DOCUMENT RESUME

ED 406 397 TM 026 261

TITLE	North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics. Technical Report No. 1.
INSTITUTION	North Carolina State Dept. of Public Instruction, Raleigh. Div. of Accountability/Testing.
PUB DATE	Aug 96
NOTE	165p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Achievement Tests; Elementary Education; Elementary School Students; *Mathematics Tests; Psychometrics; *Reading Comprehension; Reading Tests; Scaling; Scoring; State Programs; *Test Construction; *Testing Programs; Test Items; Test Reliability; Test Use; Test Validity
IDENTIFIERS	*North Carolina End of Grade Testing Program

ABSTRACT

The North Carolina End-of-Grade Testing Program is based on the assessment of higher level skills in the context of specific subject-area content. These tests inform students, parents, the community, and educators about the achievement of North Carolina students in grades three through eight in given areas. This report describes the development and psychometric properties of the end-of-grade tests in Reading Comprehension and Mathematics. The value of these tests lies primarily in the fact that the scores provide a common yardstick that is not influenced by local differences. The reading comprehension tests assess a student's ability to comprehend written material that is appropriate for the grade level and the ability to use strategies to enhance reading comprehension. The mathematics tests assess computation and mathematics applications. The tests described in this publication are administered during the last 3 weeks of the school year. For both sets of tests, the report describes: (1) item development; (2) test development; (3) scores and scales; (4) descriptive statistics and reliability; and (5) validity. Six appendixes provide sample test items and discuss aspects of test construction and scoring in detail. (Contains 26 tables, 25 figures, 20 appendix graphs, 12 appendix tables, and 37 references.) (SLD)

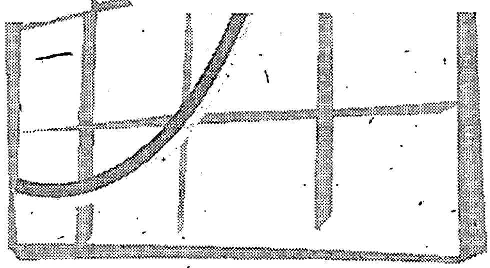
TM

Technical Report #1

ED 406 397

NORTH CAROLINA

End-of-Grade Tests



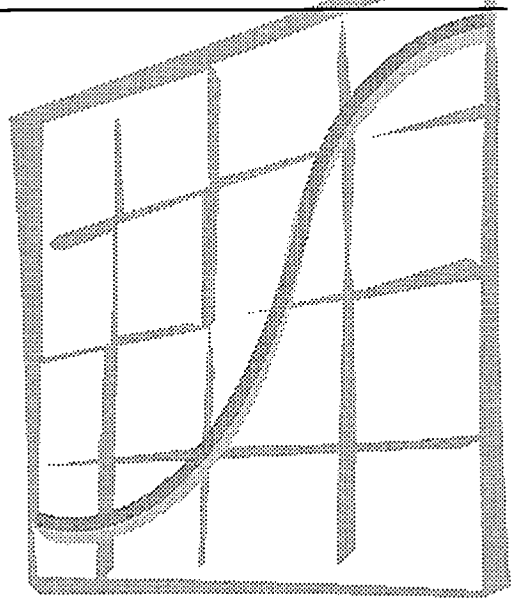
Reading Comprehension
Mathematics

end of grade testing

NCTests

Public Schools of North Carolina
State Board of Education
Department of Public Instruction
Office of Instructional and Accountability Services
Division of Accountability/Testing

End-of-Grade Tests



Reading Comprehension
Mathematics

Prepared by

Eleanor E. Sanford, Ph.D.
August 1996

Technical Support	Nada Ballator, Jeff Morgan, Laura Kramer, and Lori D. McLeod
	L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill: David Thissen, Val S.L. Williams, Mary Pommerich, Kathy Billeaud, and Lori McLeod.
Test Production	Shirley Stoll, Susan Godwin, Lynette Rivenbark, Adrienne Silvay, Patricia Atkinson, and Lisa Powell
Others	Chris Averett, Suzanne Triplett, William J. Brown, Bob Evans, Daisy Vickers, Mildred Bazemore, Gary Williamson, and Doris Tyler

For additional information contact

Eleanor E. Sanford, Ph.D.
NCDPI/Testing
301 North Wilmington Street
Raleigh, North Carolina 27601-2825
919/715-1214
e-mail: esanford@dpi.state.nc.us

Table of Contents

List of Tables

List of Figures

During this decade and for many decades to come, North Carolina students will need to move far beyond the mastery of basic skills to the mastery of higher level skills. The term "higher level skills" refers to the thinking and problem solving strategies that enable people to access, sort, and digest enormous amounts of information. It refers to the skills required to solve complex problems and to make informed choices and decisions. It also refers to advanced communication skills that enable individuals to express and share what they know and to work well with others (North Carolina End-of-Grade Testing Program: Background Information, 1993, p. 1).

The End-of-Grade Testing Program is based on the assessment of these higher level skills. When properly administered and interpreted, these 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 objectively assess their return on investment in the public schools

(*North Carolina Testing Code of Ethics*, revised 1996).

This technical report describes the development and psychometric properties of the reading comprehension and mathematics tests of the North Carolina End-of-Grade Testing Program for grades 3 through 8.

Background

The North Carolina End-of-Grade Testing Program was initiated in response to legislation passed by the North Carolina General Assembly. The following sections of the *Public School Laws* (1994) describe the legislation.

> Public School Law 115C-174.10 defines the following "purposes of testing programs in North Carolina: (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."
>
> Public School Law 115C-174.11 (c) calls for the "adoption of a system of end-of-grade testing designed to measure progress toward selected competencies, especially core academic competencies, described in the *Standard Course of Study* for appropriate grade levels. With regard to students who are identified as not demonstrating satisfactory academic progress, end-of-grade test results shall be used in developing strategies and plans for assisting those students in achieving satisfactory academic progress."

Based on these statutes, the North Carolina End-of-Grade Testing Program was developed for two purposes:

- to provide accurate measurement of individual student skills and knowledge specified in the North Carolina *Standard Course of Study*, and
- to provide accurate measurement of the knowledge and skills attained by groups of students for school, school system, and state accountability.

Scores on the end-of-grade tests are only one of many indicators of the achievement of students. The value of these tests lies primarily in the fact that the scores provide a common standard that is not influenced by local differences in achievement and expectations. The tests provide yardsticks which can be used to compare the achievement of students, schools, school systems, and the state. The assessment yardstick can be used to measure gains (or losses) in performance across time to see if educational improvement efforts at the state and local level are working.

The North Carolina End-of-Grade Testing Program includes multiple choice assessments of reading comprehension and mathematics in grades 3 through 8. Writing is assessed in grades 4 and 7 during March and, beginning in the Fall of 1996, integrated skills (open-ended format) will be assessed in grades 5 and 8. This open-ended assessment will measure a student's reading and mathematics proficiency, while integrating reading and mathematics skills in the context of social studies and science topics. Computer skills are assessed in grade 8: beginning with the class of 2001, the Quality Assurance Plan requires each student to "demonstrate computer competence" in order to graduate from high school. All of the tests are developed by the North Carolina Department of Public Instruction and are aligned with the revised North Carolina *Standard Course of Study*, and each part of the curricula is assessed in an efficient manner.

Table 1. North Carolina End-of-Grade Testing Program.

	Reading	Mathematics	Social Studies	Science	Writing	Open-Ended	Computer Skills
3							
4							
5							
6							
7							
8							

In response to Senate Bill 16 passed in 1995, the State Board of Education examined the structure and functions of the state public school system in order to improve student performance, increase local flexibility and control, and promote economy and efficiency. In May 1995, the State Board of Education issued *The New ABCs of Public Education: Accountability, curriculum Basics, and local Control and flexibility*. One of the key recommendations was to adopt a new accountability plan focused on the performance of individual public schools (rather than school systems) on the basics of reading, communication skills, and mathematics. Rather than comparing different students over time, this plan—the School-based Management and Accountability Program (legislation under consideration during the 1996 session of the North Carolina General Assembly)—will hold schools accountable for the educational growth of groups of students over time.

For school, school system, and state accountability, the multiple-choice reading comprehension and mathematics tests are used to monitor growth. The scores from a prior grade (for example, grade 5) are used as to determine a student's entering level of knowledge and skills and to determine the amount of growth during the school year (for example, grade 6). "The state will set standards for growth for the given amount of schooling, expecting at least a year's worth of growth for a year's worth of school" (North Carolina State Board of Education, 1996).

For student accountability, the grade 8 end-of-grade tests are used as a way for students to demonstrate that they have the knowledge and skills necessary to meet the reading and mathematics competency requirement for high school graduation. The Grade-Level Proficiency Guidelines, approved by the State Board of Education (February 2, 1995), establish Level III [on the end-of-grade tests] as the standard for each grade level, require LEAs to use existing funds to provide focused intervention to students who are not at Level III, provide for local decision-making regarding promotion decisions provided that test performance is taken into account, and provide for monitoring progress toward attaining these goals for students.

Related Testing Materials

The following materials have been developed to be used in conjunction with the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics.

Grade 3 Practice Test This practice test is a four-page test designed to familiarize grade 3 students with the process of taking a multiple-choice test and with recording responses on an answer grid.

NCDPI Item Bank The North Carolina Item Bank for reading and mathematics (edition A) was developed as a resource for teachers and administrators to operationalize and implement the revised North Carolina *Standard Course of Study*. The NCDPI Item Bank contains items (a) similar in format to items used by the NCDPI for the End-of-Grade Testing Program (multiple choice and open-ended) and (b) matched to the Mathematics and English Language Arts curricula. The NCDPI Item Bank can be used within an Instructional Management System (IMS).

The NCDPI Item Bank items were written by teachers, curriculum specialists, and professional item writers. The reading passages were selected from across the content areas and some longer passages were selected because more time could be spent on each passage within the classroom setting. Each item was edited, reviewed, and field tested during the Spring of 1993. Each item was administered to approximately 500 students randomly selected from across the state.

Testlets The End-of-Grade Testlets for Reading and Mathematics (Edition A) consist of mini-tests containing both multiple-choice and open-ended items designed to assist teachers in formative assessment (the ongoing assessment of students strengths and weaknesses). In mathematics, each testlet is related to a specific goal and objective of the North Carolina *Standard Course of Study* for a specific grade. In reading, each testlet is based on difficulty level (easy, medium, or hard items) and contains two to three passages (literary, content-based, and consumer/human interest). The testlets for Edition A were developed from items in the North Carolina Item Bank and were distributed in 1994.

Linking Curriculum, Instruction, and Assessment Series The Linking Series was developed to support classroom activities which reinforce the North Carolina *Teacher Handbook* for reading and mathematics and are compatible with the methods and skills being assessed on the North Carolina End-of-Grade Tests.

For reading, three topical units have been developed. Each topical unit explores a theme through a series of reading selections and classroom activities (grades 3-4: Pets, grades 5-6: Relationships Between Animals and People, and grades 7-8: Time). The reading selections mirror the End-of-Grade testing program by including literary, content-based and consumer/human interest passages. A separate student booklet contains just the reading selections for student use.

For mathematics, each Linking document explores a strand of the curriculum (i.e., measurement—goal 4 or problem-solving—goal 5) from grade 3 through grade 8. Each document summarizes the skills learned in the grade just before and just after the grade level being addressed, explains how the concepts relate to other curricular areas and other mathematical concepts, suggests classroom strategies and activities for the concepts in question, and offers procedures for assessing students' understanding of the concepts in a manner consistent with the instructional program.

Local Option Tests As part of the restructuring of the North Carolina public school in response to Senate Bill 16 in 1995, the number of state-mandated tests was cut in half to focus on the basics. The State Board of Education directed the NCDPI to develop procedures to support and facilitate the continued use of non-state-mandated tests and related services by local schools and systems. School systems have been given the option to purchase previously- or newly-developed NCDPI tests for local accountability. These assessments were designed to assess science and social studies (multiple-choice for grades 3 through 8); writing (for grades 3, 5, 6, and 8); and reading, mathematics, science, and social studies (open-ended for grades 3 through 8). Science and social studies are part of the North Carolina *Standard Course of Study* and must be taught.

The North Carolina End-of-Grade Tests were developed by the North Carolina Department of Public Instruction with technical support from the L.L. Thurstone Psychometric Laboratory at The University of North at Chapel Hill and the North Carolina Technical Advisory Group (see Appendix C for the members of the group). The tests were developed for use as achievement tests to measure the acquisition of specific subject-area content and skills associated with a particular grade in school. The purpose of these tests is twofold: (1) to improve student performance on the knowledge and skills specified in the North Carolina *Standard Course of Study*; and (2) to hold schools, school systems, and the state accountable for the education of students on the knowledge and skills specified in the North Carolina *Standard Course of Study*. Both norm-referenced (where the frame of reference is a specified population of students) and criterion-referenced (where the frame of reference is a specified content domain) interpretations of the test scores support the purpose of the North Carolina End-of-Grade Tests.

