 0.00)

c -- the asymptote or the proportion of the examinees who got the item correct, but did poorly on the overall test (the asymptote of a typical 4-choice item is 0.20)

D -- is a scaling factor, 1.7, to make the logistic function as close as possible to the normal ogive function (Hambleton, R.K. ,1983, *Applications of Item Response Theory*, p.125).

The IRT parameter estimates for each item were computed using the BILOG computer program (Muraki, Mislevy, & Bock, 1991) using the default Bayesian prior distributions for the item parameters [$a \sim \text{lognormal}(0, 0.5)$, $b \sim N(0,2)$, and $c \sim \text{Beta}(6,16)$].

2.12 Sensitivity Analysis

It is important to know the extent to which an item on a test performs differently for different students. Differential item functioning (DIF) examines the relationship between the score on an item and group membership while controlling for ability. The Mantel-Haenszel procedure examines DIF by examining ($j \times 2$) contingency tables, where j is the number of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the focus of interest and the reference group serves as a basis for comparison for the focal group (Dorans and Holland, 1993; Camilli and Shepherd, 1994). For example, females might serve as the focal group and males might serve as the reference group to determine if an item is biased towards or against females.

The Mantel-Haenszel 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 X^2 distribution has one degree of freedom (*df*) and is determined where r^2 is the Pearson correlation between the row variable and the column variable (*SAS Institute, 1994*).

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

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

2.13 Criteria for Inclusion in Item Pools

Items were flagged as exhibiting psychometric problems or bias due to ethnicity/race or gender according to the following criteria:

- “weak prediction”—the slope (a parameter) was less than 0.60,
- “guessing”—the asymptote (c parameter) was greater than 0.40,
- “ethnic” bias—the log odds ratio was greater than 1.5 (favored whites) or less than 0.67 (favored blacks), and
- “gender” bias—the log odds ratio was greater than 1.5 (favored females) or less than 0.67 (favored males).

The ethnic and gender bias flags were determined by examining the significance levels of items from several forms and identifying a typical point on the continuum of odds ratios that was statistically significant at the $\alpha = 0.05$ level. Because the tests were to be used to evaluate the implementation of the curriculum, items were not flagged on the basis of the difficulty of the item (threshold). Final average item pool parameter estimates for each of the North Carolina Reading Tests are provided below.

2.14 Item Parameter Estimates

Table 9: Average item pool parameter estimates for the NC Reading Comprehension Tests

Grade/ Course	IRT Parameters			P-value	Bias (Odds Ratio)	
	Threshold (<i>b</i>)	Slope (<i>a</i>)	Asymptote (<i>c</i>)		Ethnic/Race	Gender
3 Pre	0.171	1.054	0.176	0.586	1.108	1.022
3	-0.061	1.084	0.212	0.629	1.097	0.990
4	0.018	1.075	0.218	0.622	1.099	1.000
5	0.006	1.013	0.215	0.626	1.096	1.021
6	-0.015	1.099	0.219	0.624	1.095	0.990
7	0.259	1.053	0.216	0.569	1.069	0.994
8	0.305	1.023	0.221	0.564	1.075	1.006
10	1.255	1.116	0.216	0.412	1.009	1.108
English I	0.301	1.167	0.223	0.562	1.082	0.956

All items, statistics, and comments that were reviewed by curriculum specialists and testing consultants, and items found to be inappropriate for curricular or psychometric reasons were deleted. In addition, items flagged for exhibiting ethnic or gender bias (Mantel-Haenszel indices greater than 1.5 or less than 0.67) were then reviewed by a bias review team.

2.15 Bias Review Committee

The bias review team members, selected because of their knowledge of the curriculum area and their diversity, evaluated the items using the following questions:

- 1) Does the item contain any offensive gender, ethnic, religious, or regional content?
- 2) Does the item contain gender, ethnic, or cultural stereotyping?
- 3) Does the item contain activities that will be more familiar to one group than another?
- 4) Do the words in the item have a different meaning in one group than in another?
- 5) Could there be group differences in performance that are unrelated to proficiency in the content areas?

An answer of “yes” to any of the questions resulted in the five-digit item number being recorded on an item bias sheet along with the nature of the bias.

Items that were flagged by the bias review committee were then reviewed by curriculum specialists. If curriculum found the items measured content expected to be mastered by all students, the item was retained for test development. Items consistently identified as exhibiting bias by both review committees were deleted from the item pool.

2.16 Operational Test Construction

Once a sufficient number of items were developed for the item pools, operational tests were constructed. For NC Reading Tests three “first round” operational forms were assembled from items that were found to be psychometrically sound and measure curriculum standards as specified in the test specifications. The final item pool was based on approval by the (1) NCDPI Division of Instructional Services for curriculum purposes and (2) NCDPI Division of Accountability Services/NC Testing Program for psychometrically sound item performance. The forms for each grade and course were developed according to test specifications outlined during the initial phase of test development and the average p-value for each form was equivalent to the average p-value for the item pool.

2.17 Setting the Target p-value for Operational Tests

P-value is a measure of the difficulty of an item. P-values can range from 0 to 1. The letter “p” symbolizes the proportion of examinees that endorse an item correctly. So an item with a p-value of 0.75 was correctly endorsed by 75 percent of the students who took the item during the field test, and one might expect that roughly 75 of the 100 will examinees answer it correctly when the item is put on an operational test. An easy item has a p-value that is high—that means that a large proportion of the examinees got the item right during the field test. A difficult item has a low p-value, meaning that few examinees endorsed the item correctly during tryout.

The NCDPI psychometric staff must choose a target p-value for each operational test prior to assembling the tests. Ideally, the average p-value of a test would be 0.625, which is the theoretical average of a student getting 100 percent correct on the test and a student scoring a “chance” performance (25 percent for a 4-foil multiple-choice test). That is $(100 + 25/2)$. The target is chosen by first looking at the distribution of the p-values for a particular item pool. While the goal is to set the target as close to 0.625 as possible, it is often the case that the target p-value is set between the ideal 0.625 and the average p-value of the item pool. The average p-value of the item pool and the p-value of assembled forms are provided below for comparison.

Table 10: Comparison of p-values of item pools with p-values of assembled forms

Grade	p-Value of Item Pool	p-Value of Assembled Forms
3 Pre	0.599	0.586
3	0.629	0.642
4	0.622	0.638
5	0.626	0.640
6	0.624	0.632
7	0.569	0.572
8	0.564	0.574
10	0.412	0.412
English I	0.557	0.562

2.18 Setting the Test Administration Time

Other important considerations in the construction of the North Carolina Reading Comprehension tests were the number of items to be included on the test and the time necessary to complete testing. When assembling operational tests, the NCDPI psychometric staff reviewed field-test timing data. They determined the amount of time necessary for 98% of the students to complete the test. These data were then compared to the amount of time needed to complete previous operational administrations. In some cases it was necessary to reduce the number of items slightly so that test administration time was reasonable and comparable to previous years test administrations. For operational tests, the resulting total number of items for each grade/subject area is provided below.

Table 11: Number of items per test and time allotted by grade

Grade	Number of Items	Approximate Time Allotted (includes short breaks and general instructions)
3 Pre	32	83
3	50	130
4	50	130
5	50	130
6	56	127
7	56	127
8	56	127
10	56	127
English I	72	130

2.19 Reviewing Assembled Operational Tests

Once forms were assembled to meet test specifications, target p-values, and item parameter targets, ten to fifteen subject area teachers and curriculum supervisors then review the assembled forms. Each group of subject area teachers and curriculum supervisors worked independently of the test developers. The criteria for evaluating each group of forms included the following:

- ❑ The content of the test forms should reflect the goals and objectives of the North Carolina *Standard Course of Study* for the subject (curricular validity);
- ❑ The content of test forms should reflect the goals and objectives taught in North Carolina schools (instructional validity);
- ❑ Items should be clearly and concisely written, and the vocabulary appropriate to the target age level (item quality);
- ❑ Content of the test forms should be balanced in relation to ethnicity, gender, socioeconomic status, and geographic district of the state (test/item bias); and
- ❑ Each item should have one and only one best answer that is right; however, the distractors should appear plausible for someone who has not achieved mastery of the representative objective (one best answer).

Reviewers were instructed to take the tests (circling the correct responses in the booklet) and to provide comments and feedback next to each item. After reviewing all three forms in the set, each reviewer independently completed the survey asking for his or her opinion as to how well the tests met the five criteria listed above. During the last part of the session the group discussed the tests and made comments as a group. The ratings and the comments were aggregated for review by NCDPI curriculum specialists and testing consultants. Test development staff members, with input from curriculum staff and content experts, and editors conducted the final content and grammar check for each test form.

Chapter Three: Test Administration

3.1 Test Administration

The North Carolina Grade 3 Reading Comprehension Pretest, which measure grade 2 competencies in reading comprehension, is a multiple-choice test administered to all students in grade 3 within the first three weeks of the school year. The pretest allows schools to establish benchmarks to compare individual and group scale scores and achievement levels with the results from the regular end-of-grade test administered in the spring. In addition, a comparison of the results from the pretest and the results from the regular grade 3 end-of-grade test administration allows schools to measure growth in achievement in reading comprehension at the third-grade level for the ABCs accountability program. The grade 3 pretest measures the knowledge and skills specified for grade 2 from the reading comprehension goals and objectives of the 1998 North Carolina *Standard Course of Study*. The pretest is not designed to make student placement or diagnostic decisions in isolation.

The End-of-Grade Reading Comprehension Tests are administered to students in grades 3 through 8 as part of the statewide assessment program. The standard for grade-level proficiency is a test score at Achievement Level Three or above on both reading comprehension and mathematics tests. Effective with the 2002-2003 school year, the North Carolina End-of-Grade Reading Comprehension Tests are multiple-choice tests that measure the goals and objectives of the reading comprehension curriculum adopted in 1999 by the North Carolina State Board of Education for each grade.

The North Carolina High School Comprehensive Reading Test is administered to students in grade 10. It is a multiple-choice test that measures knowledge, skills, and competencies in reading that the typical student should have mastered by the end of the tenth grade.

All end-of-course tests are administered within the final ten days of the course to students enrolled for credit in courses where end-of-course tests are required. The purpose of end-of-course tests is to sample a student's knowledge of subject-related concepts specified in the North Carolina *Standard Course of Study* and to provide a global estimate of the student's mastery of the material in a particular content area. The reading end-of-course test (English I) was developed to provide accurate measurement of individual student knowledge and skills specified in the reading component of the North Carolina *Standard Course of Study*.

3.2 Training for Administrators

The North Carolina Testing Program uses a train-the-trainer model to prepare test administrators to administer North Carolina tests. Regional Accountability Coordinators (RACs) receive training in test administration from NCDPI Testing Policy and Operations staff at regularly scheduled monthly training sessions. Subsequently, the

RACs provide training on conducting a proper test administration to Local Education Agency (LEA) test coordinators. LEA test coordinators 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 students with limited English proficiency), test security (storing, inventorying, and returning test materials), and the Testing Code of Ethics.

3.3 Preparation for Test Administration

School test coordinators must be accessible to test administrators and proctors during the administration of secure state tests. The school test coordinator is responsible for monitoring test administrations within the building and responding to situations that may arise during test administrations. Only employees of the school system are permitted to administer secure state tests. Test administrators are school personnel who have professional training in education and the state testing program. Test administrators may not modify, change, alter, or tamper with student responses on the answer sheets or test books. Test administrators are to: thoroughly read the *Test Administrator's Manual* prior to actual test administration; discuss with students the purpose of the test; and read and study the codified North Carolina *Testing Code of Ethics*.

3.4 Test Security and Handling Materials

Compromised secure tests result in compromised test scores. To prevent contamination of test scores, the NCDPI maintains test security before, during, and after test administration at both the school system level and the individual school. School systems are also mandated to provide a secure area for storing tests. The Administrative Procedures Act 16 NCAC 6D .0302. states, in part, that

school systems 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 discussing with students or others specific items contained in the tests. Secure test materials may only be stored at each individual school for a short period prior to and after the test administration. Every effort must be made to minimize school personnel access to secure state tests prior to and after each test administration.

At the individual school, the principal shall account for all test materials received. As established by APA 16 NCAC 6D .0306, the principal shall store test materials in a secure locked area except when in use. The principal shall establish a procedure to have test materials distributed immediately prior to each test administration. 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.

3.5 Student Participation

The Administrative Procedures Act 16 NCAC 6D. 0301 requires that all public school students in enrolled grades for which the SBE adopts a test, including every child with disabilities, shall participate in the testing program unless excluded from testing as provided by 16 NCC 6G.0305(g).

Grade 3 Pretest and End of Grade Reading Comprehension Tests (Grades 3-8)

All students in membership in grade 3, including students who have been retained at grade 3 are required to participate in the Grade 3 Reading Comprehension Pretest. All students in membership in grades 3-8 are required to participate in the End-of-Grade Reading Comprehension Tests.

High School Comprehensive Tests

All students classified as tenth graders in the school system student information management system (SIMS, NCWise, etc.) must participate in the High School Comprehensive Reading Test. This also includes those students following the Occupational Course of Study (OCS) and those who are repeating grade 10.

English I End-of-Course Test

All students, including students with disabilities, enrolled in a course for credit must be administered the end-of-course test in the final ten days of the course. End-of-course tests are not required for graduation; however, students enrolled for credit in a course that has an end-of-course test must be administered the end-of-course test. Students who are repeating the course for credit must also be administered the end-of-course test. The student's most recent test score will be used for the purpose of state accountability. In addition, starting with the 2001-2002 school year, LEAs shall use results from all multiple-choice end-of-course tests (English I, Algebra I, Biology, US History, Economic Legal, and Policy Systems, Algebra II, Chemistry, Geometry, Physics, and Physical Science) as at least twenty-five percent of the student's final grade for each respective course. LEAs shall adopt policies regarding the use of end-of-course test results in assigning final grades.

3.6 Alternate Assessments

The North Carolina Testing Program currently offers the North Carolina Alternate Assessment Academic Inventory (NCAAAI) and the North Carolina Alternate Assessment Portfolio (NCAAP) as two alternate assessments for Grade 3 Pretest, the End-of-Grade Reading Comprehension Tests (grades 3-8), the High School Comprehensive Reading Test, and English I End-of-Course Test.

The NCAAAI is an assessment process in which teachers utilize a checklist to evaluate student performance on curriculum benchmarks in the areas of reading, mathematics and/or writing. Student performance data are collected at the beginning of the school year (baseline), in the middle of the school year (interim) and at the end of the school year (summative). The NCAAAI measures competencies on the North Carolina *Standard Course of Study*. The Individualized Education Program (IEP) team determines if a student is eligible to participate in the NCAAI.