The end-of-grade tests are aligned with the revised North Carolina *Standard Course of Study* and emphasize higher level thinking skills—students are expected to have knowledge of important ideas and concepts; understand and interpret events; apply knowledge, skills, and concepts; and make connections. While knowledge of facts and concepts is important, the questions on the new tests are typically at a broader level and concern major ideas that students are expected to know to be considered literate. In addition to being asked to solve problems, students are asked "how" to solve a problem or "what strategy should be used" to solve a problem. Even in reading, students are asked to explain how they determined the correct answer to a given question. Better students are able to take responsibility for their own learning. They develop an awareness of their own thinking, including attitudes, habits, and dispositions.

The North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics are administered during the last three weeks of the school year in grades 3 through 8. While the tests are designed to assess

Table 2. Administrative information for the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics.

Subject/Grade		Amount of Testing Time	Number of Items on Each Form	Number of Items for Curricular Assessment
Reading Comprehension	3 Pretest	45	28	84
	3	100	56	168
	4	100	65	195
	5	100	65	195
	6	100	65	195
	7	100	66	198
	8	100	68	204
Mathematics	3 Pretest	7/40	5/35*	120
	3	12/85	12/68*	240
	4	12/85	12/68*	240
	5	12/85	12/68*	240
	6	12/85	12/68*	240
	7	12/85	8/72*	240
	8	12/85	8/72*	240

*Number of items on the computation part/number of items on the applications part.

reading comprehension and mathematics skills and knowledge, other content areas are integrated into the assessments—the reading comprehension test includes content-based passages and the reading and interpretation of graphs and charts; the mathematics test incorporates science and social studies data and experiments as sources of data for several of the strands of the curriculum. Beginning in 1996, a reading comprehension and mathematics pretest will be administered to all grade 3 students within the first three weeks of the school year.

All of the tests are scanned and scored within the local school system. Individual "Student/Parent" reports and school and school system summary reports are printed at the local level for accountability (*Report Card, Administrative Supplement*, and School-Building Improvement Reports).

For security and curricular purposes, multiple test forms are administered in each classroom. The measurement of student achievement is attained by administering different sets of items (equivalent in difficulty) to all students. The assessment of curriculum is met by the sum of the items administered across the three forms. All forms are administered in each classroom, one form per student. The measurement afforded by the three forms of items is critical to assessing curriculum mastery at the classroom, school, school system, and state levels.

Reading Comprehension

The North Carolina End-of-Grade Test of Reading Comprehension assesses a student's ability to read and comprehend written material that is appropriate for the grade level in terms of difficulty and content. The tests assess a student's ability to use strategies which enhance reading comprehension including acquiring, interpreting, and applying information, and reading for critical analysis and evaluation. Each test form consists of ten passages and from 3 to 8 associated questions per passage.

The reading passages on the tests are chosen to reflect the variety of reading done by students in and out of the classroom. The passages tend to be longer and more complete (compared to those typically found on standardized achievement tests) and have a high interest level for students at the particular grade level. On each test form there are four literary passages (for example, narrative, fiction, drama, and poetry), four content-based passages (science, social studies, art, health, and mathematics), and two consumer/human interest passages (instructions for performing a task, short information pieces). The variety of passages on each form allows for the assessment of reading for various purposes: for literary experience, to gain information, and to perform a task.

The grade 3 pretest mirrors the grade 3 end-of-grade reading test. Each student reads five passages appropriate for grade 2 representing the three types of passages (literary, content-based, and consumer/human interest). Then each student answers 28 multiple choice items assessing goals 2 and 3 of the North Carolina *Standard Course of Study* for English Language Arts. The scores are reported on the same developmental scale as the end-of-grade reading test scores are reported. This pretest will be administered statewide for the first time at the beginning of the 1996-97 school year.

Examples of the three types of passages and associated items can be found in Appendix A.

Mathematics

The North Carolina End-of-Grade Test of Mathematics consists of two parts: mathematics computation and mathematics applications. At the student level, the two parts of the test are combined to produce one mathematics score.

The mathematics computation part assesses a student's ability to do routine computations without a calculator. In grades 3 through 7, these items are symbolic computation skills that should be mastered during the grade level. In grade 8, these items include symbolic computation skills and application skills such as estimation.

The mathematics applications part assesses a student's ability to apply mathematical principles, solve problems, and explain mathematical processes. Problems are typically posed as real situations that students at the grade level may have encountered. Students are allowed to use calculators, rulers, and protractors on this part of the test. Due to the greater proportion of application items (compared to computation items), these tests tend to require more reading than found on typical multiple-choice tests of mathematics.

The grade 3 pretest of mathematics assesses the grade 2 mathematics *Standard Course of Study*. Each student answers 5 computation items and 35 applications items. Students are allowed to use calculators and rulers on the applications part of the test. The scores are reported on the same developmental scale as the end-of-grade mathematics test scores are reported. This pretest will be administered statewide for the first time at the beginning of the 1996-97 school year.

Examples of typical items can be found in Appendix A.

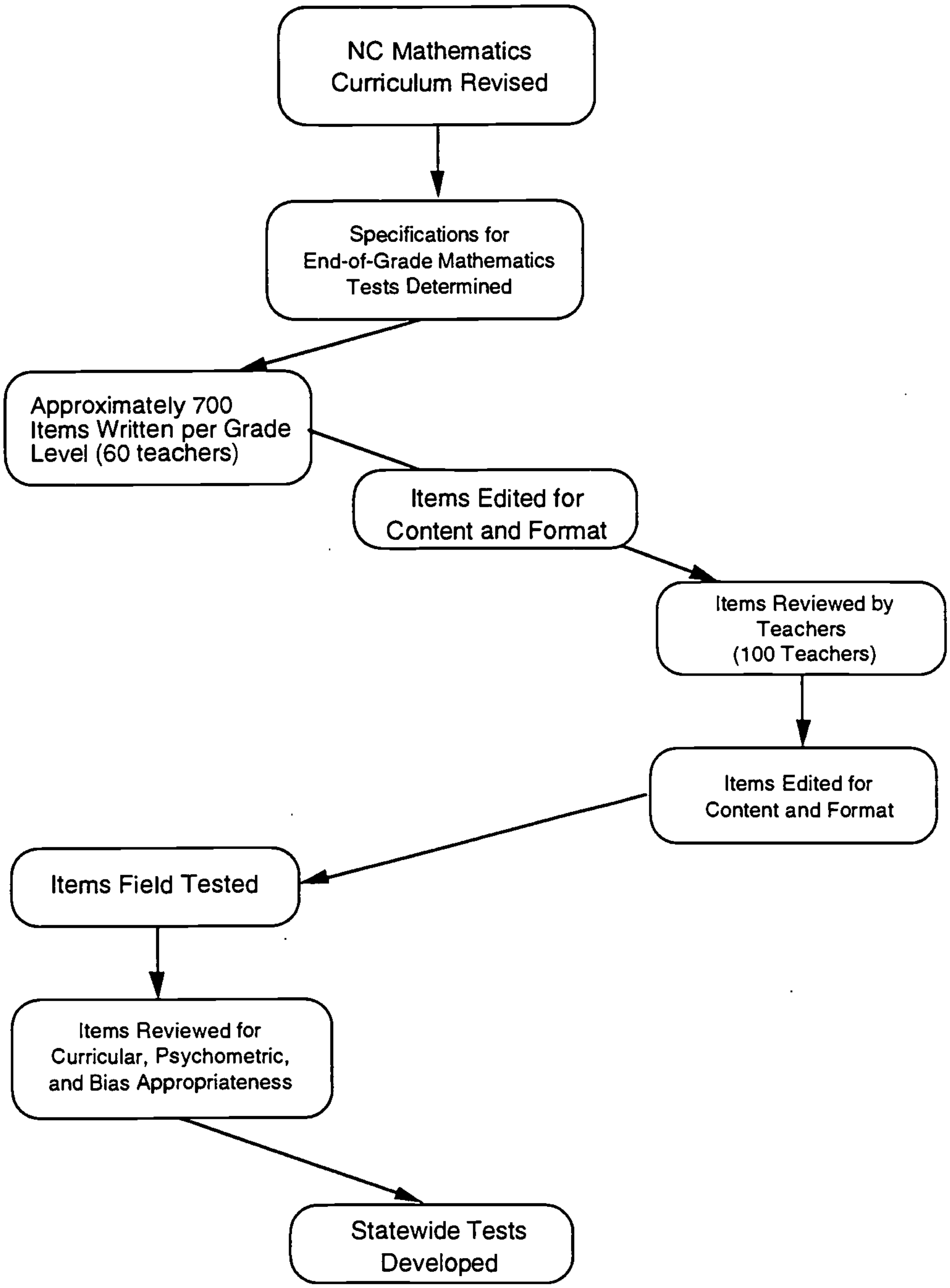


Figure 1. Test development process for the North Carolina End-of-Grade Tests.

Item development for the North Carolina End-of-Grade Tests goes through several unique stages (see flow chart to left)—content and test specification, item writing and review, field testing and analyses, and final evaluation. Each of these stages will be described in detail in the following sections. The item development phase began in the summer of 1990 during meetings with national and state curriculum specialists and continued through the fall of 1992 when the field test results were analyzed and evaluated.

Content Development

If a test is to be used to measure the degree to which a course of study has been mastered, the first step is to define the curriculum. The curricula assessed by the North Carolina End-of-Grade Tests were developed by the NCDPI/Division of Program Services, involving curriculum specialists, teachers, administrators, university professors, and others. These curricula reflect national content standards for student performance and the educational requirements in other industrialized nations. Content validity—the degree to which test items reflect the basic instructional program—is a quality commonly referenced in evaluating achievement tests. Content validity is built into a test from the beginning, and the procedures related to the content validation of the North Carolina End-of-Grade Tests are described below.

The North Carolina Mathematics curriculum is closely aligned with the National Council of Teachers of Mathematics *Curriculum and Evaluation Standards for School Mathematics* (1989) guidelines for teaching mathematics in kindergarten through grade 8. These standards call for an increase in the emphasis on the following: (1) building an understanding of numbers; (2) the meaning of the four operations; (3) the variety of ways to compute and to make estimates; (4) geometry, measurement, probability, statistics, and algebra; and (5) appropriate, individual segments of the curriculum at each grade level. The guidelines also call for a decrease in the emphasis on paper-and-pencil computation and repetition from year to year. The *Teacher Handbook for Mathematics* (NCDPI, 1992, page 3) states that

> the mathematics program proposed here is by necessity broader and more inclusive than in the past. It must develop more than vocabulary, facts, and principles; more than the ability to analyze a problem situation; more than an understanding of the logical structure of mathematics. The mathematics program must provide students with the knowledge which will enable them to distinguish fact from opinion, relevant from irrelevant data, and experimental results from proven theorems. . . . It must develop reading skills, motivation, and study habits essential for independent learning of mathematics. It must develop in students the appreciation for the connections between various branches of mathematics and how mathematics is connected to other disciplines.

The North Carolina reading curriculum is based on national trends in reading instruction, such as the International Reading Association, that see reading as the process of constructing meaning through the dynamic interaction among the reader's existing knowledge, the information suggested by the written language, and the context of the reading situation (Michigan Curriculum Review Committee/Michigan Reading Association, 1985). The *Teacher Handbook for English Language Arts* (NCDPI, 1992) states that "an effective communication skills program must be concerned with both process and content—with how students learn and what they learn" (page 5) and that "communication skills can be developed through listening to good literature because it exposes students to models of exemplary uses of language, helps develop an awareness of the power and beauty of languages, and can provide models for good writing" (page 22). Reading for information allows the reader "to make personal responses and judgements about the information as part of the reading process" and students "will become more productive if [they] are taught specific strategies for collecting data or ideas, recognizing relationships, and applying information" (page 23).

Appendix B contains the specific goals and objectives approved by the State Board of Education for mathematics (curriculum adopted in 1989) and reading (English Language Arts curriculum adopted in 1992) as the basis for instruction in grades 3 through 8.

Test Specifications

The content validity of the item pools was defined through a number of operations. First, the specifications for the reading and mathematics item pools were defined during the fall of 1990 and the spring of 1991. Working with groups of educators—NCDPI curriculum specialists, teachers, administrators, university professors, NCDPI testing consultants, the North Carolina Testing Commission, and others—test specifications were established for each of the content areas and grade levels assessed. The definition and refinement of the content specifications for the tests were continual processes.

Achievement test items can be classified along several dimensions. Two dimensions used to classify items for the end-of-grade tests are *difficulty level* and *thinking skill level.*

Difficulty level describes how hard the item is. Easy items are ones that about 70% of the examinees would answer correctly. Medium items are ones that about 50% to 60% of the students would answer correctly. Finally, hard items are ones that only about 20% or 30% of the students would answer correctly.

The other classification dimension, thinking skill level, describes the cognitive skills that a student must employ to solve the problem. One item may ask a student to classify several passages based on their genre (thinking skill: organizing); another question may ask the students to select the best procedure to use for solving a problem (thinking skill: evaluating).

In order to classify items by the thinking skill required, a framework to describe thinking skills must be used. The thinking skills framework used with the end-of-grade tests is from *Dimensions of Thinking* by Robert J. Marzano and others (1988). Many similar frameworks exist (for instance, that of Bloom), but *Dimensions of Thinking* was adopted by the North Carolina Department of Public Instruction in framing the revised *Standard Course of Study*. *Dimensions of Thinking* was developed through a collaborative process involving leading national experts in "thinking skills." The framework reflects current thinking in cognitive psychology, education, and philosophy. It provides a practical framework for curriculum development, instruction, assessment, and staff development.