The NCAAP is a yearlong assessment process that involves a representative and deliberate collection of student work/information that allows the users to make judgments about what a student knows and is able to do, and the progress that has been made in relation to the goals specified in the student's current IEP. The IEP team determines if the disability of a student is significant cognitive disability. The determination of a significant cognitive disability is a criterion for student participation in the NCAAP.

3.7 Testing Accommodations

On a case-by-case basis where appropriate documentation exists, students with disabilities and students with limited English proficiency may receive testing accommodations. The need for accommodations must be documented in a current Individualized Education Program (IEP), Section 504 Plan or LEP Plan. The accommodations must be used routinely during the student's instructional program or similar classroom assessments. 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 www.ncpublicschools.org/accountability/testing. 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.

3.8 Students with Limited English Proficiency

Per HSP-C-005, students identified as limited English proficient shall be included in the statewide testing program. Students identified as limited English proficient who have been assessed on the state-identified language proficiency test as below Intermediate High in reading may participate for up to 2 years (24 months) in U.S. schools in the NCAAAI as an alternate assessment in the areas of reading and mathematics at grades 3 through 8 and 10 and in high school courses in which an end-of-course test is

administered. Students identified as limited English proficient who have been assessed on the state-identified language proficiency test as below Superior, per HSP-A-011, in writing may participate in the NCAAAI in writing for grades 4, 7, and 10 for up to 2 years (24 months) in U.S. schools. All students identified as limited English proficient must be assessed using the state-identified language proficiency test at initial enrollment and annually thereafter during the window of February 1 to April 30. A student who enrolls after January 1 does not have to be retested during the same school year. Limited English proficient students who are administered the NCAAAI shall not be assessed off-grade level. In March 2004, the State Board of Education adopted a temporary rule to make the following changes with respect to limited English proficient students during their first year in U.S. schools.*

**Note: First year of enrollment in U.S. schools refers to the first school year that a student has been enrolled in a U.S. school. It does not refer to a 12-month period. If a student has been enrolled in any U.S. school prior to this school year, the student, regardless of his/her enrollment period would be expected to be assessed in reading and mathematics.*

Schools shall:

- continue to administer state reading and mathematics tests for LEP students who score at or above Intermediate High on the reading section of the language proficiency test during their first year in U.S. schools. Results from these assessments will be included in the ABCs and AYP.
- not require LEP students (who score below Intermediate High on the reading section of the language proficiency test) in their first year in U.S. schools to be assessed on the reading End-of-Grade tests, High School Comprehensive Test in Reading or the NC Alternate Assessment Academic Inventory (NCAAAI) for reading.
- for purposes of determining the 95% tested rule in reading, use the language proficiency test from the spring administration for these students.
- not count mathematics results in determining AYP or ABCs performance composite scores for LEP students who score below Intermediate High on the reading section of the language proficiency test in their first year in U.S. schools.
- include students previously identified as LEP, who have exited LEP identification during the last two years, in the calculations for determining the status of the LEP subgroup for AYP only if that subgroup already met the minimum number of 40 students required for a subgroup.

3.9 Medical Exclusions

In some rare cases students may be excused from the required states tests. The process for requesting special exceptions based on significant medical emergencies and/or conditions is as follows:

For requests that involve significant medical emergencies and/or conditions, the LEA superintendent or charter school director is required to submit a justification statement that explains why the emergency and/or condition prevents participation in the respective test administration during the testing window and the subsequent makeup period. The request must include the name of the student, the name of the school, the LEA code, and the name of the test(s) for which the exception is being requested. Medical documents are not included in the request to NCDPI. The request is to be based on information housed at the central office. The student's records must remain confidential. Requests must be submitted prior to the end of the makeup period for the respective test(s). Requests are to be submitted for consideration by the LEA superintendent or charter.

3.10 Reporting Student Scores

According to APA 16 NCAC 6D .0302 schools 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 tests that students will be required to take during the school year. In addition, school systems shall provide information to students and parents or guardians to advise them of the dates the tests will be administered and how the results from the tests will be used. Also, information provided to parents about the tests shall include whether the State Board of Education or local board of education requires the test. School systems shall report scores resulting from the administration of the district-wide and state-mandated tests to students and parents or guardians along with available score interpretation information within 30 days from the generation of the score at the school system level or receipt of the score and interpretive documentation from the NCDPI.

At the time the scores are reported for tests required for graduation such as competency tests and the computer skills tests, the school system shall provide information to students and parents or guardians to advise whether or not the student has met the standard for the test. If a student fails to meet the standard for the test, the students and parents or guardians shall be informed of the following at the time of reporting: (1) the date(s) when focused remedial instruction will be available and (2) the date of the next testing opportunity.

3.11 Confidentiality of Student Test Scores

State Board of Education policy states that “any written material containing the identifiable scores of individual students on tests taken pursuant to these rules shall not be

disseminated or otherwise made available to the public by any 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 Four: Scaling and Standard-Setting for the North Carolina Reading Comprehension Tests

The NC EOG and EOC Tests of Reading Comprehension scores are reported as scale scores, achievement levels, and percentiles. There are several advantages to using scale scores:

- ❑ Scale scores can be used to compare test results, when there have been changes in the curriculum and/or changes in the method of testing.
- ❑ Scale scores on pretests or released test forms can be related to scale scores used on secure test forms administered at the end of the course.
- ❑ 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.

4.1 Conversion of Test Scores

Each student's score is determined by calculating the number of items he or she answered correctly and then converting the sum to a developmental scale score. The program EOG_SCAL.LSP (developed by the L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill) is used to convert summed scores (total number of items answered correctly) to scale scores using the three item response theory parameters (threshold, slope, and asymptote) for each item. Because different items are used on each form of the test, unique score conversion tables are produced for each form of the test for each grade or subject area. For example, at grade 3 there are three End-of-Grade Reading Test forms. Therefore, three scale score conversion tables are used in the scanning and reporting program. In addition to producing scaled scores, the program also computes the standard error of measurement associated with each score.

4.2 Constructing a Developmental Scale

For Grades 3-8, a scale was constructed to measure developmental growth in skills and knowledge across grades. The original scale (1992) was constructed by administering two forms of the EOG for one grade at the next, higher grade. The next step was to analyze the linking forms and determine the differences in the distributions across grades. The individual items on each linking form were analyzed using the BIMAIN program to determine the marginal maximum likelihood estimation of the item parameters. Because all of the items were multiple-choice items, the three-parameter logistic model was used at both grades in which the linking form was administered. Item characteristic curves were developed for each item based on the IRT parameters and then the individual curves were aggregated across the test forms to develop the test characteristic curves. The test characteristic curves of the linking forms were compared from one grade to the next. Again, BIMAIN was used to determine the marginal maximum likelihood estimates of the proficiency distribution parameters. Next, the proficiency distributions were inferred based on the differences in item difficulties. The

population distributions of proficiency within grades were assumed to be Gaussian, where the grade's distribution was standard normal and the mean and standard deviation of the upper grade was estimated.

Following changes in curriculum specifications for reading, second edition tests were designed for the End-of-Grade Reading Comprehension Tests. As a result of these changes, new developmental scales were constructed for the second edition tests to provide a continuous measure of academic progress among North Carolina students. The new developmental scale was then “linked” to the first edition scale.

The table below shows the population means and standard deviations derived from the Spring 2002 item calibration for the second edition End-of-Grade Reading Comprehension Tests.

Table 12: Population means and standard deviations from the Spring 2002 item calibration for the NC Reading Comprehension Tests (n=1400 per form per grade)

NC EOG Reading Test		
Grade	Mean Scale Score	Standard Deviation
3 Pre	236.66	11.03
3	245.21	10.15
4	250.00	10.00
5	253.92	9.61
6	255.57	10.41
7	256.74	10.96
8	259.35	11.13

A comparison of the first-edition and second-edition population means and standard deviations is provided below. Note that the second edition begins with a “2” to distinguish it from the first edition scale.

Table 13: Comparison of population means and standard deviations for first and second editions of the End-of-Grade Reading Comprehension Tests

Grade	First Edition		Second Edition	
	Mean	Standard Deviation	Mean	Standard Deviation
3 Pre	139.02	8.00	236.66	11.03
3	145.59	9.62	245.21	10.15
4	149.98	9.5	250.00	10.01
5	154.74	8.21	253.92	9.61
6	154.08	9.44	255.57	10.41
7	157.81	9.09	256.74	10.96
8	159.55	8.96	259.35	11.13

The descriptive statistics shown above for each grade level provide the basis for the calculation of linear equating functions between the score-scales for the first and second editions of the reading test. When those figures are plotted, the equating lines are similar. This allows a linear equating function to be used to link the first- and second-edition scales. (For more information refer to Appendix C.)

4.3 Setting the Standards for the North Carolina Reading Tests

For tests developed under the North Carolina Testing Program, standard setting or the process of determining “cut scores” for the different achievement levels is typically accomplished through the use of “contrasting groups.” Contrasting groups is an examinee-based method of standard setting, which involves categorizing students into the various achievement levels by expert judges who are knowledgeable of students’ achievement in various domains outside of the testing situation and then comparing these judgments to students’ actual scores. For the North Carolina Reading Comprehension Tests, North Carolina teachers were considered to be expert judges under the rationale that teachers were able to make informed judgments about students’ achievement because they had observed the breadth and depth of the students’ work during the school year.

For the North Carolina Reading Comprehension Tests standard setting, approximately 95 percent of the students in each grade who participated in field testing were categorized into one of four achievement levels, with the remainder categorized as “not a clear example of any of the achievement levels.” This provided a proportional measure of the students expected to score in each of the four achievement levels. This categorization process occurred during the field test year so that proportions could be applied to scores from the first year a test is administered operationally to arrive at “cut scores” for each achievement level. Cut scores are the scores at which one achievement level ends and the next achievement level begins. Table 9 provides hypothetical percentages of contrasting-groups classifications.

Table 14: Hypothetical percentages of contrasting-groups classifications

Level I	8.22%
Level II	24.96%
Level III	43.60%
Level IV	22.74%
No Clear Category	0.48%

In contrasting-groups standard setting, scores from each grade would be distributed from lowest to highest. This distribution would then be used to set cut scores. For example, if a grade had 100,000 scale scores and those scores were distributed from lowest to highest, one would count up 8,220 (8.22%) scores from the bottom and then locate the cut-off score between Level I and Level II. Counting up the next 24,960 scores would provide the cut-off between Levels II and III. Counting up the next 43,600 scores would provide the cut-off between Levels III and IV. It should be noted that to avoid an

inflation of children categorized as Level IV, the percentage categorized as “No Clear Category” were removed from the cut score calculations. This process occurred at each grade for the NC EOG Tests of Reading Comprehension.

Since the administration of the first edition (1992) and the norming year (1998), the proportions of students in Level I have continued to decrease and the proportions of students in Levels III and IV have continued to increase. For example, from 1999 to 2000, 2% fewer children were in Level I than the year before. From 2000 to 2001 there were 1.8% fewer children in Level I than from 1999 to 2000. To continue this trend, it was anticipated that a similar percentage of fewer children would be in Level I from 2001 to 2002. Rather than develop new standards for the second edition of the NC EOG Tests of Reading comprehension, which would disrupt the continuous measure of academic progress for students, the standards for the second edition were established by maintaining the historical trends mentioned above while making use of the equated scales.

4.4 Achievement Level Descriptors

The four achievement levels in the North Carolina Student Accountability System are operationally defined below.

Table 15: Administrative Procedures Act 16 NCAC 6D .0501 (Definitions related to Student Accountability Standards)

Achievement Levels for the North Carolina Testing Program	
Level I	Students performing at this level do not have sufficient mastery of knowledge and skills in this subject area to be successful at the next grade level.
Level II	Students performing at this level demonstrate inconsistent mastery of knowledge and skills that are fundamental in this subject area and that are minimally sufficient to be successful at the next grade level.
Level III	Students performing at this level consistently demonstrate mastery of grade level subject matter and skills and are well prepared for the next grade level.
Level IV	Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.

4.5 Achievement Level Cut Scores

The achievement level score ranges for the North Carolina Reading Comprehension Tests are provided below.

Table 16: NC Reading Comprehension Tests achievement levels and corresponding scale scores

Grade/Subject	Level I	Level II	Level III	Level IV
Grade 3 Pre	211-219	220-229	230-239	240-260
3	216-229	230-239	240-249	250-272
4	223-235	236-243	244-254	255-275
5	228-238	239-246	247-258	259-277
6	228-241	242-251	252-263	264-283
7	228-242	243-251	252-263	264-287
8	231-243	244-253	254-265	266-290
10	132-150	151-162	163-174	175-201
English I	29-42	43-51	52-60	61-80

4.6 Achievement Level Trends

The percentage of students in each of the achievement levels is provided below by grade.

Table 17: Achievement level trends for Grade 3 Pretest

Grade 3 Pretest	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	*	*	11.3	11.3	10.2	9.1	8.2	7.5	7.4
Level II	*	*	23.4	23.4	22.5	21.1	20.6	19.7	19.7
Level III	*	*	39.5	40.1	41.1	41.3	42.7	42.7	43.9
Level IV	*	*	25.8	25.3	26.2	28.5	28.5	30.1	29.0

*The grade 3 pretest was not administered.