A visual representation of the framework and a brief description of each of the dimensions of thinking are presented on the following pages. The framework should be a useful reference for curriculum development, instructional design, and in-service training.

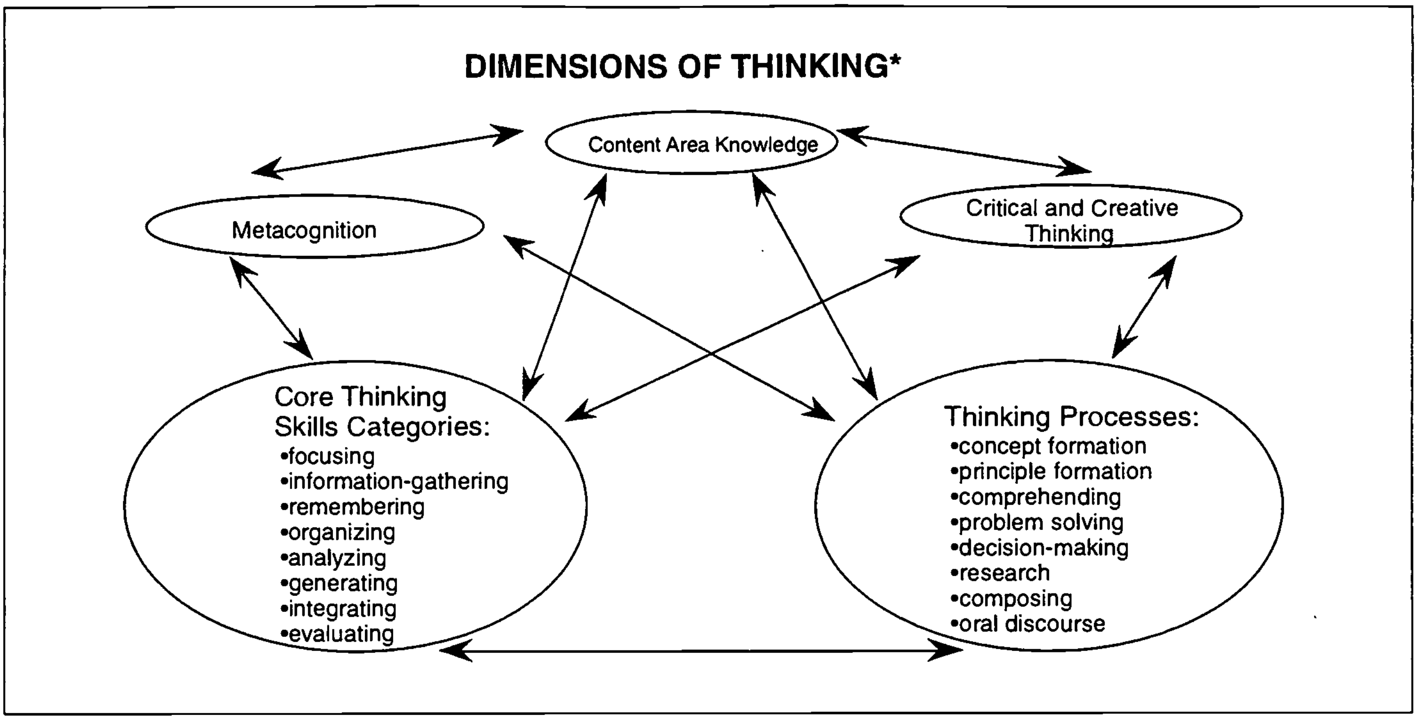


Figure 2. Thinking skills framework used with the North Carolina End-of-Grade Tests (*adapted from Robert Marzano et al., *Dimensions of Thinking*, 1988).

Metacognition Metacognition refers to awareness and control of one's thinking, including commitment, attitudes, and attention.

Critical and Creative Thinking The terms "critical" and "creative" thinking are ways of describing how we go about thinking. The two are not opposite ends of a single continuum—rather, they are complementary.

1. *Critical thinking* is "reasonable, reflective, thinking that is focused on deciding what to believe or do." Critical thinkers try to be aware of their own biases, and try to be objective and logical.
2. *Creative thinking* is "the ability to form new combinations of ideas to fulfill a need" or to get "original and otherwise appropriate results by the criteria of the domain in question."

Thinking Processes A thinking process is a relatively complex sequence of thinking skills.

1. *Concept formation*: organizing information about an entity and associating that information with a label (word).
2. *Principle formation*: recognizing relationships between or among concepts.
3. *Comprehending*: generating meaning or understanding by relating new information to prior knowledge.
4. *Problem solving*: analyzing and resolving a perplexing or difficult situation.
5. *Decision-making*: selecting from alternatives.
6. *Research*: conducting scientific inquiry.
7. *Composing*: developing a product which may be written, musical, mechanical, or artistic.
8. *Oral discourse*: talking with other people.

Core Thinking Skills A thinking skill is a relatively specific cognitive operation that can be considered a "building block" of thinking. Items are classified by the following skills because they: (1) have a sound basis in research and theoretical literature, (2) are important for students to be able to do, and (3) can be taught and reinforced in school.

Knowledge (1)

Focusing Skills—attending to selected pieces of information and ignoring others.

1. *Defining problems*: clarifying needs, discrepancies, or puzzling situations.
2. *Setting goals*: establishing direction and purpose.

Information-Gathering Skills—bringing to consciousness the relevant data needed.

3. *Observing*: obtaining information through one or more senses.
4. *Formulating questions*: seeking new information through inquiry.

Remembering Skills—storing and retrieving information.

5. *Encoding*: storing information in long-term memory.
6. *Recalling*: retrieving information from long-term memory.

Organizing (4)—arranging information so it can be used effectively.

7. *Comparing*: noting similarities and differences between or among entities.
8. *Classifying*: grouping and labeling entities on the basis of their attributes.
9. *Ordering*: sequencing entities according to a given criteria.
10. *Representing*: changing the form but not the substance of information.

Applying (5)—demonstrating prior knowledge within a new situation. The task is to bring together the appropriate information, generalizations or principles that are required to solve a problem.

Analyzing (6)—clarifying existing information by examining parts and relationships.

11. *Identifying attributes and components*: determining characteristics or parts of something.
12. *Identifying relationships and patterns*: recognizing ways in which elements are related.
13. *Identifying main idea*: identifying the central element; for example, the hierarchy of key ideas in a message or line of reasoning.
14. *Identifying errors*: recognizing logical fallacies and other mistakes and, where possible, correcting them.

Generating (7)—producing new information, meaning, or ideas.

15. *Inferring*: going beyond available information to identify what reasonably may be true.
16. *Predicting*: anticipating next events, or the outcome of a situation.
17. *Elaborating*: explaining by adding details, examples, or other relevant information.

Integrating (8)—connecting and combining information.

18. *Summarizing*: combining information efficiently into a cohesive statement.
19. *Restructuring*: changing existing knowledge structures to incorporate new information.

Evaluating (9)—assessing the reasonableness and quality of ideas.

20. *Establishing criteria*: setting standards for making judgements.
21. *Verifying*: confirming the accuracy of claims.

Mathematics The items for the mathematics item pools were specified by goal and objective, difficulty level, and thinking skill. Table 3 shows the content specifications for the mathematics test (both the computation and applications parts) by curricular strand (goal) and Table 4 shows the additional specifications for the computation part of the mathematics test at each grade level. Within goals, objectives were not weighted equally in the test specifications. Each objective was examined by the NCDPI mathematics curriculum specialists and weighted appropriately (see Appendix B for the weighting of individual objectives).

For difficulty level, 25% of the items were specified to be written at the easy level, 50% of the items were specified to be written at the medium level, and 25% of the items were specified to be written at the difficult level. The thinking skill level for each item was associated with the content objective. For example, the example item in Appendix D was written for objective 5.6, "Use proportional reasoning to solve problems," and the thinking skill was specified as "Applying." With the greater emphasis on solving problems "in context" and using "real-world" applications, the test requires more reading. While the vocabulary specific to mathematics content is used (e.g., "congruent"), every attempt has been made to have the non-content vocabulary below grade level.

Table 3. Item pool specifications for the North Carolina End-of-Grade Tests of Mathematics—Percent of items on the test assessing each strand of the curriculum.

Goal/Strand	3	4	5	6	7	8
7 Computation—Symbolic	15%	15%	15%	15%	10%	10%
1 Numeration	10%	15%	15%	11%	10%	11%
2 Geometry	10%	9%	12%	11%	10%	10%
3 Patterns/Pre-Algebra	10%	9%	10%	10%	15%	15%
4 Measurement	15%	15%	10%	10%	12%	10%
5 Problem Solving	15%	15%	15%	15%	18%	15%
6 Statistics	10%	9%	10%	15%	10%	12%
7 Computation—Context	15%	14%	12%	12%	15%	16%

Table 4. Computation skills assessed at each grade level on the North Carolina End-of-Grade Test of Mathematics.

Grade	Percent of Computation Items	Skill 1	Skill 2	Skill 3
3	15%	Add Whole Numbers	Subtract Whole Numbers	Multiply Whole Numbers
4	15%	Add/Subtract Whole Numbers	Multiply Whole Numbers	Divide Whole Numbers
5	15%	Add/Subtract Decimal Numbers	Multiply/Divide Whole Numbers	Add/Subtract Fractions (like denominators)
6	15%	Multiple/Divide Decimal Numbers	Add/Subtract Fractions (unlike denominators)	Multiply/Divide Fractions (unlike denominators)
7	10%	Add/Subtract/ Multiple/Divide Integers	Solve Ratios, Proportions, and Percents	

For grade 8, the skills assessed on the computation part of the test (10% of the whole test) include those skills that students should be able to do without the use of a calculator:

- computation within a context with decimals and percents
- computational estimation with fractions and decimals
- estimation within a context
- order of operations

In general, the "simpler" skills at one grade level are reduced and then dropped from measurement as more advanced ones become the focus of the grade level.

On the applications part of the test students are allowed to use calculators. This part assesses a student's ability to solve problems rather than apply specific procedures. Students in grades 3 through 5 are expected to have at least a simple 4-function calculator with a memory key (correct order of operations feature is desired) for use during instruction and on the test. Students in grades 6 through 8 are expected to have at least a 4-function calculator with a square-root function or algebraic logic, a fraction calculator, or a scientific calculator (not a graphing calculator) for use during instruction and on the test. In addition, on this part of the test students are expected to actually do measurement and are provided with inch/cm rulers with a leading edge (grades 3 through 8) and protractors (grades 5 through 8) to use as needed. Students are not directed to use the tools (calculator, ruler, and protractor) to solve a problem; the students must decide if and when to use the tools.

Reading Comprehension The item pools are only composed of items relating to a passage; there is no separate vocabulary section (vocabulary is assessed in context).

For all grades, the three types of passages were specified as follows: 40% literary (poetry, fiction, biographies, plays, essays), 40% content-based (science, social studies, art, health, and mathematics), and 20% consumer/human interest (recipes, directions, forms, projects, brochures, and short informational pieces relevant to the students). The passages were selected based on several criteria: they must be interesting to read, be complete (with a beginning, middle, and end), and be from sources students might actually read. By adhering to these criteria in the selection of passages, the passages tend to be longer than those typically found on tests.

The items for the reading item pools were specified to be appropriate to the passage so that the relevant aspects of the passage were assessed; specific goals and objectives were not specified in advance for each passage according to prepared test specifications. The test specifications stated that the items should reflect how a passage would actually be used in real life, e.g. no higher-level thinking skills about advertisements found in the yellow pages of the telephone book. In addition, most passages should have items spanning the goals and objectives of the English Language Arts curriculum. Goal 1, metacognitive strategies, is not assessed in grades 3 and 4. It was felt that students in these grades, while exhibiting reading strategies as they read, would not be able to explain the strategies they used. Goal 4, personal response, is not assessed by the reading comprehension multiple choice test. This goal is better assessed in an open-ended format.

Item Writing and Review

Selection of Reading Passages Reading passages were selected during the winter and spring of 1991 by NCDPI Instructional and Testing consultants and other curriculum specialists. For each grade level (3 through 8) 100 passages were selected: 40 literary, 40 content-based (art, science, health, mathematics, and social studies), and 20 consumer/human interest. The selected passages were ones that would generally be read by students, would be interesting to students, and were appropriate content for a reading comprehension test.

For each passage, a frame was written as an introduction to the passage. Each frame was designed to stir interest and support comprehension, while at the same time not reveal something in the passage that the reader should discover for himself or herself.

The readability of the passages was determined by the Fry index and the Degrees of Reading Power (DRP) index. Literary passages were assigned to a grade level based on content and the Fry index (generally on grade level). Content-based passages were assigned to a grade level based on the content match with the North Carolina *Standard Course of Study* and the Fry index (generally one grade level above). Consumer/human interest passages were assigned to a grade level based on content.