Table 18: Achievement level trends for Grade 3

Grade 3	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	12.9	11.3	11.0	8.6	6.9	6.2	5.7	4.2	3.9
Level II	23.7	23.9	23.2	19.8	19.5	19.4	17.9	16.0	13.5
Level III	37.2	37.9	37.6	36.3	36.7	38.0	38.4	38.8	37.1
Level IV	26.2	26.9	28.3	35.3	36.9	36.4	38.0	41.0	45.5

Table 19: Achievement level trends for Grade 4

Grade 4	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	10.8	9.0	9.9	7.9	7.4	7.0	6.1	4.7	4.2
Level II	25.1	21.6	22.4	21.2	21.2	21.0	19.4	18.2	12.0
Level III	41.6	44.8	42.9	41.5	43.8	42.3	43.2	44.7	41.9
Level IV	22.6	24.6	24.8	29.4	27.6	29.7	31.3	32.4	41.8

Table 20: Achievement level trends for Grade 5

Grade 5	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	8.0	8.9	7.6	6.1	5.0	4.4	3.4	2.7	1.8
Level II	23.8	24.6	21.6	18.8	19.3	16.6	13.9	12.8	9.5
Level III	41.3	41.3	41.4	40.4	43.1	41.0	43.2	44.5	45.0
Level IV	26.9	25.3	29.4	34.8	32.7	38.1	39.4	40.0	43.7

Table 21: Achievement level trends for Grade 6

Grade 6	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	8.2	7.0	6.6	5.0	4.3	4.1	3.3	2.2	1.7
Level II	24.1	20.5	20.7	16.7	14.6	14.9	13.8	11.4	8.2
Level III	42.5	43.0	40.5	40.7	39.8	38.1	40.5	39.2	34.5
Level IV	25.1	29.6	32.2	37.7	41.3	42.9	42.4	47.2	55.6

Table 22: Achievement level trends for Grade 7

Grade 7	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	8.4	9.0	8.6	5.4	3.9	4.5	3.2	2.7	2.9
Level II	24.5	22.5	20.6	17.7	13.6	14.8	15.5	14.0	13.3
Level III	38.6	38.8	36.9	38.3	37.4	35.1	33.3	32.4	31.1
Level IV	28.5	29.7	34.0	38.6	45.0	45.6	48.0	50.9	52.7

Table 23: Achievement level trends for Grade 8

Grade 8	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	8.2	8.8	9.0	5.4	5.4	4.8	5.3	4.2	4.5
Level II	24.2	23.5	22.1	18.3	17.0	14.6	15.2	13.5	11.3
Level III	40.1	38.7	38.4	37.6	37.9	36.5	36.8	35.7	34.1
Level IV	27.5	29.1	30.5	38.7	39.7	44.1	42.7	46.6	50.1

Table 24: Achievement level trends for Grade 10 High School Comprehensive Test

Grade 10	1995	1996	1997	1998	1999	2000	2001	2002	2003
Level I	*	*	*	11.9	8.8	8.8	9.5	*	8.3
Level II	*	*	*	32.5	30.2	29.4	28.9	*	27.0
Level III	*	*	*	41.0	45.2	45.4	44.9	*	47.5
Level IV	*	*	*	14.6	15.9	16.4	16.7	*	17.2

4.7 Percentile Ranking

The percentile rank for each scale score is the percentage of scores less than or equal to that score. If the percentile formula is applied to the frequency distribution of scores for grade three reading, then a score of 260 would have a percentile rank of 89th. Eighty-nine percent of students scored at or below a score of 260. The percentile rank provides information about a student's score on a test relative to other students in the norming year. The percentile ranks for the scores on the North Carolina EOG Tests of Reading Comprehension were calculated based on the 2002 administration of the tests.

Score reports are generated at the local level to depict achievement for individual students, classrooms, schools, and local education agencies. These data can be disaggregated by subgroups of gender and race/ethnicity as well as other demographic variables collected during the test administration. Demographic data are reported on variables such as free/reduced lunch status, LEP status, migrant status, Title I status, disability status, and parents' levels of education. The results are reported in aggregate at the state level usually at the end of June of each year. The NCDPI uses these data for school accountability and to satisfy other federal requirements such as Adequate Yearly Progress (AYP) and federal *No Child Left Behind* mandates.

Chapter Five: Reports

5.1 Use of Test Scores Reports Provided by the North Carolina Testing Program

The North Carolina Testing Program provides reports at the student level, school level, and state level. The North Carolina 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 together with other scores and indicators. Test data help educators understand educational patterns and practices. Data analysis of test scores for decision-making purposes should be based upon disaggregation of data by student demographics and other student variables as well as an examination of grading practices in relation to test scores, growth trends, and goal summaries for state-mandated tests.

5.2 Reporting by Student

The state provides scoring equipment in each school system so that administrators can score all state-required multiple-choice tests. This scoring generally takes place within two weeks after testing so the individual score report can be given to the student and parent before the end of the school year.

Each student in grades 3-8 who takes the end-of-grade tests is given a “Parent/Teacher Report”. This single sheet provides information on that student’s performance on the reading and mathematics tests. A flyer titled, “Understanding Your Child’s EOG Score,” is provided with each “Parent/Teacher Report.” This publication offers information for understanding student scores as well as suggestions on what parents and teachers can do to help students in the areas of reading and mathematics.

The student report also shows how that student’s performance compared to the average scores for the school, the school system and the state. A four level grade scale is used for the tests:

- Achievement Level I represents insufficient mastery of the subject.
- Achievement Level II is inconsistent mastery of the subject.
- Achievement Level III is consistent mastery and the minimum goal for students.
- Achievement Level IV is superior mastery of the subject.

Student achieving at Level III or Level IV are considered to be at or above grade level. Achievement Level III is the level students must score to be considered proficient and to pass to the next grade under state Student Accountability Standards for grades 3, 5, and 8.

5.3 Reporting by School

Since 1997, the student performance on end-of-grade tests for each elementary and middle school has been released by the state through the ABCs School Accountability.

High school student performance began to be reported in 1998 in the ABCs School Accountability. For each school, parents and others can see: the actual performance for groups of students at the school in reading, mathematics, and writing; the percentage of students tested; whether the school met or exceeded goals that were set for it; and the status designated by the state.

Some schools that do not meet their goals and that have low numbers of students at grade level receive help from the state. Other schools, where goals have been reached or exceeded, receive bonuses for the certified staff and teacher assistants in that school. Local school systems received their first results under No Child Left Behind (NCLB) in July 2003 as part of the state's ABCs accountability program. Under NCLB, each school is evaluated according to whether or not it met Adequate Yearly Progress (AYP). AYP is not only a goal for the school overall, but also for each subgroup of students in the school. Every subgroup must meet its goal for the school to meet AYP.

AYP is only one part of the state's ABCs accountability model. Complete ABCs results are released in September and will show how much growth students in every school made as well as the overall percentage of students who are proficient. The ABCs report is available on the Department of Public Instruction web site at <http://abcs.ncpublicschools.org/abcs/>. School principals also can provide information about the ABC report to parents.

5.4 Reporting by the State

The state reports information on student performance in various ways. The North Carolina Report Cards provide information about K-12 public schools (including charters and alternative schools) for schools, school systems and the state. Each report card includes a school or district profile and information about student performance, safe schools, access to technology and teacher quality.

As a participating state in the National Assessment of Educational Progress (NAEP), North Carolina student performance is included in annual reports released nationally on selected subjects. The state also releases state and local SAT scores each summer.

Chapter Six: Descriptive Statistics and Reliability

6.1 Means and Standards Deviations for the First Operational Administration

The second editions of the North Carolina Reading Comprehension Tests were administered for the first time in the spring of 2003. Descriptive statistics for the first operational year of the North Carolina EOG Tests of Reading Comprehension are provided below along with operational administration population demographics.

Table 25: Descriptive statistics by grade for the first operational administration of the North Carolina Reading Comprehension Tests

Year	Grade/Course	N students tested	Mean	Standard Deviation
2003	3 Pre	102,397	239	9.93
2003	3	102,349	248	9.07
2003	4	100,483	252	8.68
2003	5	103,597	257	8.03
2003	6	104,816	259	8.55
2003	7	104,959	261	9.07
2003	8	102,192	264	9.06
1998	10	79,266	163	10.15
2003	English I	102,794	58	7.64

6.2 Population Demographics for the First Operational Administration

Table 26: Population demographics for the first operational administration of the North Carolina Reading Comprehension Tests

Grade/Course	N students tested	Male	Female	American Indian	Black	White	Other	% LEP
3 Pre	102,397	51.5	48.5	1.5	28.9	54.1	15.51	< .1
3	102,349	51.5	48.5	1.5	30.2	56.8	11.5	.2
4	100,483	50.4	47.8	1.4	29.8	56.5	12.3	.2
5	103,597	51.1	48.9	1.5	30.6	58.0	9.9	.3
6	104,816	51.5	48.5	1.4	30.8	58.6	9.2	.3
7	104,959	48.4	51.6	1.4	30.5	59.5	8.6	.3
8	102,192	50.6	49.4	1.4	29.8	60.4	8.4	.4
10	79,266	49.63	50.36	1.4	28.6	65.49	4.48	< .1
English I	102,794	50.6	49.4	1.4	29.4	61.3	7.9	< .1

6.3 Scale Score Frequency Distributions

The following tables present the frequency distributions of the developmental scale scores from the first statewide administration of the North Carolina Reading Comprehension Tests. The frequency distributions are not smooth because of the conversion from raw scores to scales scores. Due to rounding in the conversion process, sometimes two raw scores in the middle of the distribution convert to the same scale score resulting in the appearance of a “spike” in that particular scale score.

Figure 4: Scale score frequency distribution for the 2003 Grade 3 Reading Comprehension Pretest

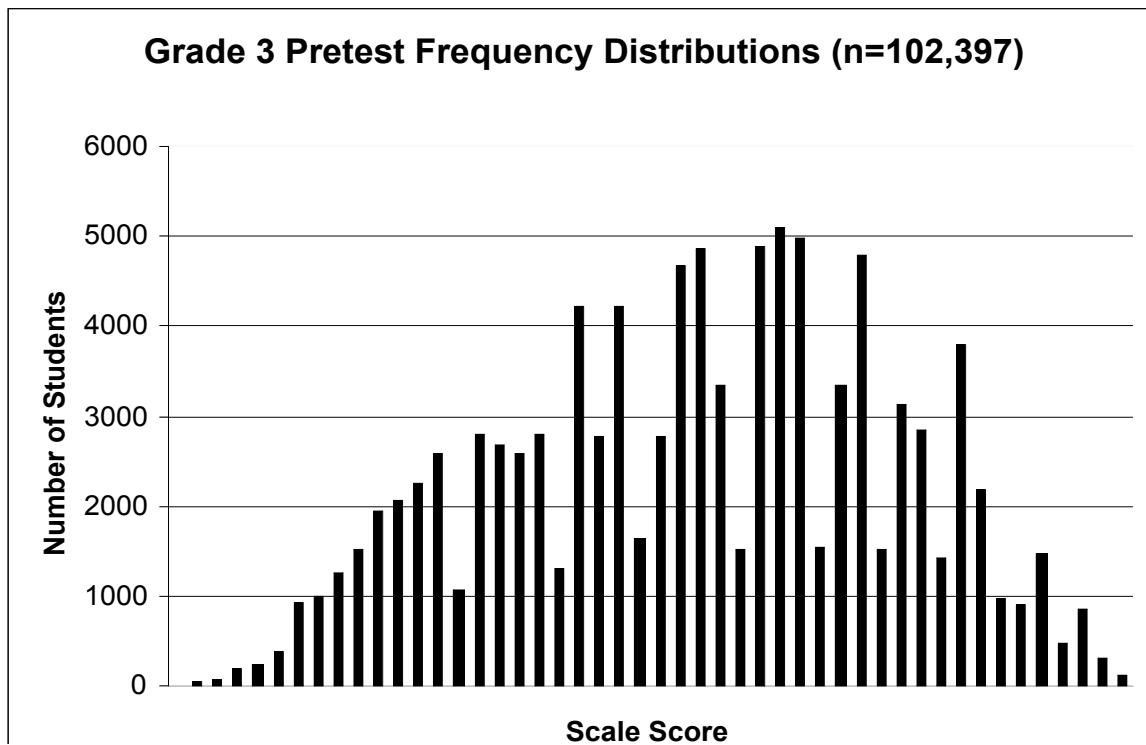


Figure 5: Scale score frequency distribution for the 2003 Grade 3 Reading Comprehension Test

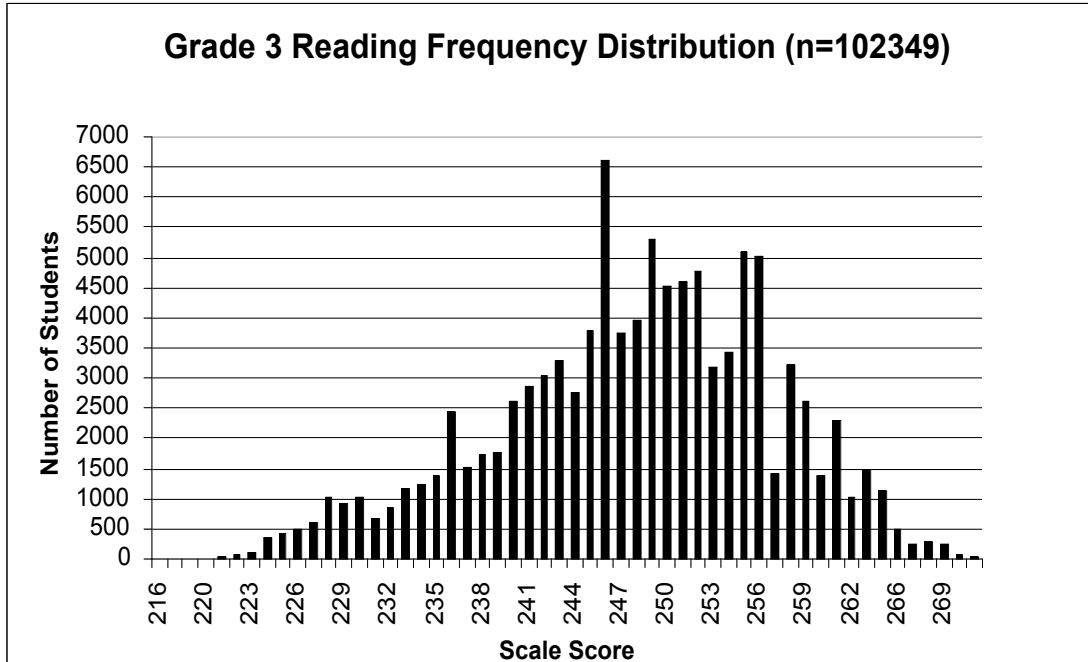


Figure 6: Scale score frequency distribution for the 2003 Grade 4 Reading Comprehension Test

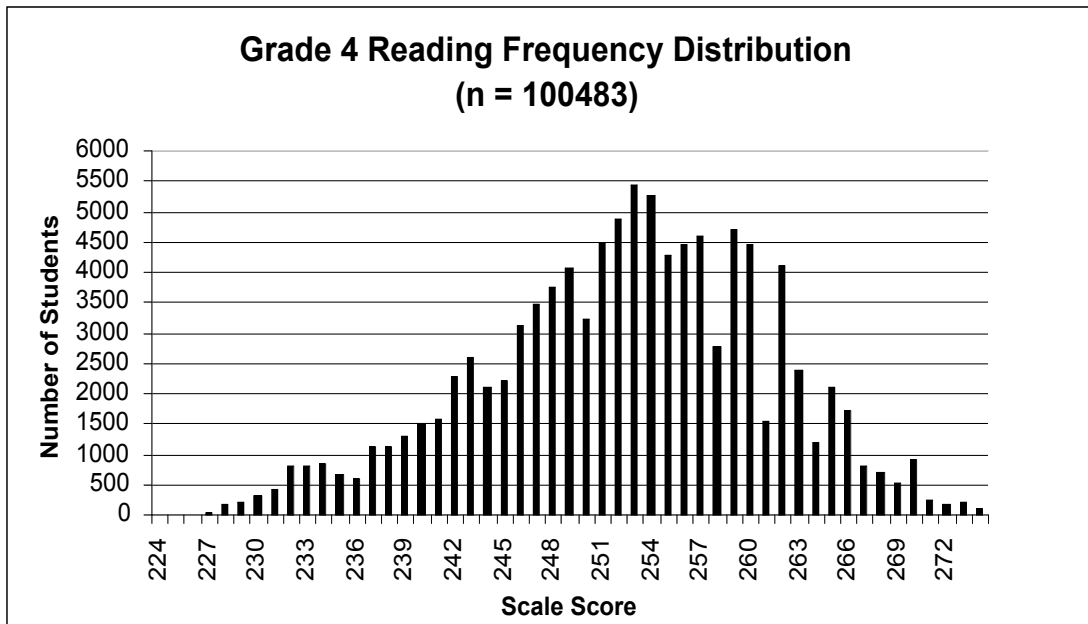


Figure 7: Scale score frequency distribution for the 2003 Grade 5 Reading Comprehension Test