Selection and Training of Item Writers Item writers were nominated by the NCDPI curriculum specialists from across the state because of their knowledge of the curriculum and exemplary teaching status. Twelve teachers and curriculum specialists at each grade (for a total of 72) were trained in the technical aspects of mathematics item writing, and 12 teachers and curriculum specialists at each grade (for a total of 72) were trained to write reading items. Each item writer was sent a training packet specially designed for the grade and subject matter for ehich they were to write items. The materials consisted of a videotape (and a script) of *How to Write Multiple Choice Achievement Test Items* developed by the NCDPI, three packets of work materials related to the specific content area (i.e., grade 6 mathematics), and a copy of *Guidelines for Bias-Free Publishing* developed by CTB/McGraw-Hill. As described by Haladyna (1994), the "best-answer" format of multiple choice questions was used in these tests. This format is well suited for testing a student's ability to evaluate (Marzano's highest thinking skill level).

Each item writer was asked to submit 10 items as part of the training phase. These 10 items were evaluated and, based on the evaluations by content specialists, the item writers were asked to develop additional items. Item writers received feedback on the 10 items they had developed to help them better develop the additional items requested. Since the reading curriculum was new, item writers were brought to two central locations for further training on the curriculum and item development (March 1991).

The use of classroom teachers from across the state as item writers helped to insure that instructional validity was maintained because the background would be drawn from their classroom experience.

Item Writing and Review Each item writer was contracted to write 75 items during the winter of 1990 and spring of 1991. Mathematics item writers wrote 75 items from across the curriculum (all seven goals) and reading item writers wrote 75 relevant questions for 10 passages. For reading, the goals/objectives were not specified for the item writers in advance; instead, the item writers were asked to develop the relevant and important questions related to the passage, while addressing as many aspects of the curriculum as possible. The reading item writers were instructed to write items assessing goals 1, 2, and 3 of the curriculum, but primarily objectives 2.2 and 2.3. The writers were asked to write 8 to 10 items for each literary and content-based passage and 4-6 items for each consumer/human interest passage. All total, 7,383 mathematics items and 7,550 reading items were developed. See Appendix D for a completed Item Specification Form for a sample item.

Table 5. Number of items written and passages selected by grade level.

Grade	Items Written for Mathematics	Passages Selected for Reading	Items Written for Reading
3	1,167	135	954
4	1,257	150	1,137
5	1,269	135	1,311
6	1,264	159	1,353
7	1,218	133	1,377
8	1,208	171	1,418

Next, the item pools were analyzed by curriculum specialists and classroom teachers to ensure that the items were valid representations of the objectives for which they were written. For each subject area and grade level, 10 to 15 individuals met in June 1991 to review the items (and passages for reading). Item reviewers learned about and discussed the revised *Standard Course of Study* for the items they would be reviewing, the end-of-grade testing program, and the test development process in general. Each item reviewer received a copy of the document *How to Review Multiple Choice Achievement Test Items* (developed by NCDPI) which described the criteria to be used to evaluate each item. During review training, each of the criteria was discussed, example items were used to show how each of the criteria could be met, and example items were discussed that did not meet the criteria for inclusion in the item pools.

The criteria for evaluating each item included the following:

- conceptual: objective match, fair representation, lack of cultural bias, clear statement, single problem, one best answer, common context in foils, each foil credible;
- language: appropriate for age; correct punctuation, spelling, and grammar; lack of excess words; no stem/foil clues; no negative in foils;
- format: logical order of foils; familiar presentation style, print size, and type; correct mechanics and appearance; equal length foils; and
- diagram: necessary, clean, relevant, unbiased.

The evaluation of each reading passage and the associated questions also included the following:

- For what grade levels is the passage appropriate?
- For the grade level to which the passage is currently assigned, is it easy, medium, or hard?
- Is the passage interesting to read and does it have a beginning, middle, and end?
- Is the frame acceptable for the passage?
- Do all of the objectives fit well with the passage, or should one or more not be used and substituted with another objective? Please explain.
- Do the items adequately cover the major content of the passage? Are the most important ideas included? Please explain.

Each item was reviewed by at least four individuals. See Appendix D for a sample completed Item Review Form and the summary of the teacher comments.

During the summer of 1991, the results from the item reviews were aggregated by item (and passage). The results were then examined item-by-item by exemplary teachers, curriculum specialists, and test development staff. Based on the comments from the reviewers, items were revised and/or rewritten, item-objective matches were examined and changed where necessary, and frames and diagrams for reading passages were refined. Throughout the item writing and review process 1,123 mathematics items (112 to 258 per grade) and 2,923 reading items (458 to 511 per grade) were deleted from the item pools. The large number of deleted reading items resulted from the revision of the reading curriculum from a skills-based curriculum to a holistic reading curriculum; teachers had a much harder time writing quality items at the level required by the test specifications with the revised curriculum.

During the fall and winter of 1991, additional passages were found and additional items were developed where necessary. These items were reviewed by curriculum specialists and test development staff based on the same criteria as the first review.

Field Testing

During the winter of 1992, the items at each grade level were collected into 10 test forms for field testing (except for grades 4 and 5 mathematics where there were 11 forms of the tests due to an abundance of items). Although the forms were not the final forms of the North Carolina End-of-Grade Tests, they were organized according to the specifications for the final tests as shown in Table 3 for the mathematics test and the discussion on page 15 for the reading comprehension test. Each mathematics field test contained 80 items (8 or 12 symbolic computation items and 72 or 68 applications items). Each reading field test contained 10 passages; because more items were field tested for each reading passage than would eventually be used, in grades 7 and 8 there were only nine passages on each field test form. The number of reading items per form ranged from 58 to 61 in grade 3 to 69 to 73 items in grade 8.

Table 6. Number of items and passages field tested in May 1992 by grade level.

Grade	Mathematics Items Field Tested	Reading Passages Field Tested	Reading Items Field Tested
3	800	81	496
4	868	83	627
5	867	93	707
6	800	92	710
7	800	83	710
8	800	90	716

One of the goals of the development of the reading comprehension and mathematics End-of-Grade Tests was to follow the curriculum for each area as closely as possible. Both the Mathematics and the English Language Arts curricula are developmental in nature. In order to establish developmental scales that spanned grade 3 to grade 8 for the two tests, the typical amount of growth during a school year that a student could exhibit on the tests needed to be established. In mathematics, 2 forms at each grade level were also administered at the next higher grade level (for example, forms 1 and 2 of grade 4 were administered as forms 11 and 12 in grade 5). In reading, due to the unknown effect of using passages (i.e., testlets), all 10 forms for a grade level were also administered at the grade below and the grade above (for example, forms 1 through 10 of grade 4 were administered as forms 11 through 20 in grade 3 and forms 21 through 30 at grade 5).

Next, test administration instructions were written, distribution procedures were organized, and administrators were trained to conduct the field test administration. The test administration organization used to administer statewide tests in North Carolina was employed to do the field testing. The administration of the field test forms followed the routine eventually expected to be used when the statewide tests were administered.

Samples of students were selected to take the reading and mathematics field test forms in May 1992. To insure broad representation, schools were selected from across the state and were representative of the state based on ethnic/racial characteristics of the student population and geographic location. At least one grade in every school in the state was sampled for one of three field tests: multiple choice mathematics, multiple choice reading, or open-ended. Tables 7 and 8 show the characteristics of the field test samples at each grade level for mathematics and reading (number of students tested, number of schools sampled, and percent of students in the field test samples that were idendified as limited English proficient). While every effort was made to ensure full participation, only limited modifications for exceptional children were available for use during the field tests (e.g., dictation to a proctor/scribe, magnification devices, student marks in the test book, multiple test sessions, scheduled extended time, and testing in a separate room). Modifications not availiable during field testing included braille and large-print versions of the test books.

Table 7. Characteristics of the mathematics field test samples (May 1992) by grade level.

Grade	Number of Students Tested	Number of Schools Sampled	Percent of Students Tested Identified as LEP
3	10,525	157	1.1
4	12,202	184	1.5
5	11,665	183	0.7
6	10,926	104	1.9
7	11,136	89	2.2
8	11,026	83	1.6

Table 8. Characteristics of the reading field test samples (May 1992) by grade level.

Grade	Number of Students Tested	Number of Schools Sampled	Percent of Students Tested Identified as LEP
3	19,102	283	1.0
4	27,350	384	1.1
5	26,863	408	1.0
6	27,528	260	1.2
7	26,240	195	1.6
8	17,274	137	1.4

Item Analysis and Selection

The field test data for all items were analyzed by the NCDPI using both the classical measurement model and the three-parameter logistic item response theory model. Item statistics and descriptive information (item number, passage number, field test form and item number, curriculum objective, and answer key) were printed on labels and attached to the item record for each item. The item record contains the statistical, descriptive, and historical information for an item; a copy of the item itself as it was field tested; any comments by reviewers; and the psychometric notations. Each item has a separate item record. See Appendix D for the item record form of a sample item.

Classical Measurement Analyses For each item the p-value (percent correct), the standard deviation of the p-value, 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.

Item Response Theory Analyses Classical test theory has two basic shortcomings: (1) the use of item indices whose values depend on the particular group of examinees from which they were obtained, and (2) the use of examinee ability estimates that depend on the particular choice of items selected for a test. The basic premises of item response theory (IRT) overcome these shortcomings by predicting the performance of an examinee on a test item based on a set of underlying abilities. 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 three-parameter logistic model (3PL) 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 one-parameter model (Rasch) only takes into account the difficulty of the item. The 3PL model has three assumptions: (1) unidimensionality—only one ability is assessed by the set of items, (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 reflects the true relationship among the unobservable variable (ability) and the observable variable (item response). The equation for the three-parameter logistic model is

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad \text{where } i = 1, 2, ..., n \qquad \text{(Equation 1)}$$

$P_i(\theta)$—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—a scaling factor to make the logistic function as close as possible to the normal ogive function (equals 1.7)

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

The following figures show the item characteristic curves for a typical 4-option multiple-choice item and several items from the end-of-grade reading comprehension and mathematics item pools.

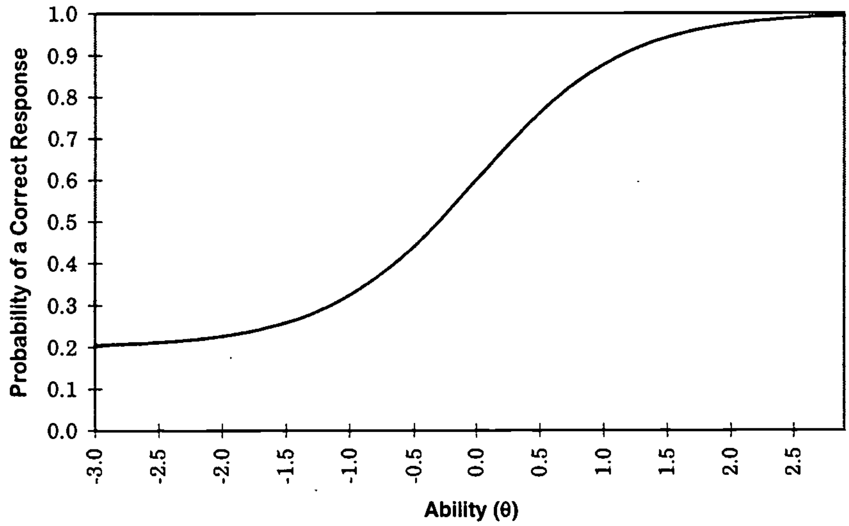


Figure 3. Item characteristic curve of a typical 4-option multiple-choice item (a = 1.00, b = 0.00, and c = 0.20).

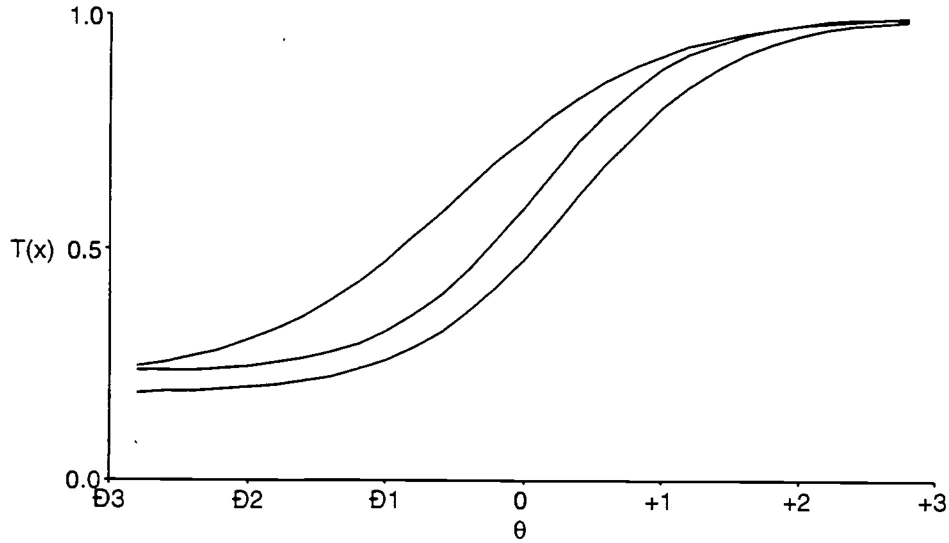


Figure 4. Item characteristic curve of reading item #500R2 (a = 1.096, b = 0.078, and c = 0.23). This item was field tested at multiple grades in order to vertically equate the tests.