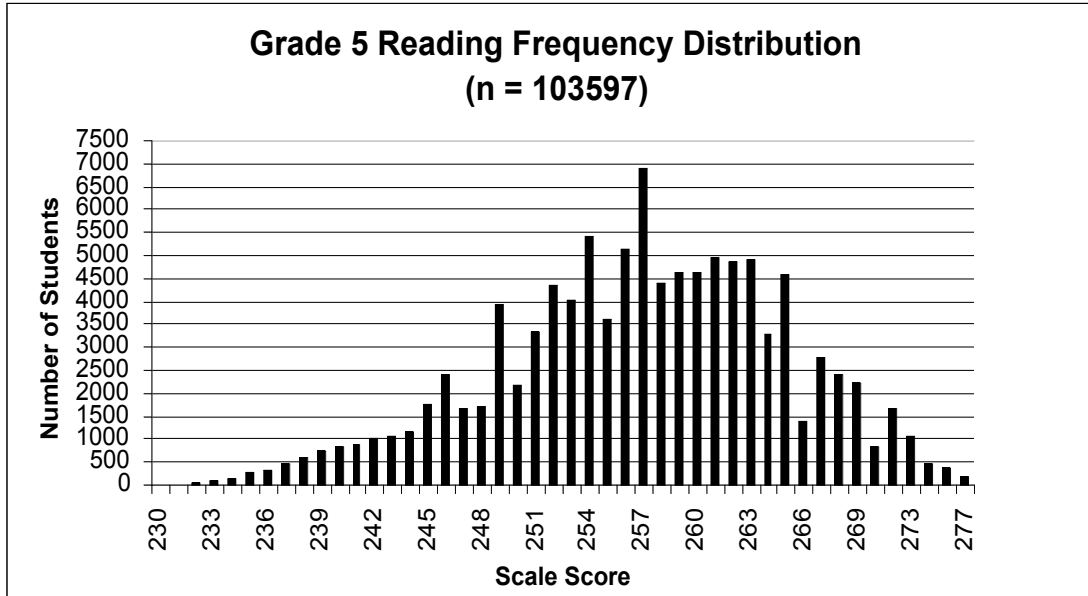


Figure 8: Scale score frequency distribution for the 2003 Grade 6 Reading Comprehension Test

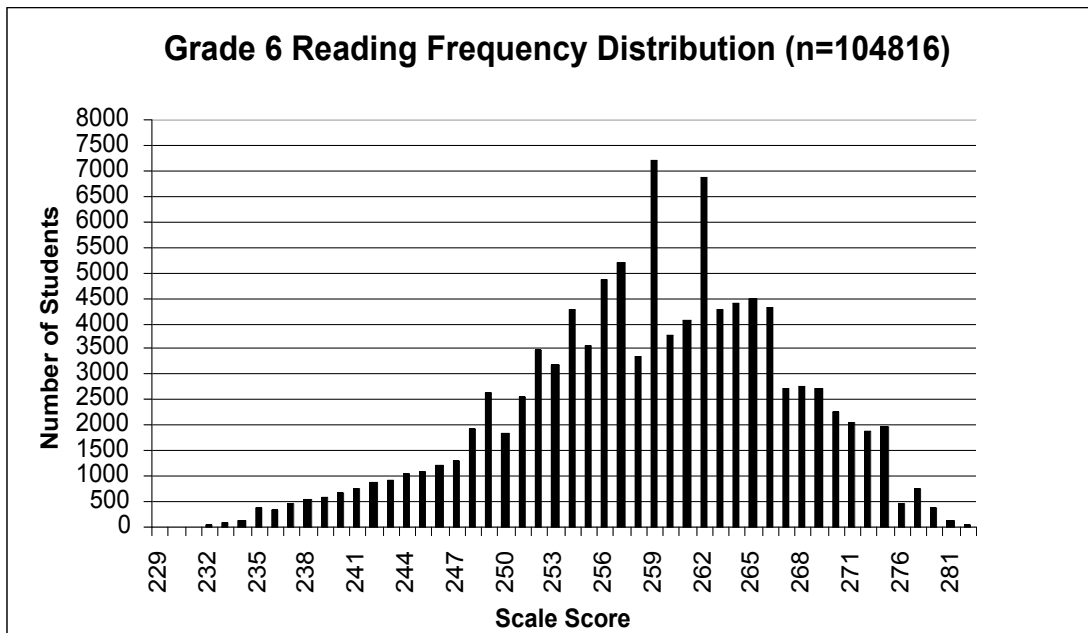


Figure 9: Scale score frequency distribution for the 2003 Grade 7 Reading Comprehension Test

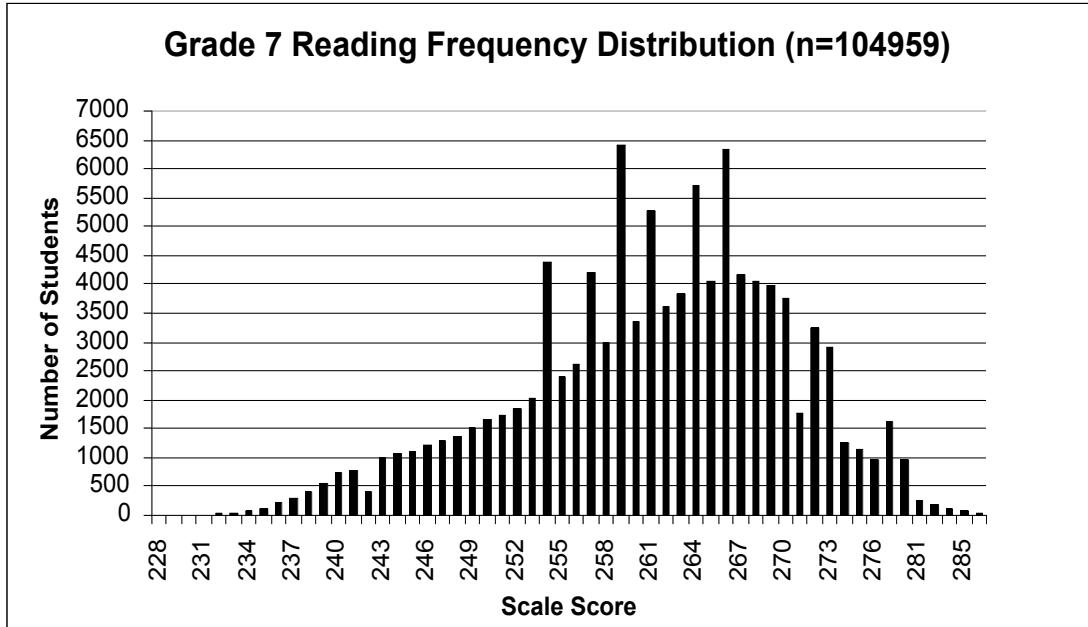


Figure 10: Scale score frequency distribution for the 2003 Grade 8 Reading Comprehension Test

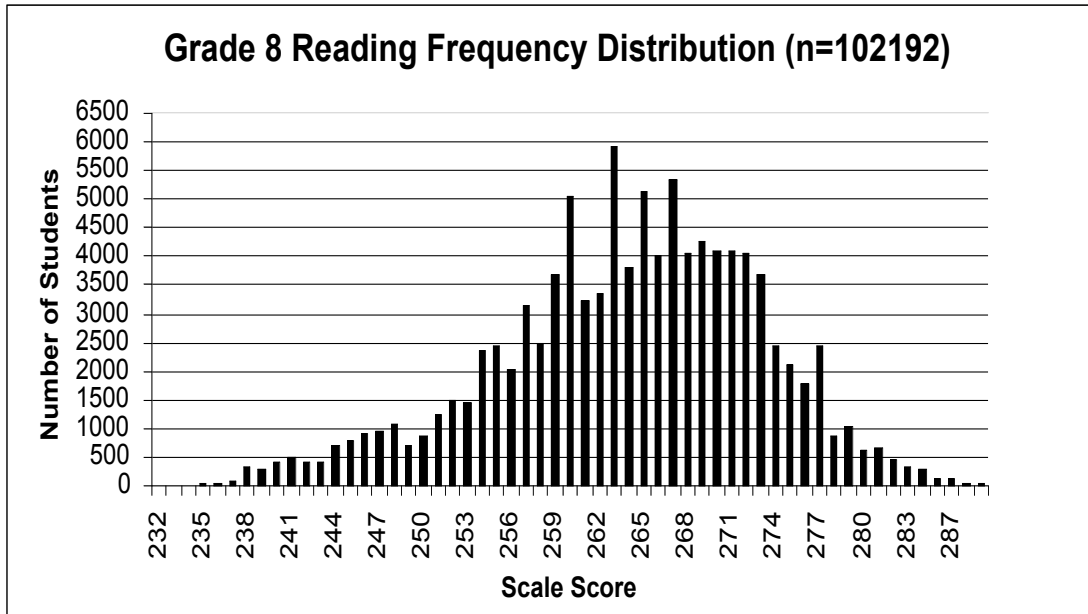


Figure 11: Scale score frequency distribution for the 1998 High School Comprehensive Reading Test

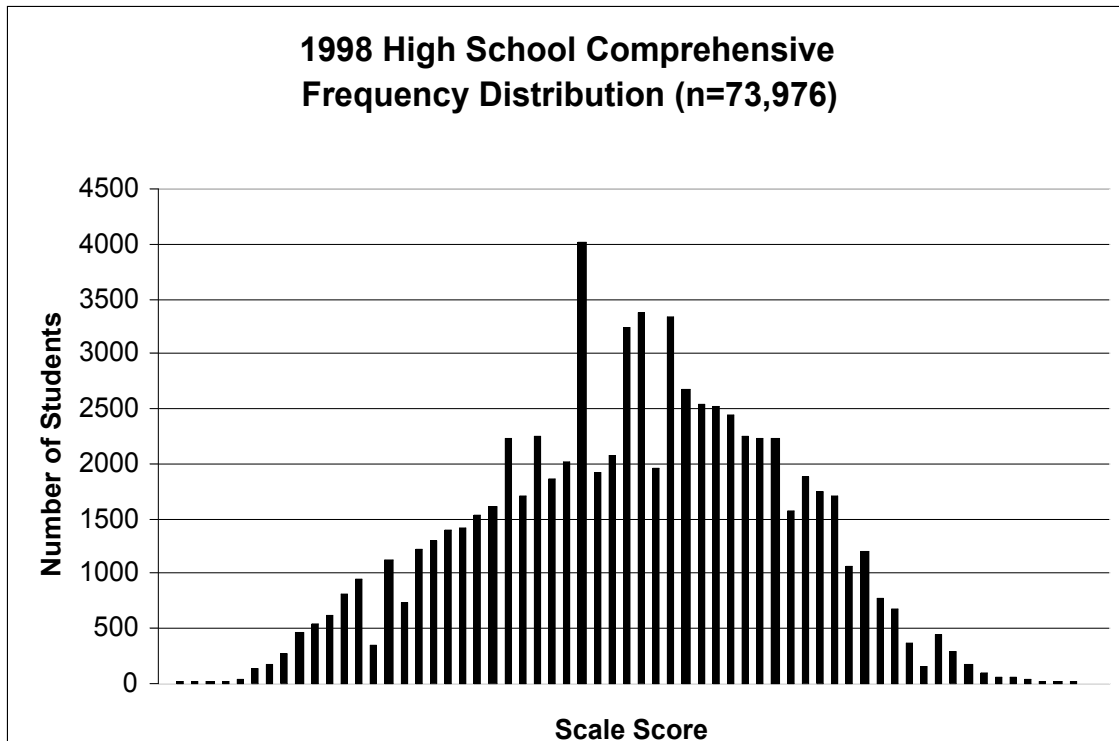
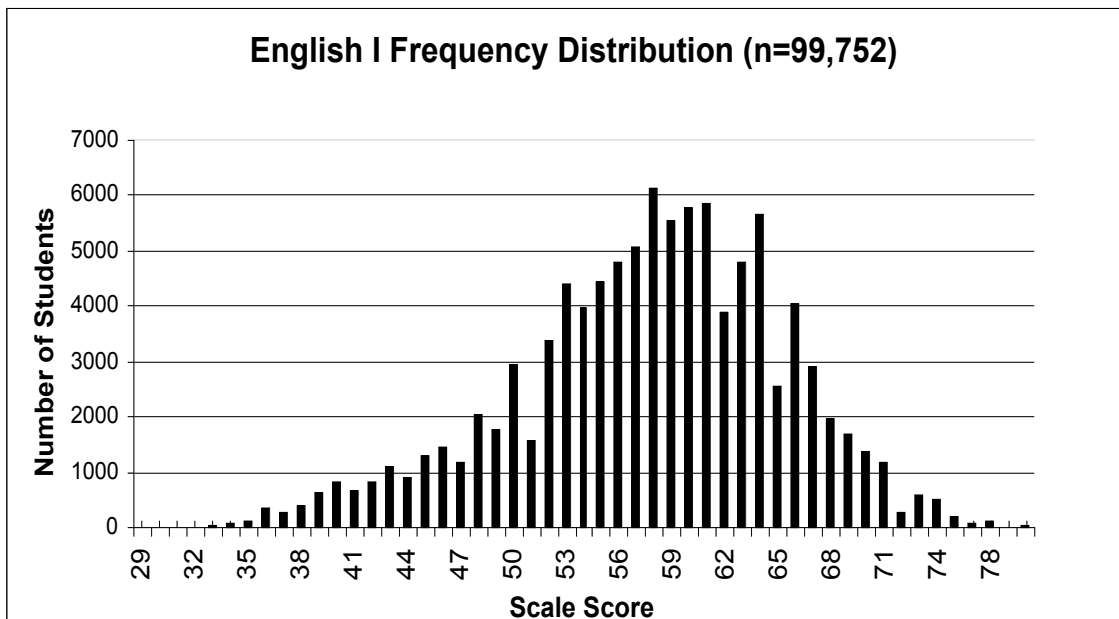


Figure 12: Scale score frequency distribution for the 2003 English I Reading Test



6.4 Reliability of the North Carolina Reading Tests

Reliability refers to the consistency of a measure when the testing procedure is repeated on a population of individuals or groups. Three broad categories of reliability coefficients are recognized as appropriate indices for establishing reliability in tests: (a) coefficients derived from the administration of parallel forms in independent testing sessions (alternate-form coefficients); (b) coefficients obtained by administration of the same instrument on separate occasions (test-retest or stability coefficients); and (c) coefficients based on the relationships among scores derived from individual items or subsets of the items within a test, all data accruing from a single administration of the test. The last coefficient, commonly known as the internal consistency coefficient (*Standards*, p.), is the coefficient used to establish reliability for the North Carolina Reading Comprehension Tests.