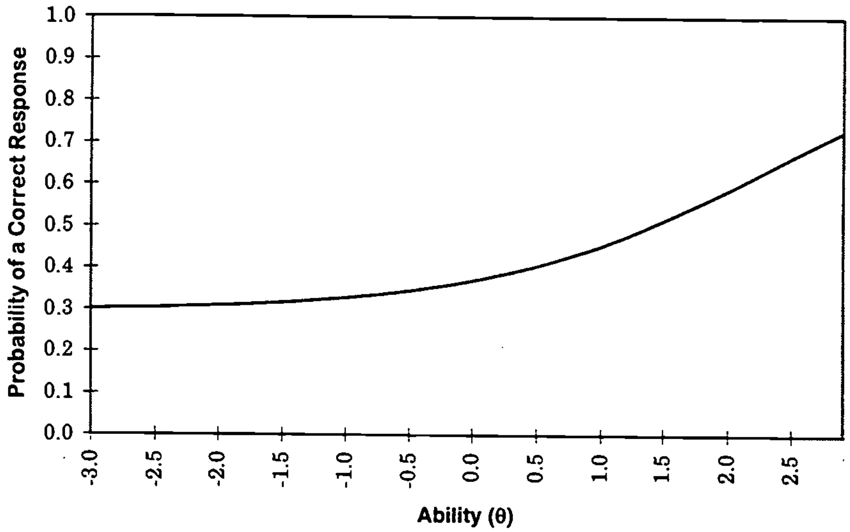


Figure 5. Item characteristic curve of mathematics item #80311 that exhibited a low slope ($a = 0.527$, $b = 2.387$, and $c = 0.295$). This item was flagged as exhibiting "Weak Prediction."

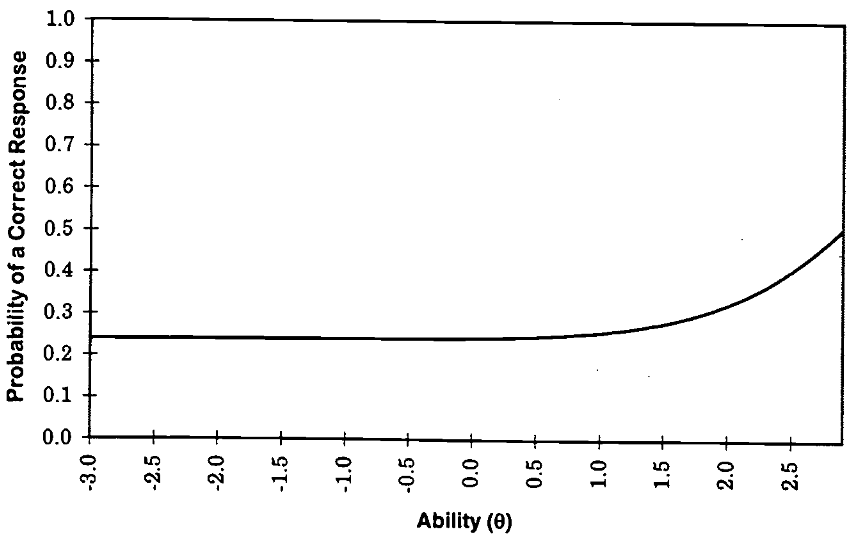


Figure 6. Item characteristic curve of mathematics item #6R1 that was difficult, but was retained for test development ($a = 0.95$, $b = 3.277$, and $c = 0.239$).

Bias Analyses 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 2×2 contingency tables, where j is the number of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the 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).

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

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

where r^2 is the Pearson correlation between the row variable and the column variable (SAS Institute, 1985).

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

$$\alpha_j = \frac{p_{Rj}/q_{Rj}}{p_{Fj}/q_{Fj}} = \frac{\Omega_{Rj}}{\Omega_{Fj}} \qquad \text{(Equation 3)}$$

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

For the end-of-grade tests, males (approximately 50.8% of the population) and blacks (approximately 29% of the population) were defined as the focal groups and females (approximately 49.2% of the population) and whites (approximately 65.9% of the population) were defined as the reference groups.

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).

Review Of Item Pools During the field test each test administrator was asked to review one field test form (the top one in the set of materials they received for administration). The teachers were asked to respond to two questions and to make any general comments concerning the item.

1. Instruction: Which of the following describes the concept or skill measured by this item?
   - **Basic** to instruction this year and taught to most students in this class. When coding an item as "basic," consider whether classroom instruction has been sufficient such that the item is a fair test question for most students in the class;
   - **Enrichment** material taught only to advanced students this year; or
   - Not considered part of the curriculum this year so **not taught**.

2. Item Quality: In your opinion, is this item appropriate for the end-of-grade test? In judging whether an item is appropriate, the reviewer should take into consideration conceptual quality, language quality, format and graphics quality, and cultural bias. For any item that the reviewer believes needs revision or is not appropriate, the space next to the bubble should be used to write a comment or suggestion for improvement.
   - **Yes**, if the item is okay;
   - **Revise**, if the item needs additional work; or
   - **No**, if the item should not be used in any form.

All items, statistics, and comments were reviewed by curriculum specialists and testing consultants, and items that were not deemed appropriate for curricular or psychometric reasons were deleted.

Items flagged for exhibiting ethnic and/or gender bias (Mantel-Haenszel indices greater than 1.5 or less than 0.67) were then reviewed by a group of individuals (not the developers) that represented various minority groups. The individuals on the bias review team were selected because of their minority group membership or their experience with exceptional students and because of their knowledge of the curriculum area. The members of the team were provided copies of the item records of items flagged as being biased (each item record included a copy of the item, the curricular objective the item was assessing, and the various item statistics from the field test). The team members were asked to review individually each item in terms of the following questions:

- Does the item contain any offensive gender, ethnic, and/or regional content?
- Does the item contain gender, ethnic, or cultural stereotyping?
- Does the item contain activities that will be more familiar to one group than another?
- Do the words in the item have a different meaning in one group than in another?
- Could there be group differences in performance that are *unrelated* to proficiency in the content area?

The team members were then instructed that if their answer was "yes" to any of the questions for a particular item, that they should record the 5-digit item number and check the appropriate column(s) on the Item Bias Review Sheet (see Appendix F for samples of the bias review materials). If an item was deemed to be biased, the team members were also asked to explain their decision. Items that were consistently identified as exhibiting bias were deleted from the item pool and were not used in the development of any tests. All of the biased items that were flagged, but not rejected by the bias review team, were examined by the curriculum specialists. If all students were expected to master the content of the item, then the item was retained for test development (e.g., the item 2×1 is biased in favor of females, the log odds ratio is 1.563).

Final average item pool parameter estimates for reading comprehension and mathematics are presented in Tables 9 and 10 respectively.

Table 9. Average item pool parameter estimates for the North Carolina End-of-Grade Test of Reading Comprehension by Grade.

Grade	IRT Parameters			P-value	Bias (Odds Ratio)	
	Threshold (*b*)	Slope (*a*)	Asymptote (c)		Ethnic/Race	Gender
3	0.096	1.220	0.209	0.592	1.026	1.039
4	0.143	1.192	0.213	0.583	1.028	1.037
5	0.186	1.170	0.219	0.577	1.023	1.041
6	0.202	1.150	0.225	0.577	1.032	1.039
7	0.242	1.161	0.222	0.570	1.031	1.045
8	0.139	1.168	0.227	0.595	1.041	1.053

Table 10. Average item pool parameter estimates for the North Carolina End-of-Grade Test of Mathematics by grade.

Grade	3PL IRT Parameters			P-value	Bias (Odds Ratio)	
	Threshold (*b*)	Slope (*a*)	Asymptote (c)		Ethnic/Race	Gender
3	-0.085	1.063	0.220	0.618	1.027	1.056
4	0.184	1.037	0.221	0.577	1.032	1.056
5	0.809	1.039	0.214	0.470	1.034	1.041
6	0.993	1.061	0.211	0.434	1.039	1.040
7	1.239	1.119	0.211	0.392	1.045	1.027
8	1.226	1.080	0.212	0.398	1.033	1.028

Dimensionality of Item Pools The dimensionality of the end-of-grade tests was examined by the L.L Thurstone Laboratory at the University of North Carolina at Chapel Hill. The analyses involved the application of weighted least squares analysis of the polychoric correlations computed among item parcels constructed from the binary items on the tests. This procedure was used to avoid the well-documented artifacts that may arise when binary data are factor-analyzed (Mislevy, 1986), and the less well-documented difficulties that arise with full-information item factor analysis (Bock, Gibbons, and Muraki, 1988).

For all of the mathematics end-of-grade field tests, item parcels (Cattell, 1956) were constructed of the items measuring each curricular goal that the items were developed to measure. For the reading end-of-grade field tests, item parcels were constructed of the items associated with each passage on the field test. In all cases, item-parcel scores were computed as the summed score (number correct) for each parcel. Polychoric correlations were computed then among the item-parcel scores using the computer software PRELIS (Joreskog & Sorbom, 1986), and the weighted least squares analysis was done with LISREL (Joreskog & Sorbom, 1988). When open-ended responses were analyzed with the multiple-choice item parcels, the Schmid and Leiman (1957) representation of heirarchical factor analysis was used to explore the relative sizes of the unique factors for the open-ended and multiple-choice items over and above the general factor for the test.

For mathematics, generally a single-common-factor model fit the data very well, as measured by the likelihood-ratio goodness of fit criterion. Additional analyses were performed to determine whether it would be wise to produce single score for combined performance on the multiple-choice items and the open-ended items. While loadings on a unique factor for the open-ended items was significantly different from zero, these loadings tended to be very small, generally of the order of 0.2, while loadings for the multiple choice item parcels and the open-ended items on the general factor tended to be around 0.7. It was concluded that, for mathematics, even including the open-ended items, the tests were very nearly unidimensional. In order to aid in the interpretation of test scores, two subscores were developed for the mathematics multiple choice tests—computation and applications. Tests were constructed to be equivalent at the total score level and at each of the subscore levels.

For the reading multiple-choice tests, generally a single-common-factor model fit the data very well, as measured by the likelihood-ratio goodness of fit criterion. When similar analyses concerning the inclusion of open-ended reading items were performed, in general, the open-ended items exhibited larger loadings on a factor unique to the open-ended items (around 0.5). Conversely, all of the multiple-choice item parcels (passage scores) had loadings on the general factor around 0.7. It was concluded that for reading, the open-ended items measured sufficiently different aspects of individual differences such that separate scores for the open-ended portion of the test were justified.

For each end-of-grade test, reading comprehension or mathematics, six forms were prepared for administration to students in each grade, 3 through 8. Three forms were developed during the Fall of 1992 for administration the first year (May 1993) and three forms were developed during the Fall of 1993 for administration in May 1994. In order to assure that each set of three forms for a subject and grade was equivalent (both within and between the sets), the item pools were randomly split in half and the second half of each item pool was put away and not used during the development of the first set of three forms.

Items for the mathematics tests and passages for the reading tests were selected using a modified domain sampling model, with the various forms equivalent. In the modification used here, the domain of items for each test was limited to those items that had satisfactory psychometric characteristics and curricular approval, and were approved by the bias review team. This was determined by the analyses of the item field-test data and the reviews by curriculum specialists and testing consultants. Some items that did not meet the psychometric criteria were used in test development because these were items that assessed new parts of the curricula (typically the items had low slopes because the material had not been completely taught to students).

The three forms of each subject and grade test were developed according to the test specifications delineated during the initial phase of development (see Table 3, the discussion on page 15, and Appendix B) and the average p-value for each test or subtest (mathematics computation and applications) was equivalent to the average p-value of the entire item pool. The item parameters for each form (threshold and slope) were examined to determine if they were approximately equivalent. If the average item parameters were not equivalent, then some items were replaced with other items from the item pools to insure equivalence.

Table 11. Average p-value for each part of the mathematics test and the reading test by grade level.

Grade	Math Computation Average P-value	Math Applications Average P-value	Reading Average P-value
3	0.83	0.600	0.562
4	0.80	0.535	0.570
5	0.65	0.440	0.560
6	0.51	0.423	0.564
7	0.46	0.380	0.560
8	0.45	0.390	0.585

After each test was assembled into forms (3 forms for each of six grades in reading and mathematics), the forms were reviewed by five to ten grade level and subject teachers and curriculum supervisors. Each group (separate for grade and subject) met in Raleigh for one day during the fall of 1992 and winter of 1993 (also during the fall of 1993 for the second set of forms) and worked independently of the test developers. The test review groups discussed the revised *Standard Course of Study* for the test they would be reviewing, the end-of-grade testing program, and the test development process in general. During review training, each of the criteria for evaluating the tests was discussed.

The criteria for evaluating each group of three forms included the following:

- that the content of the test forms should reflect the goals and objectives of the North Carolina *Standard Course of Study* for the subject and grade level (curricular validity);
- that the content of the test forms should reflect the goals and objectives taught in North Carolina schools (instructional validity);
- that the items should be clearly and concisely written, and the vocabulary appropriate to the target age level (item quality);
- that the 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
- that each item should have one and only one answer that is right; however, the distractors should appear plausible for someone who has not achieved mastery of the representative objective (one best answer).