6.5 Internal Consistency of the North Carolina Reading Tests

Internal-consistency reliability estimates examine the extent to which the test measures a single basic concept. Internal-consistency reliability estimates examine the extent to which the test measures a single basic concept. One procedure for determining the internal consistency of a test is coefficient alpha (a). Coefficient alpha sets an upper limit to the reliability of tests constructed in terms of the domain sampling model. The formula for coefficient alpha is:

$$r_{xx} = \left(\frac{N}{N-1} \right) \left(\frac{S^2 - \sum s^2}{S^2} \right)$$

where

r_{xx} = coefficient alpha

N = Number of items constituting the instrument

S^2 = Variance of the summated scale scores

$\sum s^2_i$ = The sum of the variances of the individual items that constitute this scale (Hatcher & Stepanski, 1994, *Using SAS System for Univariate and Multivariate Statistics*).

If any use is to be made of the information from a test, then test results must be reliable. The North Carolina Testing Program follows industry standards and maintains a reliability coefficient of at least 0.85 on multiple-choice tests.

The following table presents the coefficient alpha indices averaged across forms by grade.

Table 27: Reliability indices for the NC Reading Comprehension Tests

Grade	Coefficient Alpha
3 Pre	.8153
3	.9245
4	.9243
5	.9180
6	.9367
7	.9214
8	.9167
10	.8832
English I	.8964

6.6 Standard Error of Measurement

The information provided by the standard error of measurement for a given score is important because it assists in determining the accuracy of an examinee's obtained score. It allows a probabilistic statement to be made about an individual's test score. For example, if a score of 100 has an SEM of plus or minus 2, then one can say that a student obtained a score of 100 which is accurate within plus or minus two points with 95 percent confidence, or the 95 percent confidence interval for a score of 100 is 98-102.

The standard error of measurement ranges for scores on the North Carolina Reading Comprehension Tests are provided below. For students with scores within two standard deviations of the mean (95% of the students), standard errors are typically 2 to 3 points. Students with scores that fall outside of two standard deviations (above the 97.5 percentile and below the 2.5 percentile) have standard errors of measurement of approximately 4 to 6 points. This is typical as scores become more extreme due to less measurement precision associated with those extreme scores.

Table 28: Ranges of standard error of measurement for scale scores by grade.

Grade/Subject	Standard Error of Measurement (Range)
3 Pre	3-6
3	2-5
4	2-6
5	2-6
6	2-6
7	2-6
8	2-6
10	3-6
English I	2-5

Chapter Seven: Evidence of Validity

7.1 Evidence of Validity

The validity of a test is the degree to which evidence and theory support the interpretation of test scores. Validity provides a check on how well a test fulfills its function. For all forms of test development, the validity of the test is an issue to be addressed from the first stage of development through analysis and reporting of scores. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed test score interpretations. Those interpretations of test scores are evaluated rather than the test itself. Validation, when possible, should include several types of evidence and the quality of the evidence is of primary importance (AERA, APA, NCME, 1985). For the North Carolina EOG and EOC Tests of Reading Comprehension, evidence of validity is provided through content relevance, response processes, relationship of test scores to other external variables, and maintaining consistency in the testing environment.

7.2 Content Validity

Evidence of content validity begins with an explicit statement of the constructs or concepts being measured by the proposed test. Interpretation of test scores refers to constructs or concepts the test is proposed to measure. The constructs or concepts measured by the NC EOG Tests of Reading Comprehension have four basic strands: cognition, interpretation, critical stance, and connections.

All items developed for the EOG are done so to measure those four concepts with particular focus on assessing students' ability to process information and engage in higher order thinking. The tables below provide the major goals or concepts measured by each of the NC EOG and EOC Tests of Reading Comprehension and the percentage of items by each of the four strands. The purpose of the test specification summary is to show the distribution of items across the curriculum.

Table 29: Grade 3 Reading Comprehension Test Specifications

Grade 3	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	18.7	37.3%
Interpretation	18.3	36.7%
Critical Stance	9.7	19.3%
Connections	3.3	6.7%
Totals	50	100%

Table 30: Grade 4 Reading Comprehension Test Specifications

Grade 4	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	19.7	39.3%
Interpretation	19.3	38.7%
Critical Stance	9.0	18.0%
Connections	2.0	4.0%
Totals	50	100%

Table 31: Grade 5 Reading Comprehension Test Specifications

Grade 5	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	17.7	35.3%
Interpretation	19.7	39.3%
Critical Stance	10.0	20.0%
Connections	2.7	5.3%
Totals	50	100%

Table 32: Grade 6 Reading Comprehension Test Specifications

Grade 6	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	16.3	29.2%
Interpretation	22.3	39.9%
Critical Stance	13.7	24.4%
Connections	3.7	6.5%
Totals	56	100%

Table 33: Grade 7 Reading Comprehension Test Specifications

Grade 7	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	14.5	25.9%
Interpretation	23.5	42.0%
Critical Stance	15.0	26.8%
Connections	3.0	5.4%
Totals	56	100%

Table 34: Grade 8 Reading Comprehension Test Specifications

Grade 8	Average Number of Items per Form	Average Percentage of Items Per Form
Cognition	16.3	29.2%
Interpretation	22.3	39.9%
Critical Stance	14.0	25.0%
Connections	3.3	6.0%
Totals	56	100%

7.3 Criterion-Related Validity

Analysis of the relationship of test scores to variables external to the test provide another important source of validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs.

Criterion-related validity of a test indicates the effectiveness of a test in predicting an individual's behavior in a specific situation. The criterion for evaluating the performance of a test can be measured at the same time (concurrent validity) or at some later time (predictive validity). For the North Carolina Reading Comprehension Tests, teachers' judgments of student achievement, expected grade, and assigned achievement levels all serve as sources of evidence of concurrent validity. The Pearson correlation coefficient is

used to provide a measure of association between the scale score and those variables listed above. The correlation coefficients for the North Carolina Reading Comprehension Tests range from 0.49 to 0.65, indicating a moderate to strong correlation between scale scores and their associated variables.

Table 35: Pearson Correlation Coefficients for the NC Reading Comprehension Tests (Grade 3 Pretest and Grades 3-8)

Variables	Grade 3 Pretest	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Assigned Achievement Level by Expected Grade	0.54	0.63	0.60	0.57	0.49	0.46	0.44
Teacher Judgment of Achievement by Assigned Achievement Level	0.43	0.63	0.59	0.60	0.57	0.56	0.57
Expected Grade by Scale Score	0.58	0.67	0.65	0.63	0.53	0.51	0.49
Teacher Judgment of Achievement by Scale Score	0.46	0.67	0.64	0.65	0.62	0.61	0.61

Table 36: Pearson Correlation Coefficients for English I Test

Variables	English I
Assigned Achievement Level by Expected Grade	0.56
Teacher Judgment of Achievement by Assigned Achievement Level	0.51
Expected Grade by Scale Score	0.59
Teacher Judgment of Achievement by Scale Score	0.53

Chapter Eight: Quality Control Procedures

Quality control procedures for the North Carolina testing program are implemented throughout all stages of testing. This includes quality control for test development, test administration, scores analysis, and reporting.

8.1 Quality Control Prior to Test Administration

Once test forms have been assembled, they are reviewed by a panel of subject experts. Once the review panel has approved a test form, test forms are then configured to go through the printing process. Printers send a blue-lined form back to NCDPI Test Development staff to review and adjust if necessary. Once all test answer sheets and booklets are printed, the test project manager must conduct a spot check of test booklets to ensure that all test pages are included and test items are in order.

8.2 Quality Control in Data Preparation and Test Administration

Student background information must be coded before testing begins. The school system may elect to either: (1) pre-code the answer sheets, (2) direct the test administrator to code the Student Background Information, or (3) direct the students to code the Student Background Information. For the North Carolina multiple-choice tests, the school system may elect to pre-code some or all of the Student Background Information on SIDE 1 of the printed multiple-choice answer sheet. The pre-coded responses come from the schools' SIMS/NCWISE database. Pre-coded answer sheets provide schools with the opportunity to correct or update information in the SIMS/NCWISE database. In such cases, the test administrator ensures that the pre-coded information is accurate. The test administrator must know what information will be pre-coded on the student answer sheets to prepare for the test administration. Directions for instructing students to check the accuracy of these responses are located in test administrator manuals. All corrections for pre-coded responses are provided to a person designated by the school system test coordinator to make such corrections. The students and the test administrator must not change, alter, or erase pre-coding on students' answer sheets. To ensure that all students participate in the required tests and to eliminate duplications, students, regardless of whether they take the multiple-choice test or an alternate assessment, are required to complete the student background information on the answer sheets.

When tests and answer sheets are received by the local schools, they are kept in a locked, secure location. Class roster are reviewed for accuracy by the test administrator to ensure that students receive their answer sheets. During test administration at the school level, proctors and test administrators circulate throughout the test facility (typically a classroom) to ensure that students are using the bubble sheets correctly. Once students have completed their tests, answer sheets are reviewed and where appropriate cleaned by local test coordinators (removal of stray marks, etc.).

8.3 Quality Control in Data Input

All answer sheets are then sent from individual schools to the Local Test Coordinator, where they are scanned in a secure facility. The use of a scanner provides the opportunity to program in a number of quality control mechanisms to ensure that errors overlooked in the manual check of data are identified and resolved. For example, if the answer sheet is unreadable by the scanner, the scanner stops the scan process until the error is resolved. In addition, if a student bubbles in two answers for the same question, the scan records the student's answer as a (*) indicating that the student has answered twice.

8.4 Quality Control of Test Scores

Once all tests are scanned they are then sent through a secure system to the Regional Accountability Coordinators who check to ensure that all schools in all LEA's have completed and returned student test scores. The Regional Accountability Coordinators also conduct a spot check of data and then send the data through a secure server to the North Carolina Department of Public Instruction Division of Accountability. Data are then imported into a file and cleaned. When a portion of the data are in, NCDPI runs a CHECK KEYS program to flag areas where answer keys may need second check. In addition, as data come into the NCDPI Division of Accountability, Reporting Section staff import and clean data to ensure that individual student files are complete.

8.5 Quality Control in Reporting

Scores can only be reported at the school level after NCDPI issues a certification statement. This is to ensure that school, district, and state-level quality control procedures have been employed. The certification statement is issued by the NCDPI Division of Accountability. The following certification statement is an example:

“The department hereby certifies the accuracy of the data from the North Carolina end-of-course tests for Fall 2004 provided that all NCDPI-directed test administration guidelines, rules, procedures, and policies have been followed at the district and schools in conducting proper test administrations and in the generation of the data. The LEAs may generate the required reports for the end-of-course tests as this completes the certification process for the EOC tests for the Fall 2004 semester.”

Definition of 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 typical four choice item is 0.20 but can vary somewhat by test. (For math it is generally 0.15 and for social studies it is generally 0.22).
Biserial correlation	The relationship between an item score (right or wrong) and a total test score.
Common Curriculum	Objectives that are unchanged between the old and new curricula
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 examine individual items for bias.
Operational Test	Test is administered statewide with uniform procedures and 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 Forms	Covers the same curricular material as other forms
Percentile	The score on a test below which a given percentage of scores fall.
Pilot Test	Test is administered as if it were “the real thing” but has limited associated reporting or stakes for examinees or schools.

Quasi-equated	Item statistics are available for items that have been through item tryouts (although they could change after revisions); and field test forms are developed using this information to maintain similar difficulty levels to the extent possible.
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 numbers of items from each objective to appear on test and 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.
WINSCAN Program	Proprietary computer program that contains the test answer keys and files necessary to scan and score state multiple-choice tests. Student scores and local reports can be generated immediately using the program.

References

- Gregory, Robert J. (2000) *Psychological Testing: History, Principles, and Applications*. Needham Heights: Allyn & Bacon.
- Hinkle, Wiersma, Jurs, *Applied Statistics for the Behavioral Sciences*, year, pp.69-70)
- Marzano, R.J., Brandt, R.S., Hughes, C.S., Jones, B.F., Presseisen, B.Z., Stuart, C./, & Suhor, C. (1988). *Dimensions of Thinking*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Millman, J., and Greene, J. (1993). *The Specification and Development of Tests of Achievement and Ability*. In Robert Linn (ed.), *Educational Measurement* (pp. 335-366). Phoenix: American Council on Education and Oryx Press.

Additional Resources

- Anastasi, A. (1982). *Psychological Testing*. New York: Macmillan Publishing Company, Inc.
- Berk, R.A. (1984). *A Guide to Criterion-Referenced Test Construction*. Baltimore: The Johns Hopkins University Press.
- Berk, R.A. (1982). *Handbook of Methods for Detecting Test Bias*. Baltimore: The Johns Hopkins University Press.
- Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage Publications, Inc.
- Dorans, N.J. & Holland, P.W. (1993). DIF Detection and description: Mantel-Haenszel and standardization. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp 35-66). Hillsdale, NJ: Lawrence Erlbaum.
- Haladyna, T.M. (1994). *Developing and Validating Multiple-Choice Test Items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.
- Holland, P.W. & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

- Joreskog, K.J. & Sorbom, D. (1986). PRELIS: A program for multivariate data screening and data summarization. Chicago, IL: Scientific Software, Inc.
- Joreskog, K.J. & Sorbom, D. (1988). LISREL 7: A guide to the program and applications. Chicago, IL: SPSS, Inc.
- Kubiszyn, T. & Borich, G. (1990). Educational Testing and Measurement. New York: HarperCollins Publishers.
- Muraki, E., Mislevy, R.J., & Bock, R.D. PC-Bimain Manual. (1991). Chicago, IL: Scientific Software, Inc.
- North Carolina Department of Public Instruction. (1993). North Carolina End-of-Grade Testing Program: Background Information. Raleigh, NC: Author.
- North Carolina Department of Public Instruction. (1996). North Carolina Testing Code of Ethics. Raleigh, NC: Author.
- Nunnally, J. (1978). Psychometric Theory. New York: McGraw-Hill Book Company.
- Rosenthal, R. & Rosnow, R.L. (1984). Essentials of behavioral research: Methods and data analysis. New York: McGraw-Hill Book Company.
- SAS Institute, Inc. (1985). The FREQ Procedure. In SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: Author.
- Traub, R.E. (1994). Reliability for the social sciences: Theory and applications. Thousand Oaks, CA: Sage Publications, Inc.

Appendix A: Item Development Guidelines

Content Guidelines

1. Items must be based on the goals and objectives outlined in the North Carolina *Standard Course of Study* in Reading Comprehension and written at the appropriate grade level.
2. To the extent possible, each item written should measure a single concept, principle, procedure, or competency.
3. Write items that measure important or significant material instead of trivial material.
4. Keep the testing vocabulary consistent with the expected grade level of students tested.
5. Avoid writing stems based on opinions.
6. Emphasize higher-level thinking skills using the taxonomy provided by the NCDPI.

Procedural Guidelines

7. Use the best answer format.
8. Avoid writing complex multiple-choice items.
9. Format the items vertically, not horizontally.
10. Avoid errors of grammar, abbreviations, punctuation, and spelling.
11. Minimize student reading time.
12. Avoid tricky or misleading items.
13. Avoid the use of contractions.
14. Avoid the use of first or second person.

Stem Construction Guidelines

15. Items are to be written in the question format.
16. Ensure that the directions written in the stems are clear and that the wording lets the students know exactly what is being tested.
17. Avoid excessive verbiage when writing the item stems.
18. Word the stems positively, avoiding any negative phrasing. The use of negatives such as NOT and EXCEPT is to be avoided.
19. Write the items so that the central idea and the phrasing are included in the stem instead of the foils.
20. Place the interrogative as close to the item foils as possible.

General Foil Development

21. Each item must contain four foils (A, B, C, D).
22. Order the answer choices in a logical order. Numbers should be listed in ascending or descending order.

23. Each item should contain foils that are independent and not overlapping.
24. All foils in an item should be homogeneous in content and length.
25. Do not use the following as foils: all of the above, none of the above, I don't know.
26. Word the foils positively, avoiding any negative phrasing. The use of negatives such as NOT and EXCEPT is to be avoided.
27. Avoid providing clues to the correct response. Avoid writing items with phrases in the stem (slang associations) that are repeated in the foils.
28. Also avoid including ridiculous options.
29. Avoid grammatical clues to the correct answer.
30. Avoid specific determiners since they are so extreme that they are seldom the correct response. To the extent possible, specific determiners such as ALWAYS, NEVER, TOTALLY, and ABSOLUTELY should not be used when writing items. Qualifiers such as *best*, *most likely*, *approximately*, etc. should be bold and italic.
31. The correct response for items written should be evenly balanced among the response options. For a 4-option multiple-choice item, each correct response should be located at each option position about 25 percent of the time.
32. Items should contain one and only one best (correct) answer.

Distractor Development

33. Use plausible distractors. The best (correct) answer must clearly be the best (correct) answer and the incorrect responses must clearly be inferior to the best (correct) answer. No distractor should be obviously wrong.
34. To the extent possible, use the common errors made by students as distractors. Give your reasoning for incorrect choices on the back of the item spec sheet.
35. Technically written phrases may be used, where appropriate, as plausible distractors.
36. True phrases that do not correctly respond to the stem may be used as plausible distractors where appropriate.
37. The use of humor should be avoided.

Appendix B: Scale Score Frequency Distribution Tables

Grade 3 Pretest Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
213	10	0.01	10	0.01
214	42	0.04	52	0.05
215	83	0.08	135	0.13
216	190	0.19	325	0.32
217	230	0.22	555	0.54
218	369	0.36	924	0.9
219	929	0.91	1853	1.81
220	1007	0.98	2860	2.79
221	1266	1.24	4126	4.03
222	1514	1.48	5640	5.51
223	1955	1.91	7595	7.42
224	2068	2.02	9663	9.44
225	2251	2.2	11914	11.64
226	2586	2.53	14500	14.16
227	1077	1.05	15577	15.21
228	2791	2.73	18368	17.94
229	2688	2.63	21056	20.56
230	2580	2.52	23636	23.08
231	2790	2.72	26426	25.81
232	1301	1.27	27727	27.08
233	4228	4.13	31955	31.21
234	2777	2.71	34732	33.92
235	4220	4.12	38952	38.04
236	1625	1.59	40577	39.63
237	2777	2.71	43354	42.34
238	4678	4.57	48032	46.91
239	4862	4.75	52894	51.66
240	3340	3.26	56234	54.92
241	1514	1.48	57748	56.4
242	4893	4.78	62641	61.17
243	5103	4.98	67744	66.16
244	4976	4.86	72720	71.02
245	1551	1.51	74271	72.53
246	3339	3.26	77610	75.79
247	4795	4.68	82405	80.48
248	1524	1.49	83929	81.96
249	3141	3.07	87070	85.03
250	2853	2.79	89923	87.82
251	1414	1.38	91337	89.2
252	3804	3.71	95141	92.91
254	2192	2.14	97333	95.05
255	972	0.95	98305	96
256	890	0.87	99195	96.87
257	1476	1.44	100671	98.31

259	467	0.46	101138	98.77
260	843	0.82	101981	99.59
263	307	0.3	102288	99.89
264	109	0.11	102397	100

Frequency Missing 403

Grade 3 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
216	1	0	1	0
217	3	0	4	0
219	9	0.01	13	0.01
220	8	0.01	21	0.02
221	36	0.04	57	0.06
222	66	0.06	123	0.12
223	122	0.12	245	0.24
224	365	0.36	610	0.6
225	412	0.4	1022	1
226	494	0.48	1516	1.48
227	590	0.58	2106	2.06
228	1013	0.99	3119	3.05
229	935	0.91	4054	3.96
230	1020	1	5074	4.96
231	684	0.67	5758	5.63
232	836	0.82	6594	6.44
233	1184	1.16	7778	7.6
234	1255	1.23	9033	8.83
235	1379	1.35	10412	10.17
236	2445	2.39	12857	12.56
237	1513	1.48	14370	14.04
238	1717	1.68	16087	15.72
239	1776	1.74	17863	17.45
240	2633	2.57	20496	20.03
241	2860	2.79	23356	22.82
242	3035	2.97	26391	25.79
243	3278	3.2	29669	28.99
244	2767	2.7	32436	31.69
245	3800	3.71	36236	35.4
246	6608	6.46	42844	41.86
247	3746	3.66	46590	45.52
248	3947	3.86	50537	49.38
249	5286	5.16	55823	54.54
250	4538	4.43	60361	58.98
251	4595	4.49	64956	63.47
252	4776	4.67	69732	68.13
253	3194	3.12	72926	71.25
254	3412	3.33	76338	74.59
255	5108	4.99	81446	79.58
256	5016	4.9	86462	84.48
257	1418	1.39	87880	85.86
258	3208	3.13	91088	89
259	2610	2.55	93698	91.55
260	1362	1.33	95060	92.88
261	2283	2.23	97343	95.11
262	1010	0.99	98353	96.1

263	1496	1.46	99849	97.56
265	1117	1.09	100966	98.65
266	511	0.5	101477	99.15
267	241	0.24	101718	99.38
268	286	0.28	102004	99.66
269	241	0.24	102245	99.9
271	63	0.06	102308	99.96
272	41	0.04	102349	100

Frequency Missing 4204

Grade 4 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
224	1	0	1	0
225	2	0	3	0
226	11	0.01	14	0.01
227	45	0.04	59	0.06
228	163	0.16	222	0.22
229	203	0.2	425	0.42
230	314	0.31	739	0.74
231	406	0.4	1145	1.14
232	810	0.81	1955	1.95
233	820	0.82	2775	2.76
234	849	0.84	3624	3.61
235	665	0.66	4289	4.27
236	600	0.6	4889	4.87
237	1125	1.12	6014	5.99
238	1138	1.13	7152	7.12
239	1315	1.31	8467	8.43
240	1498	1.49	9965	9.92
241	1577	1.57	11542	11.49
242	2266	2.26	13808	13.74
243	2590	2.58	16398	16.32
244	2115	2.1	18513	18.42
245	2207	2.2	20720	20.62
246	3138	3.12	23858	23.74
247	3485	3.47	27343	27.21
248	3764	3.75	31107	30.96
249	4072	4.05	35179	35.01
250	3218	3.2	38397	38.21
251	4501	4.48	42898	42.69
252	4886	4.86	47784	47.55
253	5432	5.41	53216	52.96
254	5280	5.25	58496	58.21
255	4267	4.25	62763	62.46
256	4458	4.44	67221	66.9
257	4611	4.59	71832	71.49
258	2786	2.77	74618	74.26
259	4690	4.67	79308	78.93
260	4446	4.42	83754	83.35
261	1531	1.52	85285	84.88
262	4111	4.09	89396	88.97
263	2394	2.38	91790	91.35
264	1180	1.17	92970	92.52
265	2089	2.08	95059	94.6
266	1732	1.72	96791	96.33
267	819	0.82	97610	97.14
268	685	0.68	98295	97.82
269	522	0.52	98817	98.34

270	916	0.91	99733	99.25
271	257	0.26	99990	99.51
272	191	0.19	100181	99.7
273	200	0.2	100381	99.9
275	102	0.1	100483	100

Frequency Missing 4103

Grade 5 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
230	1	0	1	0
231	11	0.01	12	0.01
232	25	0.02	37	0.04
233	82	0.08	119	0.11
234	143	0.14	262	0.25
235	258	0.25	520	0.5
236	327	0.32	847	0.82
237	479	0.46	1326	1.28
238	585	0.56	1911	1.84
239	741	0.72	2652	2.56
240	838	0.81	3490	3.37
241	881	0.85	4371	4.22
242	1002	0.97	5373	5.19
243	1056	1.02	6429	6.21
244	1143	1.1	7572	7.31
245	1745	1.68	9317	8.99
246	2428	2.34	11745	11.34
247	1657	1.6	13402	12.94
248	1720	1.66	15122	14.6
249	3918	3.78	19040	18.38
250	2178	2.1	21218	20.48
251	3322	3.21	24540	23.69
252	4368	4.22	28908	27.9
253	4024	3.88	32932	31.79
254	5430	5.24	38362	37.03
255	3617	3.49	41979	40.52
256	5155	4.98	47134	45.5
257	6880	6.64	54014	52.14
258	4393	4.24	58407	56.38
259	4650	4.49	63057	60.87
260	4625	4.46	67682	65.33
261	4936	4.76	72618	70.1
262	4845	4.68	77463	74.77
263	4902	4.73	82365	79.51
264	3281	3.17	85646	82.67
265	4561	4.4	90207	87.07
266	1373	1.33	91580	88.4
267	2799	2.7	94379	91.1
268	2423	2.34	96802	93.44
269	2233	2.16	99035	95.6
270	846	0.82	99881	96.41
271	1648	1.59	101529	98
273	1068	1.03	102597	99.03
274	449	0.43	103046	99.47
276	368	0.36	103414	99.82
277	183	0.18	103597	100

Frequency Missing 4220

Grade 6 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
229	1	0	1	0
230	5	0	6	0.01
231	13	0.01	19	0.02
232	39	0.04	58	0.06
233	99	0.09	157	0.15
234	123	0.12	280	0.27
235	368	0.35	648	0.62
236	316	0.3	964	0.92
237	444	0.42	1408	1.34
238	553	0.53	1961	1.87
239	600	0.57	2561	2.44
240	688	0.66	3249	3.1
241	770	0.73	4019	3.83
242	864	0.82	4883	4.66
243	930	0.89	5813	5.55
244	1054	1.01	6867	6.55
245	1085	1.04	7952	7.59
246	1201	1.15	9153	8.73
247	1278	1.22	10431	9.95
248	1941	1.85	12372	11.8
249	2625	2.5	14997	14.31
250	1824	1.74	16821	16.05
251	2568	2.45	19389	18.5
252	3464	3.3	22853	21.8
253	3202	3.05	26055	24.86
254	4280	4.08	30335	28.94
255	3568	3.4	33903	32.35
256	4845	4.62	38748	36.97
257	5199	4.96	43947	41.93
258	3338	3.18	47285	45.11
259	7213	6.88	54498	51.99
260	3780	3.61	58278	55.6
261	4073	3.89	62351	59.49
262	6858	6.54	69209	66.03
263	4282	4.09	73491	70.11
264	4412	4.21	77903	74.32
265	4496	4.29	82399	78.61
266	4326	4.13	86725	82.74
267	2705	2.58	89430	85.32
268	2780	2.65	92210	87.97
269	2707	2.58	94917	90.56
270	2267	2.16	97184	92.72
271	2048	1.95	99232	94.67
272	1869	1.78	101101	96.46
274	1957	1.87	103058	98.32
276	441	0.42	103499	98.74

277	759	0.72	104258	99.47
280	383	0.37	104641	99.83
281	113	0.11	104754	99.94
283	62	0.06	104816	100

Frequency Missing 3649

Grade 7 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
228	2	0	2	0
229	4	0	6	0.01
230	4	0	10	0.01
231	4	0	14	0.01
232	23	0.02	37	0.04
233	44	0.04	81	0.08
234	70	0.07	151	0.14
235	112	0.11	263	0.25
236	231	0.22	494	0.47
237	282	0.27	776	0.74
238	401	0.38	1177	1.12
239	560	0.53	1737	1.65
240	719	0.69	2456	2.34
241	762	0.73	3218	3.07
242	390	0.37	3608	3.44
243	1001	0.95	4609	4.39
244	1051	1	5660	5.39
245	1104	1.05	6764	6.44
246	1217	1.16	7981	7.6
247	1275	1.21	9256	8.82
248	1346	1.28	10602	10.1
249	1501	1.43	12103	11.53
250	1648	1.57	13751	13.1
251	1727	1.65	15478	14.75
252	1842	1.75	17320	16.5
253	2041	1.94	19361	18.45
254	4383	4.18	23744	22.62
255	2382	2.27	26126	24.89
256	2633	2.51	28759	27.4
257	4183	3.99	32942	31.39
258	2998	2.86	35940	34.24
259	6423	6.12	42363	40.36
260	3364	3.21	45727	43.57
261	5271	5.02	50998	48.59
262	3610	3.44	54608	52.03
263	3814	3.63	58422	55.66
264	5708	5.44	64130	61.1
265	4042	3.85	68172	64.95
266	6327	6.03	74499	70.98
267	4169	3.97	78668	74.95
268	4043	3.85	82711	78.8
269	3984	3.8	86695	82.6
270	3762	3.58	90457	86.18
271	1774	1.69	92231	87.87
272	3250	3.1	95481	90.97
273	2899	2.76	98380	93.73