Each of the criteria was evaluated on the following scale: to a superior degree (4), to a high degree (3), to an average degree (2), to a low degree (1), and not at all (0). Reviewers were also given space to make additional comments related to each of the five criteria and any general comments.

Reviewers worked as a group to review the three forms of each test. They were instructed to first actually take the tests (circling the correct response in the booklet) and 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 was allowed to discuss the tests and make comments as a group. The ratings of the tests were completed anonymously. The ratings and the comments were aggregated for review by NCDPI curriculum specialists and testing consultants. The ratings for the tests are shown in Tables 12 and 13.

Table 12. Average test review ratings for the North Carolina End-of-Grade Test of Reading Comprehension by grade.

Grade	Curricular Validity (Q #1)	Instructional Validity (Q #2)	Item Quality (Q #3)	Test/Item Bias (Q #4)	One Best Answer (Q #5)
3	3.3	2.0	2.6	3.0	2.5
4	3.0	2.4	1.9	2.3	2.3
5	3.2	3.0	2.1	3.2	2.4
6	2.9	2.5	2.5	2.6	2.3
7	2.8	2.0	2.1	1.9	2.2
8	2.7	2.1	2.1	3.0	2.3

Table 13. Average test review ratings for the North Carolina End-of-Grade Test of Mathematics by grade.

Grade	Curricular Validity (Q #1)	Instructional Validity (Q #2)	Item Quality (Q #3)	Test/Item Bias (Q #4)	One Best Answer (Q #5)
3	3.3	2.0	3.3	3.5	3.0
4	3.2	2.3	2.9	2.8	3.0
5	3.4	2.4	2.7	3.0	2.7
6	3.3	2.5	2.4	3.1	3.0
7	3.0	1.8	2.5	2.5	3.0
8	2.8	1.7	2.5	3.3	2.9

Developmental Scales

One of the main goals of the development of the North Carolina End-of-Grade Tests was to match the content at each grade level *and* the overall philosophy of the English Language Arts and mathematics curricula. The philosophy of each area is developmental in nature—the skills and knowledge needed at one grade level are built on the skills and knowledge acquired at the previous grade level. Therefore, a developmental scale was constructed to measure growth in skills and knowledge throughout the grades.

The first step in the process of developing these scales was to determine the typical amount of growth that occurs during a school year in reading comprehension and mathematics. The field test procedure for administering the same forms at multiple grades (linking forms) was described on page 18.

The next step in the development process was to analyze the linking forms and determine the differences in the distributions across the 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, the three-parameter logistic model was used at both grades the linking form was administered (the grade the items were developed for and the grade higher for mathematics and the grade higher and lower for reading). Figure 7 below graphically models this procedure—item characteristic curves are developed for each item based on the IRT parameters and then the individual curves are aggregated across the test form to develop the test characteristic curve.

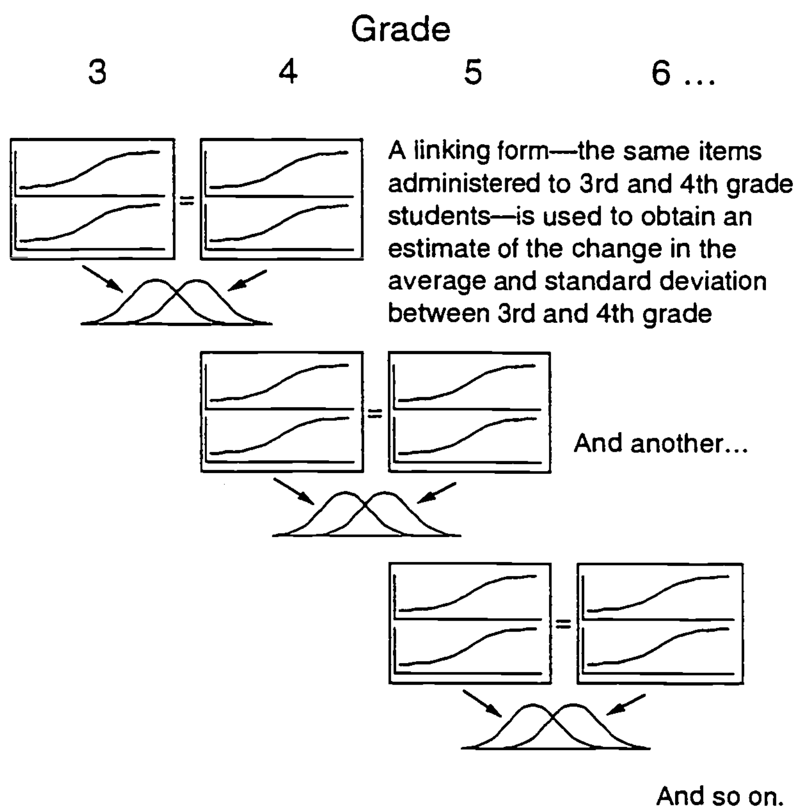


Figure 7. Graphical model of the examiniation of linking forms to determine changes in achievement over one year.

The next steps in the process were conducted at the L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill and reported in an unpublished manuscript by Williams, Pommerich, and Thissen (1996). First, 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 changes in 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, θ_G~N(0,1), and the mean and the standard deviation of the upper grade was estimated, θ_U~N(μ_U,σ_U). Bimain equates only the slopes (*a*s) and asymptotes (*c*s) for the two groups, and then estimates separate thresholds (*b*s) for each using a common population distribution; then with μ_G set to 0 and σ_G set to 1.0, it adjusts μ_U and σ_U based upon the estimated differences in the difficulties of the items. Figure 8 shows how the distributions for the linked forms of the mathematics test compare across the grade levels.

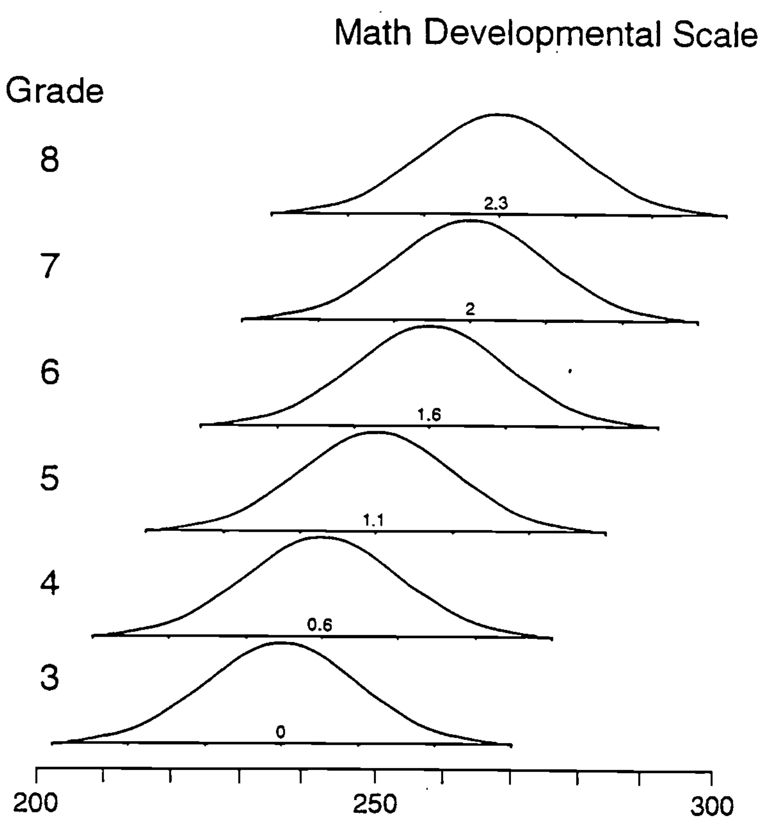


Figure 8. Graphical presentation of the changes in the test difficulties across the grades based on the grade 3 mean of 0 and standard deviation of 1.

A similar procedure was used with the reading tests. Instead of only estimating the upper grade's population distribution in relation to the grade's distribution, the lower grade's distribution was also estimated, $\theta_L \sim N(\mu_L, \sigma_L)$.

Tables 14 and 15 show the linking of forms for reading and mathematics across the six grades, the scale of the latent proficiency for reading and mathematics and the operational scales for reading and mathematics. The growth between grades was approximately one-half standard deviation, therefore, a developmental scale was appropriate for the reading and mathematics test scores.

Table 14. Scaling results for the North Carolina End-of-Grade Test of Reading Comprehension.

Grade	Scaling Results from Linking Forms—μ_U (σ_U)	Latent Proficiency Scaling Mean (SD)	Operational Scaling Mean (SD for forms)
3	0 (1.0)	141.125 (10.744)	141.125 (10.22–10.23)
4	0.43 (0.97)	145.883 (10.471)	145.883 (10.00–10.03)
5	0.83 (0.93)	150.000 (10.000)	150.000 (9.49–9.54)
6	1.06 (0.94)	152.364 (10.047)	152.364 (9.53–9.58)
7	1.34 (0.91)	155.031 (9.758)	155.031 (9.29–9.31)
8	1.53 (0.92)	157.013 (9.908)	157.013 (9.45–9.47)

Table 15. Scaling results for the North Carolina End-of-Grade Test of Mathematics.

Grade	Scaling Results from Linking Forms—μ_U (σ_U)	Latent Proficiency Scaling Mean (SD)	Operational Scaling Mean (SD for forms)
3	0 (1.0)	137.347 (11.785)	137.347 (11.36–11.37)
4	0.53 (0.91)	144.169 (10.697)	144.169 (10.24–10.26)
5	1.07 (0.85)	150.000 (10.000)	150.000 (9.33–9.38)
6	1.65 (0.87)	155.878 (10.248)	155.878 (9.43–9.46)
7	2.08 (0.92)	160.851 (10.791)	160.851 (9.60–9.74)
8	2.36 (0.93)	164.120 (10.950)	164.120 (9.90–9.96)

Tables 14 and 15 show the final mean scores and standard deviations (operational scaling results) used with the End-of-Grade Testing Program. The decision was made that the developmental scale scores for reading comprehension and mathematics should range from 100 to 200 such that grade 5 for both subjects would have a mean of 150 and a standard deviation of 10. With a standard deviation of 10, one point on the scale corresponds to 0.1 standard deviations—typical of most standardized achievement tests.

The scales for reading comprehension and mathematics, while both ranging from 100 to 200, are not directly comparable. From the results in Tables 14 and 15, one can see that reading and mathematics do not grow at the same rate (scaling results from linking forms). For reading comprehension, the average proficiency of the population changes from 0.43 in grade 4 to 1.53 in grade 8; for mathematics, the average proficiency of the population changes from 0.53 in grade 4 to 2.36 in grade 8.

Scores

Developmental Scale 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. Because the sum of the number of items answered correctly is easy to understand and interpret, even though more sophisticated scoring options based on the IRT item parameters could be used (such as pattern scoring), it was felt that this was the best way for the results of the North Carolina End-of-Grade Tests to be reported.

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 IRT parameters for each item. The scale scores produced by this program give essentially the same results as the scale scores produced by pattern scoring (easier items answered correctly count less than harder items answered correctly). 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 for each subject area. For example, at grade 3 there are three mathematics forms and 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 summed raw score. See Appendix E for further information concerning the methodology used to convert summed scores to scaled scores.

Figures 9 and 10 show the progression across grades of the reading comprehension and mathematics developmental scales. The shaded graphics show 95% of the scores for each grade, with the mean represented by the white line. The notches in the graphs are one standard deviation from the mean, and the tails of each graph end two standard deviations from the mean. The scales allow the performance of individual students, groups of students, schools, school systems, and the state to be compared across grades.

Percentile Ranks In addition to scale scores, the percentile ranks associated with each scale score within each grade and subject are also reported at the individual level. The percentile rank for each scale score is the percentage of students at that grade level who obtained scores lower than that scale score. The percentile ranks provide relative information on the performance of students. The percentile ranks for the scores on the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics were calculated based on the May 1993 administration of the tests. The percentile tables are published in *State Norms Tables for North Carolina* (NCVDPI, 1995). Within a grade, meaningful comparisons can be made between the percentile ranks associated with the reading comprehension and mathematics scores.