274	1260	1.2	99640	94.93
275	1131	1.08	100771	96.01
276	968	0.92	101739	96.93
277	1606	1.53	103345	98.46
279	973	0.93	104318	99.39
281	261	0.25	104579	99.64
282	200	0.19	104779	99.83
284	99	0.09	104878	99.92
285	59	0.06	104937	99.98
287	22	0.02	104959	100

Frequency Missing 3627

Grade 8 Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
232	5	0	5	0
233	3	0	8	0.01
234	8	0.01	16	0.02
235	22	0.02	38	0.04
236	44	0.04	82	0.08
237	99	0.1	181	0.18
238	317	0.31	498	0.49
239	309	0.3	807	0.79
240	398	0.39	1205	1.18
241	477	0.47	1682	1.65
242	396	0.39	2078	2.03
243	425	0.42	2503	2.45
244	705	0.69	3208	3.14
245	791	0.77	3999	3.91
246	917	0.9	4916	4.81
247	937	0.92	5853	5.73
248	1060	1.04	6913	6.76
249	718	0.7	7631	7.47
250	871	0.85	8502	8.32
251	1226	1.2	9728	9.52
252	1475	1.44	11203	10.96
253	1443	1.41	12646	12.37
254	2362	2.31	15008	14.69
255	2460	2.41	17468	17.09
256	2011	1.97	19479	19.06
257	3157	3.09	22636	22.15
258	2490	2.44	25126	24.59
259	3677	3.6	28803	28.19
260	5039	4.93	33842	33.12
261	3209	3.14	37051	36.26
262	3334	3.26	40385	39.52
263	5928	5.8	46313	45.32
264	3820	3.74	50133	49.06
265	5146	5.04	55279	54.09
266	4018	3.93	59297	58.03
267	5347	5.23	64644	63.26
268	4056	3.97	68700	67.23
269	4279	4.19	72979	71.41
270	4086	4	77065	75.41
271	4091	4	81156	79.42
272	4041	3.95	85197	83.37
273	3701	3.62	88898	86.99
274	2433	2.38	91331	89.37
275	2100	2.05	93431	91.43
276	1764	1.73	95195	93.15
277	2427	2.37	97622	95.53

278	877	0.86	98499	96.39
279	1028	1.01	99527	97.39
280	615	0.6	100142	97.99
281	643	0.63	100785	98.62
282	459	0.45	101244	99.07
283	343	0.34	101587	99.41
284	275	0.27	101862	99.68
286	139	0.14	102001	99.81
287	110	0.11	102111	99.92
289	40	0.04	102151	99.96
290	41	0.04	102192	100

Frequency Missing 5586

High School Comprehensive Scale Score Frequency Distribution (1998)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
132	3	0	3	0
133	14	0.02	17	0.02
134	13	0.02	30	0.04
135	13	0.02	43	0.06
136	20	0.03	63	0.09
137	37	0.05	100	0.14
138	145	0.2	245	0.33
139	173	0.23	418	0.57
140	263	0.36	681	0.92
141	462	0.62	1143	1.55
142	543	0.73	1686	2.28
143	630	0.85	2316	3.13
144	810	1.09	3126	4.23
145	944	1.28	4070	5.5
146	358	0.48	4428	5.99
147	1118	1.51	5546	7.5
148	744	1.01	6290	8.5
149	1213	1.64	7503	10.14
150	1297	1.75	8800	11.9
151	1399	1.89	10199	13.79
152	1411	1.91	11610	15.69
153	1540	2.08	13150	17.78
154	1616	2.18	14766	19.96
155	2233	3.02	16999	22.98
156	1714	2.32	18713	25.3
157	2252	3.04	20965	28.34
158	1867	2.52	22832	30.86
159	2018	2.73	24850	33.59
160	4011	5.42	28861	39.01
161	1927	2.6	30788	41.62
162	2066	2.79	32854	44.41
163	3242	4.38	36096	48.79
164	3377	4.56	39473	53.36
165	1955	2.64	41428	56
166	3333	4.51	44761	60.51
167	2670	3.61	47431	64.12
168	2535	3.43	49966	67.54
169	2514	3.4	52480	70.94
170	2450	3.31	54930	74.25
171	2245	3.03	57175	77.29
172	2239	3.03	59414	80.32
173	2230	3.01	61644	83.33
174	1563	2.11	63207	85.44
175	1883	2.55	65090	87.99
176	1743	2.36	66833	90.34
177	1699	2.3	68532	92.64

178	1069	1.45	69601	94.09
179	1195	1.62	70796	95.7
180	772	1.04	71568	96.74
181	678	0.92	72246	97.66
182	361	0.49	72607	98.15
183	153	0.21	72760	98.36
184	437	0.59	73197	98.95
185	296	0.4	73493	99.35
186	181	0.24	73674	99.59
187	97	0.13	73771	99.72
188	49	0.07	73820	99.79
189	65	0.09	73885	99.88
190	32	0.04	73917	99.92
191	13	0.02	73930	99.94
192	27	0.04	73957	99.97
194	14	0.02	73971	99.99
197	2	0	73973	100
199	3	0	73976	100

Frequency Missing = 5471

English I Scale Score Frequency Distribution (2003)

Scale Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
29	4	0	4	0
30	5	0.01	9	0.01
31	2	0	11	0.01
32	7	0.01	18	0.02
33	28	0.03	46	0.05
34	77	0.08	123	0.12
35	122	0.12	245	0.25
36	361	0.36	606	0.61
37	289	0.29	895	0.9
38	382	0.38	1277	1.28
39	646	0.65	1923	1.93
40	812	0.81	2735	2.74
41	657	0.66	3392	3.4
42	815	0.82	4207	4.22
43	1103	1.11	5310	5.32
44	921	0.92	6231	6.25
45	1310	1.31	7541	7.56
46	1438	1.44	8979	9
47	1161	1.16	10140	10.17
48	2039	2.04	12179	12.21
49	1779	1.78	13958	13.99
50	2940	2.95	16898	16.94
51	1556	1.56	18454	18.5
52	3376	3.38	21830	21.88
53	4416	4.43	26246	26.31
54	3969	3.98	30215	30.29
55	4439	4.45	34654	34.74
56	4789	4.8	39443	39.54
57	5088	5.1	44531	44.64
58	6122	6.14	50653	50.78
59	5544	5.56	56197	56.34
60	5779	5.79	61976	62.13
61	5847	5.86	67823	67.99
62	3911	3.92	71734	71.91
63	4805	4.82	76539	76.73
64	5658	5.67	82197	82.4
65	2539	2.55	84736	84.95
66	4050	4.06	88786	89.01
67	2901	2.91	91687	91.91
68	1966	1.97	93653	93.89
69	1709	1.71	95362	95.6
70	1368	1.37	96730	96.97
71	1185	1.19	97915	98.16
72	282	0.28	98197	98.44
73	578	0.58	98775	99.02

74	527	0.53	99302	99.55
76	216	0.22	99518	99.77
77	86	0.09	99604	99.85
78	102	0.1	99706	99.95
79	17	0.02	99723	99.97
80	29	0.03	99752	100
Frequency Missing			3044	

Appendix C: Developmental Scale Report

The Developmental Scale for the North Carolina End of Grade Reading Tests, Second Edition

David Thissen, Michael C. Edwards, Cheryl D. Coon, and Carol M. Woods
L.L. Thurstone Psychometric Laboratory
The University of North Carolina at Chapel Hill
November 15, 2002

Following changes in the North Carolina curricular specifications for Reading, a second edition of the North Carolina End of Grade tests in reading has been designed, and an item tryout was administered as a field test in the Spring of 2002. This report describes the use of data from that item tryout to construct a developmental scale for the second edition of the North Carolina End of Grade tests in Reading.

The basis of a developmental scale is the specification of the means and standard deviations for scores on that scale for each grade-level. In the case of the North Carolina End of Grade tests the grade levels range from the grade 3 pretest (administered in the Fall to students in the third grade) through grade 8. The data from which the scale-score means and standard deviations are derived make use of special test forms (called “linking forms”) that are administered to students in adjacent grades. The difference in performance between grades on these forms is used to estimate the difference in proficiency between grades. The second edition of the NC End of Grade Reading tests used item response theory (IRT) to compute these estimates following procedures described by Williams, Pommerich, and Thissen (1998). The population means and standard deviations derived from the Spring 2002 item calibration for the NC End of Grade Reading tests are shown in Table 1.

Table 1. Population means and standard deviations derived from the Spring 2002 item calibration for the NC End of Grade Reading tests, second edition.

Grade	Population	
	Mean	Standard Deviation
3 Pretest	236.66	11.03
3	245.21	10.15
4	250.00	10.00
5	253.92	9.61
6	255.57	10.41
7	256.74	10.96
8	259.35	11.13

The values for the developmental scale shown in Table 1 are based on IRT estimates of differences between adjacent-grade means and ratios of adjacent-grade standard deviations computed using the computer program MULTLOG (Thissen, 1991); the

estimates from MULTILOG were cross-checked against parallel estimates computed using the software IRTLRFID (Thissen, 2001). In the computation of estimates using either software system, the analysis of data from adjacent grades arbitrarily sets the mean and standard deviation of the population distribution of the lower grade to values of zero (0) and one (1), respectively; the values of the mean (μ) and standard deviation (σ) of the higher grade are estimated making use of the item response data and the three-parameter logistic IRT model (Thissen and Orlando, 2001). Table 2 shows the average difference between adjacent-grade means (μ) in units of the standard deviation of the lower grade, and ratios between adjacent-grade standard deviations (σ), derived from the Spring 2002 item calibration for the NC End of Grade Reading tests. The values in Table 2 were converted into the final scale, shown in Table 1, by (arbitrarily) setting the average scale score in grade 4 to be 250, with a standard deviation of 10, and then computing the values for the other grades such that the differences between the means for adjacent grades, in units of the standard deviation of the lower grade, are the same as those shown in Table 2.

Table 2. Average difference between adjacent-grade means (μ) in units of the standard deviation of the lower grade, and ratios between adjacent-grade standard deviations (σ), derived from the spring, 2002 item calibration for the NC End of Grade Reading tests, second edition.

Grades	Average μ Difference	Average s Ratio	Replications
3P-3	0.77	0.92	3
3-4	0.47	0.99	10
4-5	0.39	0.96	10
5-6	0.17	1.08	19
6-7	0.11	1.05	5
7-8	0.24	1.02	10

The averages shown in Table 2 are over between three and 19 replications of the between-grade difference; the numbers of replications for each grade pair are also shown in Table 2. Each replication is based on a (different) linking form from among the item tryout forms administered in the Spring 2002 field test. (There were different numbers of forms for different between-grade comparisons, because in some pairs of grades item tryout forms were administered in more than one grade for other purposes in addition to linking; all available data were used to construct the new developmental scale.) The sample size for each linking form was approximately 1400 students in each grade.

Table 3 shows, for each adjacent-grade pair, the values of the average difference between adjacent-grade means (μ) in units of the standard deviation of the lower grade, and ratios of adjacent-grade standard deviations (σ), derived from the Spring 2002 item calibration for the NC End of Grade Reading tests for each replication. In Table 3 the values for each grade-pair are in decreasing order of the estimate of the difference between the means. The standard deviations of the mean-differences range from 0.03 to 0.08, which are

values close to 0.04, the theoretical standard error of the mean for differences between samples of about 1400. The standard deviation ratios in Table 3 vary extremely little.

In addition to the linking forms between adjacent grades that were used to construct the developmental scale, there were several linking forms that were administered in grades 4, 5, and 6, and several other forms that were administered in grades 5, 6, and 7. IRT analyses of those “triplet” forms yield estimates of the difference over the two-year spans between grades 4 and 6, and between grades 5 and 7. In Table 4 the averages of those two-year links are compared with the corresponding combined average one-year links (from Table 2); the results show good agreement.

Table 3. Replications of the average difference between adjacent-grade means (μ) in units of the standard deviation of the lower grade, and ratios between adjacent-grade standard deviations (σ), derived from the Spring 2002 item calibration for the NC End of Grade Reading tests, second edition.

Grade 3P-3		Grades 3-4		Grades 4-5		Grades 5-6		Grades 6-7		Grades 7-8	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
0.80	0.90	0.57	0.95	0.51	0.91	0.35	1.08	0.17	1.06	0.32	0.94
0.78	0.93	0.54	1.00	0.45	0.93	0.30	1.07	0.13	1.03	0.31	1.01
0.74	0.93	0.49	0.96	0.42	0.99	0.25	1.06	0.10	1.06	0.26	0.97
		0.49	1.06	0.41	0.93	0.22	1.03	0.09	1.04	0.24	0.99
		0.48	0.98	0.38	0.93	0.22	1.04	0.06	1.08	0.23	1.02
		0.47	1.00	0.37	0.93	0.21	1.06			0.21	0.98
		0.45	0.95	0.36	0.99	0.20	1.06			0.21	1.06
		0.45	0.97	0.36	0.99	0.20	1.04			0.21	1.09
		0.40	0.99	0.33	1.06	0.17	1.05			0.20	1.07
		0.38	1.01	0.33	0.94	0.17	1.11			0.19	1.03
						0.15	1.15				
						0.15	1.08				
						0.14	1.11				
						0.11	1.11				
						0.11	1.12				

0.09	1.07
0.08	1.07
0.07	1.16
0.06	1.11

Table 4. Comparison between the average difference between means (μ) in units of the standard deviation of the lower grade grades, for grades 4 and 6, and 5 and 7, and ratios between the corresponding standard deviations (σ), with values obtained by combining the adjacent-grade estimates.