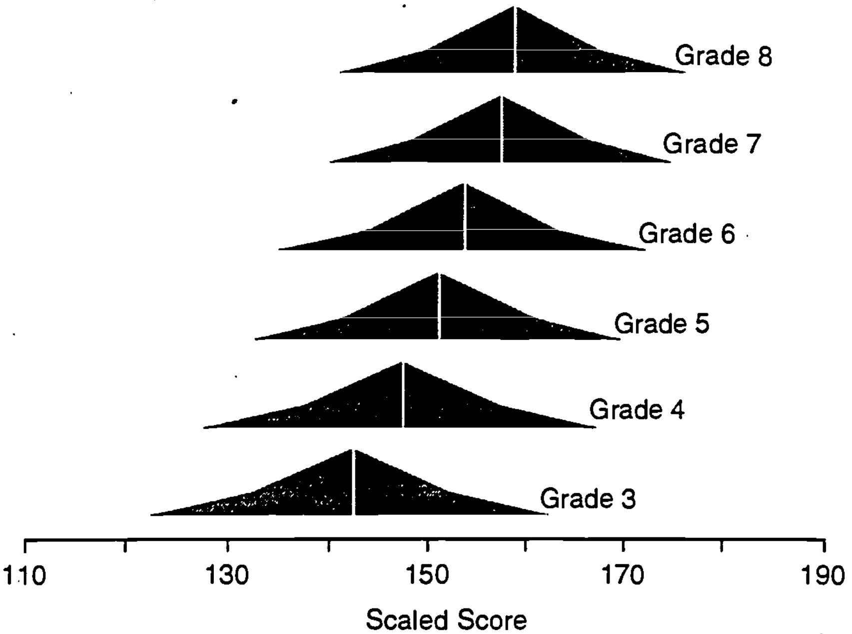


Figure 9. Grade distributions on the developmental scale for the North Carolina End-of-Grade Test of Reading Comprehension—1993 Forms A, B, and C.

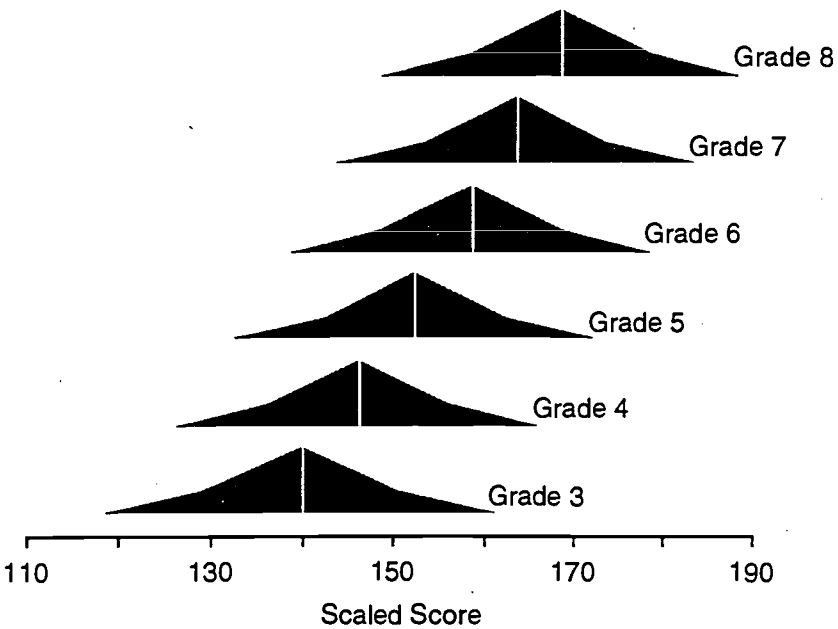


Figure 10. Grade distributions on the developmental scale for the North Carolina End-of-Grade Test of Mathematics—1993 Forms A, B, and C.

Descriptive Statistics

The North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics were first administered in May 1993—Forms A, B, and C. Three additional equivalent forms—Forms D, E, and F—were administered for the first time in May 1994. Table 16 presents the descriptive statistics from the first administration of the tests in May 1993 and Table 17 presents the descriptive statistics from the second administration of the tests in May 1994 (the first administration of forms D, E, and F).

Table 16. Descriptive Statistics for the North Carolina End-of-Grade Tests 1993 Administration—Forms A, B, and C.

Grade	N	Mean Scale Score	Mean by Form (Range)	Standard Deviation
Reading				
3	85,381	142.7	142.2–143.0	9.9
4	84,811	147.1	146.7–147.6	9.6
5	85,337	151.5	150.9–151.9	9.0
6	84,278	154.0	153.2–154.7	9.1
7	83,868	157.0	156.2–158.3	8.6
8	80,833	158.7	158.0–159.3	8.9
Mathematics				
3	85,026	139.9	139.3–140.1	11.3
4	84,453	146.1	145.6–146.4	10.5
5	84,999	152.3	152.2–152.4	9.7
6	83,683	158.3	158.2–158.5	10.1
7	83,143	164.1	163.6–164.5	10.0
8	80,032	168.3	167.7–168.6	10.6

Table 17. Descriptive Statistics for the North Carolina End-of-Grade Tests
1994 Administration—Forms D, E, and F.

Grade	N	Mean Scale Score	Mean by Form (Range)	Standard Deviation
Reading				
3	88,301	142.8	142.4–142.7	10.0
4	85,311	147.9	147.3–148.0	9.3
5	85,330	151.7	151.0–152.0	8.9
6	85,813	154.4	153.9–155.2	9.1
7	84,852	157.3	156.6–157.8	8.7
8	82,985	159.7	158.8–160.0	8.6
Mathematics				
3	88,414	140.0	139.6–140.1	11.5
4	85,363	147.2	146.4–147.3	10.7
5	85,384	153.5	153.0–153.5	10.0
6	85,850	159.4	159.1–159.3	10.2
7	84,768	164.8	164.7	10.4
8	82,793	169.0	168.5–169.3	11.0

Of special significance to the comparison of student scores across time, and scores in general across time, is the equivalence of the test forms. All six forms developed for a subject/grade (for example, grade 3 mathematics, Forms A, B, C, D, E, and F) were equated to the mean derived from the latent proficiency scaling of the tests (see Tables 14 and 15) using the EOG_SCAL.LSP program. From an examination of Tables 16 and 17, the differences between the mean scores across the forms are at or near zero and are always less than the standard error of measurement for the test (see Tables 19 and 20).

Figures 11 through 22 present the frequency distributions of the developmental scale scores from the May 1993 administration of the tests. The frequency distributions are not smooth because of the conversion from raw score to scale score. Due to rounding in the conversion process, sometimes two raw scores in the middle of the distribution convert to the same scale score.

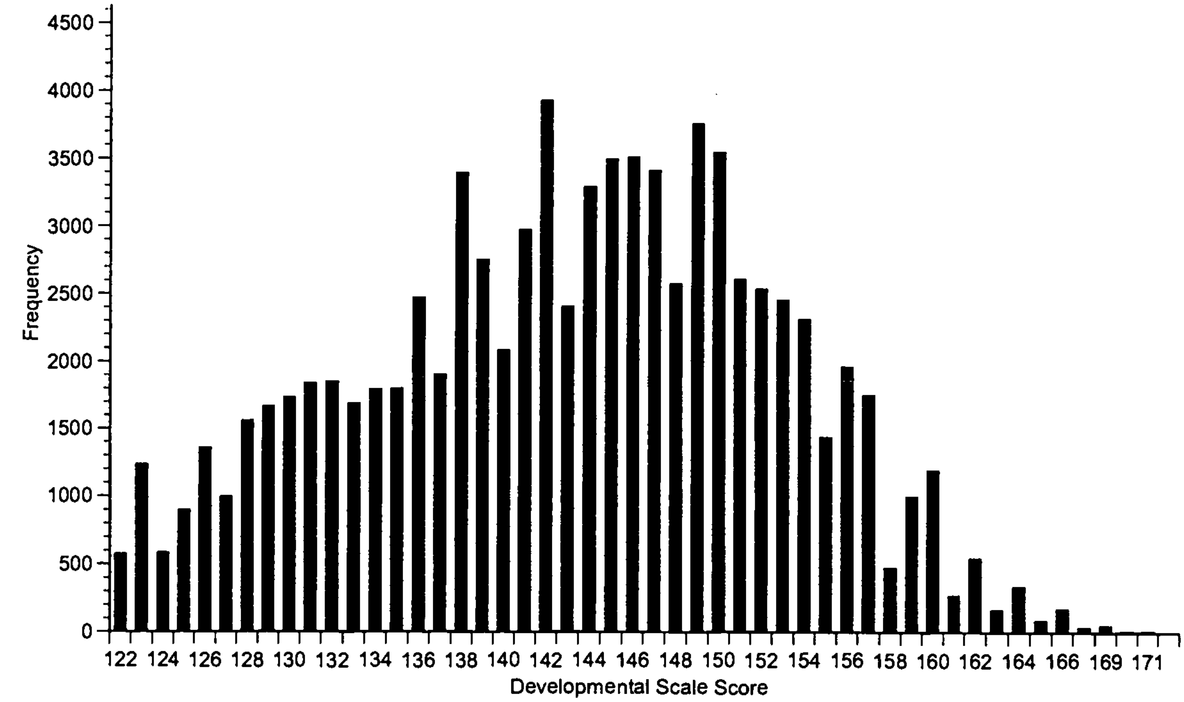


Figure 11. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 3, Forms A, B, and C (N = 85,381).

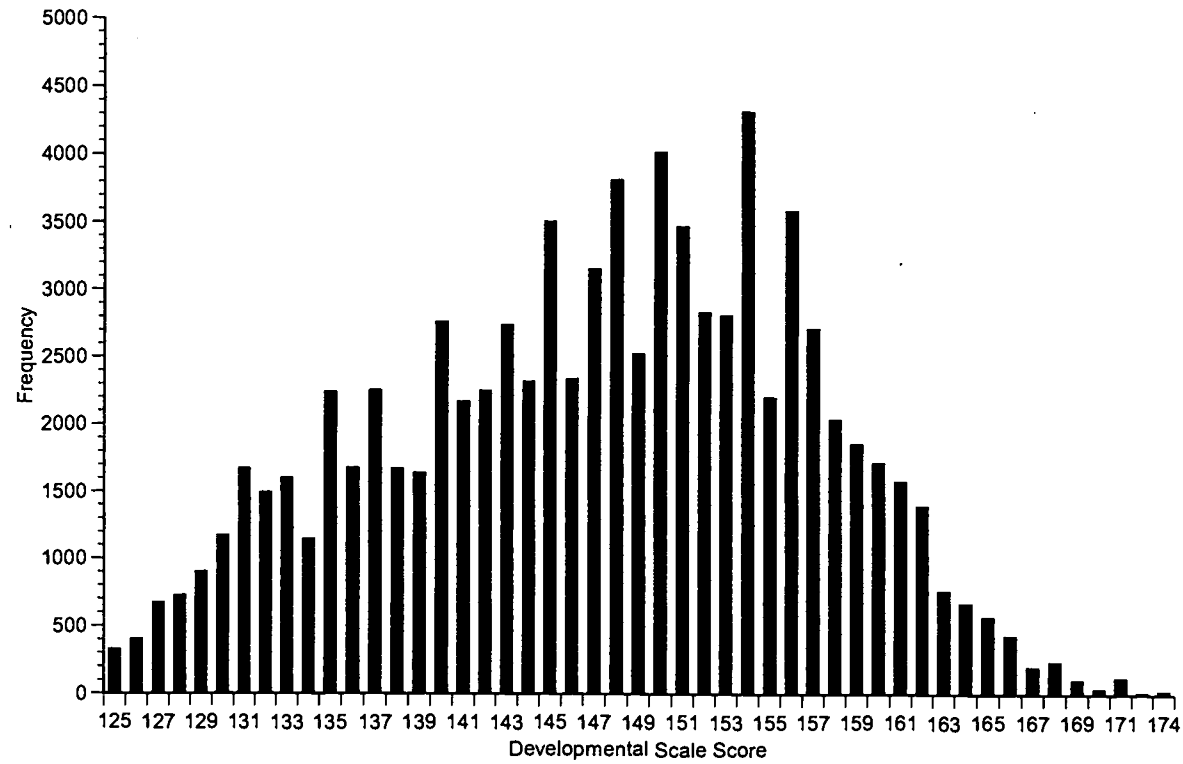


Figure 12. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 4, Forms A, B, and C (N = 84,811).

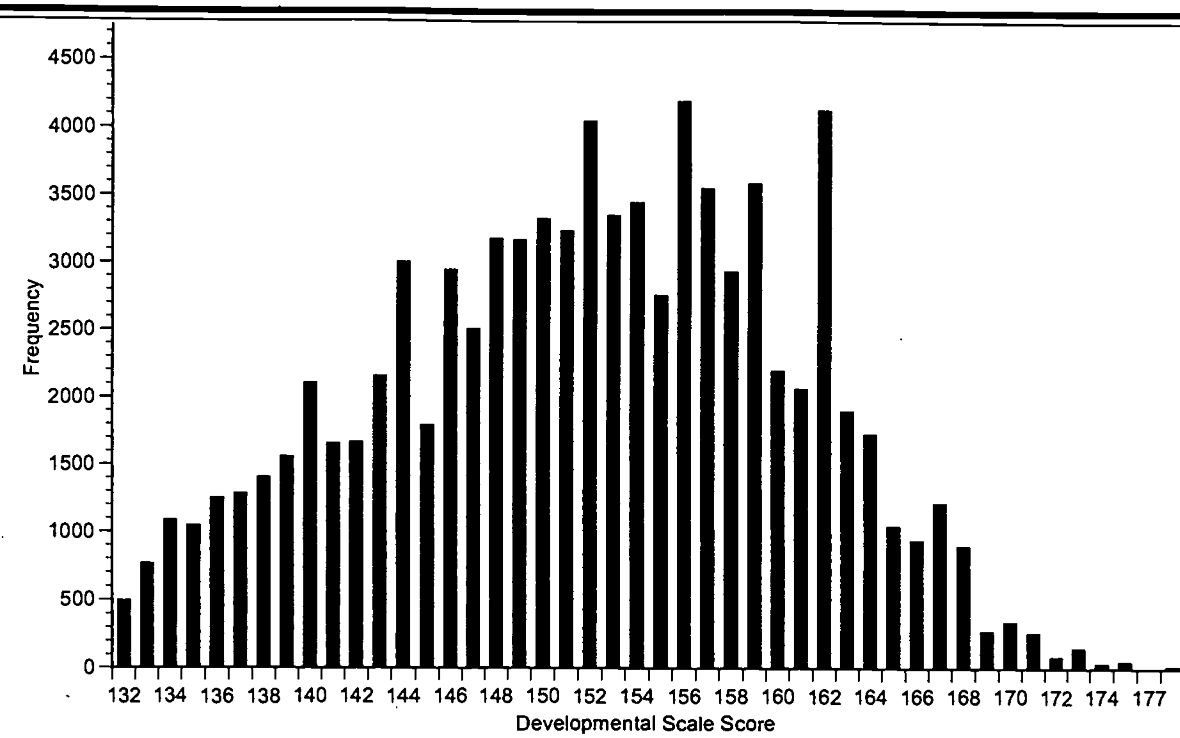


Figure 13. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 5, Forms A, B, and C (N = 85,337).