Two-Year Links			One-Year Links Combined		
Grades	Mean	Standard Deviation	Grades	Mean	Standard Deviation
4-6	0.54	1.05	4-5 & 5-6	0.55	1.04
5-7	0.31	1.13	5-6 & 6-7	0.29	1.13

Comparison with and Linkage to the First Edition Scale

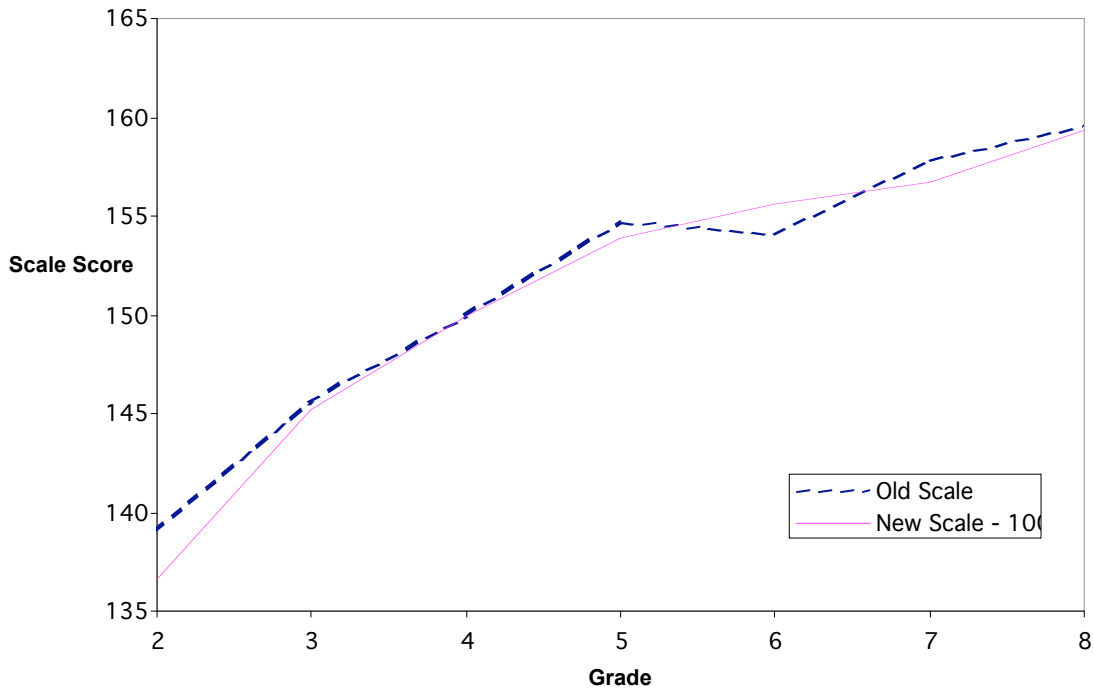
To provide a basis for linkage of the second edition developmental scale with that of the first edition, one form of the first edition was administered to a random sample of the students in the Spring 2002 field test, spiraled with the item-tryout forms for the second edition. Table 5 shows a comparison of the population means and standard deviations for the second edition with the averages and standard deviations for the scale scores obtained from the spiraled administration of the first edition. For ease of comparison of the two scales, Figure 1 shows the two sets of averages plotted together, with 100 subtracted from the *2nn* values of the new scale so they use approximately the same range. The forms of the developmental scales for the first and second editions of the Reading test are very similar. [The careful reader will note that, in Table 5, the second edition standard deviations are somewhat larger than those for the first edition. That is due to the fact that the standard deviations for the second edition are the values for the population distribution and those for the first edition are standard deviations of the scale scores themselves; the latter must be somewhat smaller than the former for IRT scale scores.]

Table 5. Comparison of the population means and standard deviations for the second edition with the averages and standard deviations obtained from the spiraled administration of the first edition in the Spring 2002 item calibration for the NC End of Grade Reading tests.

Grade	First Edition		Second Edition	
	Mean	Standard Deviation	Mean	Standard Deviation
3 Pretest	139.02	8.00	236.66	11.03
3	145.59	9.62	245.21	10.15

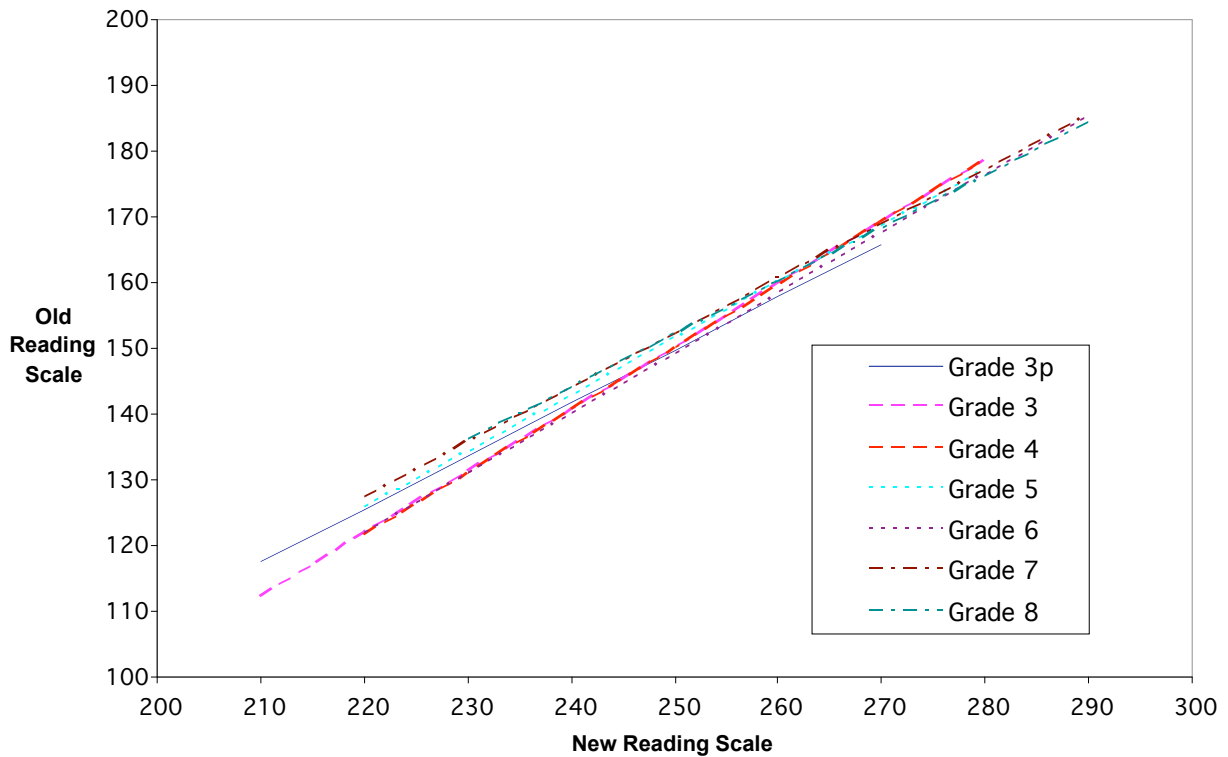
4	149.98	9.50	250.00	10.01
5	154.74	8.21	253.92	9.61
6	154.08	9.44	255.57	10.41
7	157.81	9.09	256.74	10.96
8	159.55	8.96	259.35	11.13

Figure 1. Reading Developmental Scales from Pretest



The values in Table 5 provide the basis for the calculation of linear equating functions between the score-scales for the first and second editions of the reading test; those seven functions (one for each grade-level) are shown in Figure 2. For the most part, the equating lines in Figure 2 exhibit patterns that are as one would expect: The functions for grades 3 and 4 are nearly collinear, as are those for grades 5-8. The slope of the equating functions for grades 3 and 4 is steeper than those for grades 5-8; that is to be expected given the decelerating nature of the growth curves.

Figure 2. Pretest-Based Linear Equating



Anomalies

There are, nevertheless, some anomalies in the item tryout data as they relate to the first edition of the test. One anomaly is that the linear equating function for the grade 3 pretest is somewhat different from those for grades 3 and 4. That is probably related to the fact that the second edition developmental curve exhibits greater growth between the grade 3 pretest and the grade 3 (Spring) testing than do the scores on the first edition (see Table 5 and Figure 1).

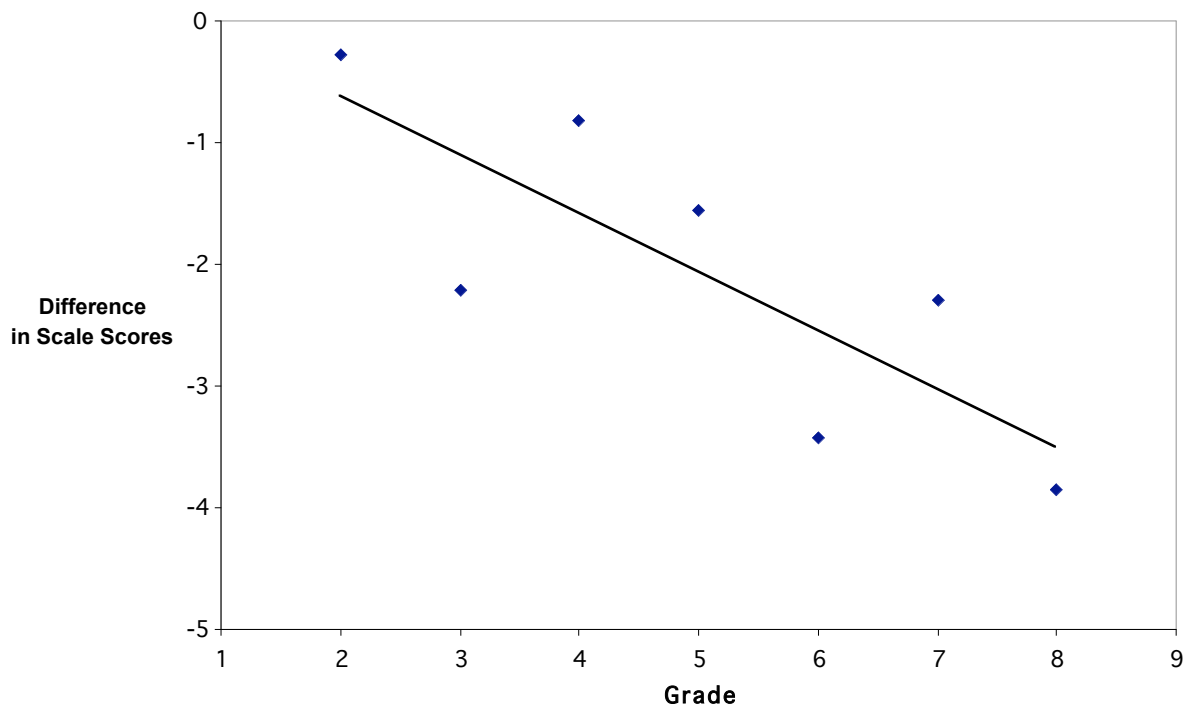
A second anomaly in the first-edition data from the item tryout involves the data for grade 6, for which the average score for the sample that was administered the first edition is lower than the average score for grade 5 (see Table 5 and Figure 1). This was not the case for the statewide operational administration of the first edition in the Spring of 2002; see Table 6 for a comparison of the pretest sample means and standard deviations with the statewide operational values.

Table 6. Comparison of the population means and standard deviations for the second edition with the averages and standard deviations obtained from the spiraled administration of the first edition in the spring 2002 item calibration for the NC End of Grade Reading tests.

Grade	Item Tryout 2002		Statewide Data, spring 2002	
	Mean	Standard Deviation	Mean	Standard Deviation
3 Pretest	139.02	8.00	139.3	8.5
3	145.59	9.62	147.8	8.9
4	149.98	9.50	150.8	9.0
5	154.74	8.21	156.3	7.9
6	154.08	9.44	157.5	9.1
7	157.81	9.09	160.1	8.3
8	159.55	8.96	163.4	7.8

These anomalies suggest that there is some possibility that the first-edition data from the field test samples may be somewhat unusual for the grade 3 pretest and/or for grade 3, and likely for grade 6. The most likely scenario is that the grade 3 and grade 6 samples happened to score a little low on the first edition scale (largely by chance?). Figure 3 shows a plot of the differences between the average first-edition scale scores from the pretest sample and those from operational administration in spring 2002. We note the usual trend that those differences increase from roughly zero for the grade 3 pretest to about 4 scale-score points for grade 8; as students become more aware of the lack of consequences associated with item tryout administrations, scores decrease. However, we also note that the pretest averages for grades 3 and 6 are unusually low.

Figure 3. Difference in Scale Scores, Pretest - Operational



Discussion, Future Tasks, and Potential Challenges

The newly-constructed developmental scale for the second edition of the NC End of Grade Reading test is very similar in form to the developmental scale that has been in use for the first edition. The new scale also exhibited a good deal of consistency across replications in the construction of its components. These facts suggest that the new developmental scale itself is reliable in the sense that it is consistent across replications and valid in the sense that it is in agreement with the independently-constructed first-edition scale.

Tasks that remain to be completed include the creation of scoring tables for the second edition forms that are to be administered in the Spring of 2002, and the construction of “equating” tables relating scores on the second edition’s *2nn* scale with scores on the first edition’s *1nn* scale.

Construction of scoring tables awaits the assembly of the forms themselves. Then, using the item parameters from the item tryout and the population means and standard deviations in Table 1, we will construct scoring tables using the procedures described by Thissen, Pommerich, Billeaud, and Williams (1995) and Thissen and Orlando (2001). These procedures yield tables that translate summed scores into corresponding IRT scale scores on the developmental scale.

A side-effect of the construction of those scoring tables is that the algorithm provides model-based estimates of the proportions of the item tryout samples that would have obtained each summed score (and hence, each scale score). Those score-proportions can be matched with the observed score distributions on the first-edition forms that were included in the item tryout, yielding equipercentile equating tables that may be used *with caution* in the computations of score-changes that are involved in the accountability system. (The linear equating functions in Figure 2 will likely not be used for that purpose, unless the equipercentile computations yield results that are very similar to those linear relations.)

Caution will be required in the use of those (as yet unavailable) linking tables, because the anomalies cited above for the data from grade 3 and/or grade 6 may yield score translation tables that are consistently too high or too low for those grades.

After the new second edition forms are constructed, additional information will also be available for the assembly of the final linking tables between the first-edition and second-edition score scales. That additional information takes the form of the consistency in the patterns of the matched cut scores between the EOG Levels I, II, III, and IV across grades. It is possible that a combination of smoothing of the trends of those cut scores across grades with the within-grade equipercentile data may yield results that will “smooth over” the unusual results from the item tryout data.

However, there are no guarantees that there will not be surprising results the first time the second-edition reading tests are administered. It is simply not possible to administer different tests, based on different curricula, in two successive years and expect the results to be in all senses as-expected, or even equatable. Nevertheless, the NC accountability system requires that that be done. The new second edition scale reported herein, combined with subsequent equipercentile equating of the first edition scale with the to-be-constructed operational forms for the second edition, represents our best attempt to facilitate that goal.

Thissen, D. (1991). *MULTILOG user's guide—Version 6*. Chicago, IL: Scientific Software, Inc.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Unpublished ms.

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.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V.S.L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

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.

Appendix D: Testing Code of Ethics

Testing Code of Ethics (16 NCAC 6D .0306)

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.

Testing Code of Ethics (16 NCAC 6D .0306)

.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) disaggregation 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.