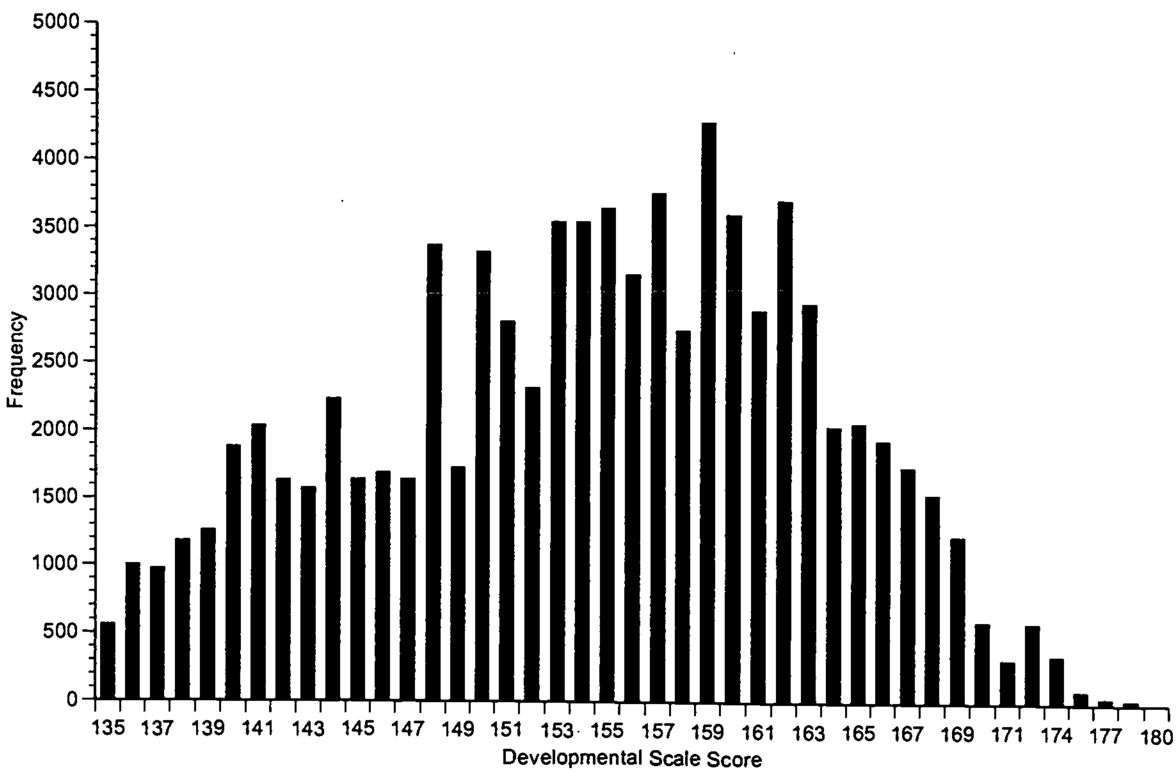


Figure 14. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 6, Forms A, B, and C (N = 84,278).

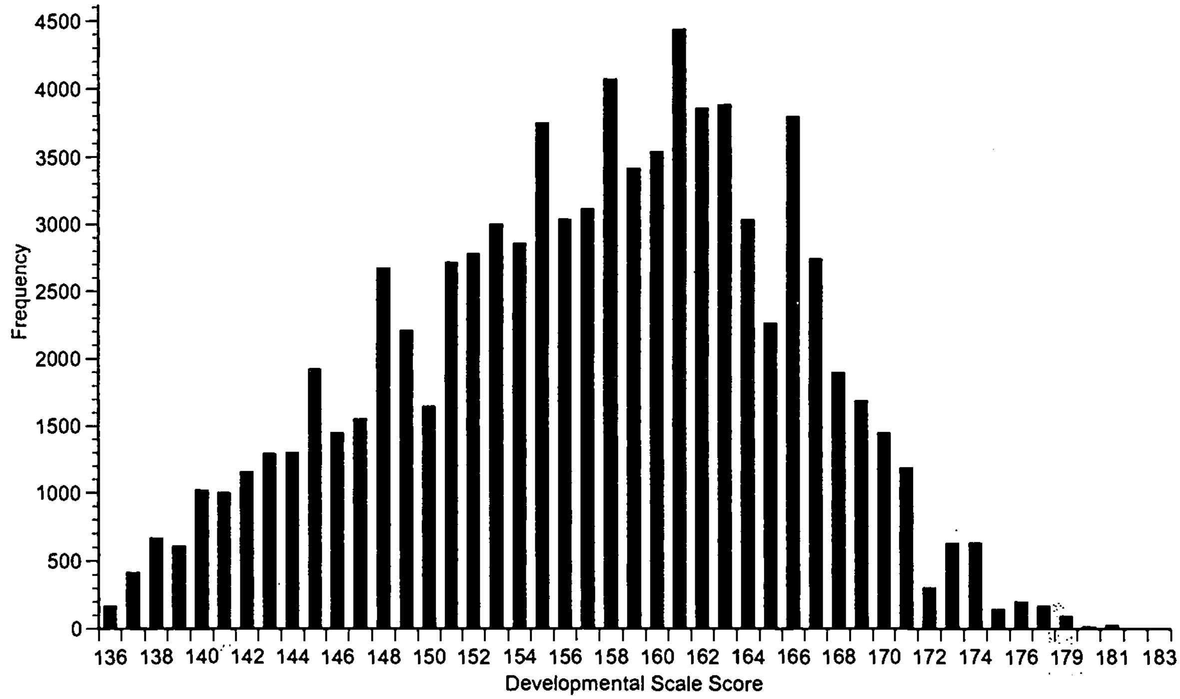


Figure 15. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 7, Forms A, B, and C (N = 83,868).

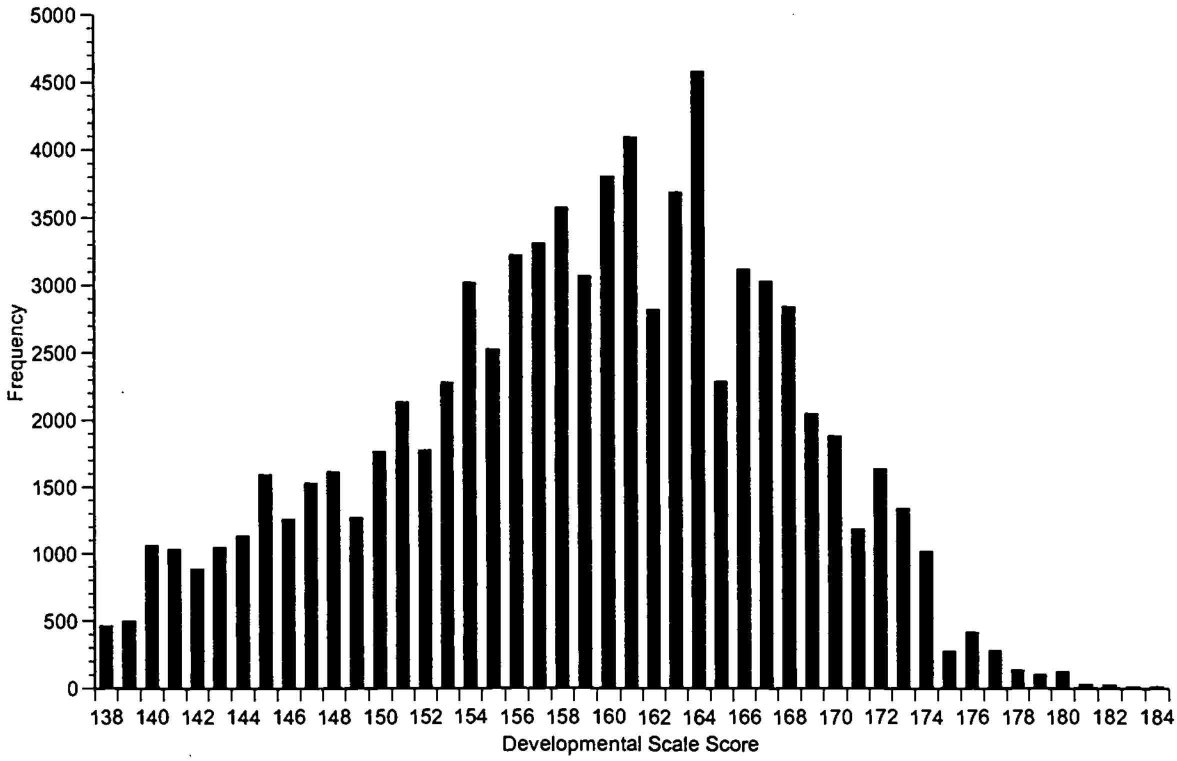


Figure 16. Frequency distribution of scores on the North Carolina End-of-Grade Test of Reading Comprehension—Grade 8, Forms A, B, and C (N = 80,833).

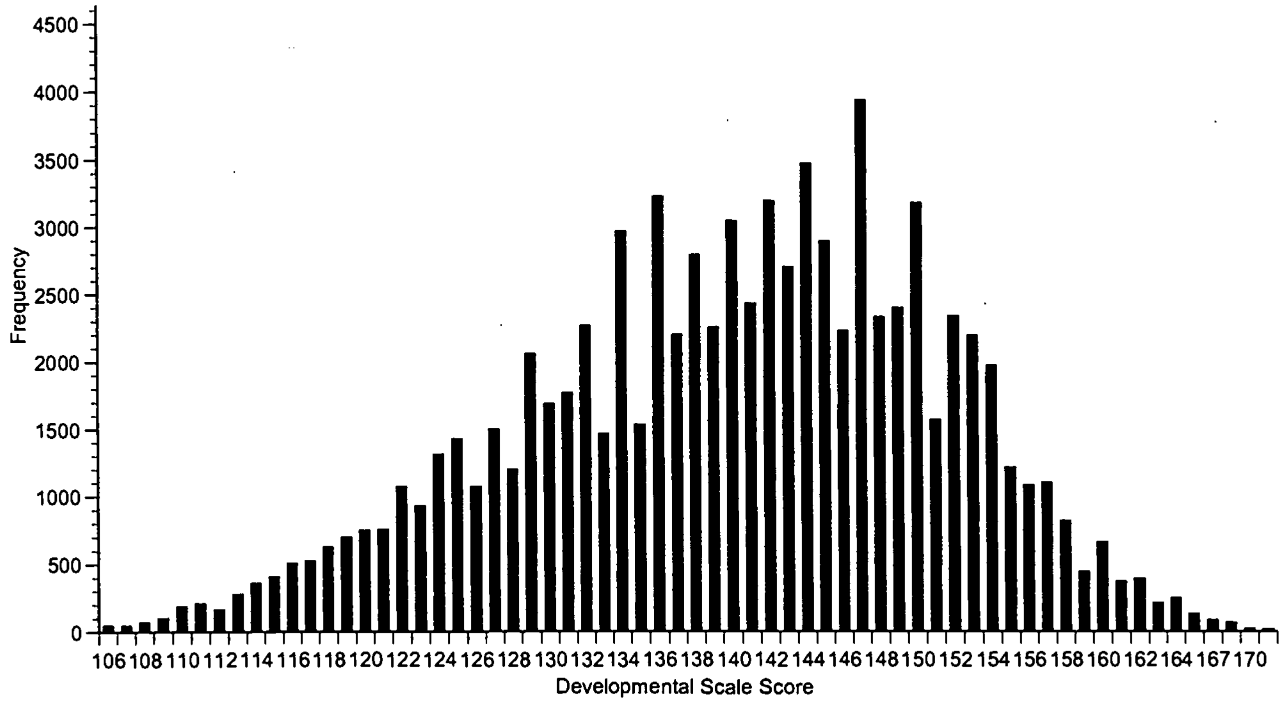


Figure 17. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 3, Forms A, B, and C (N = 85,026).

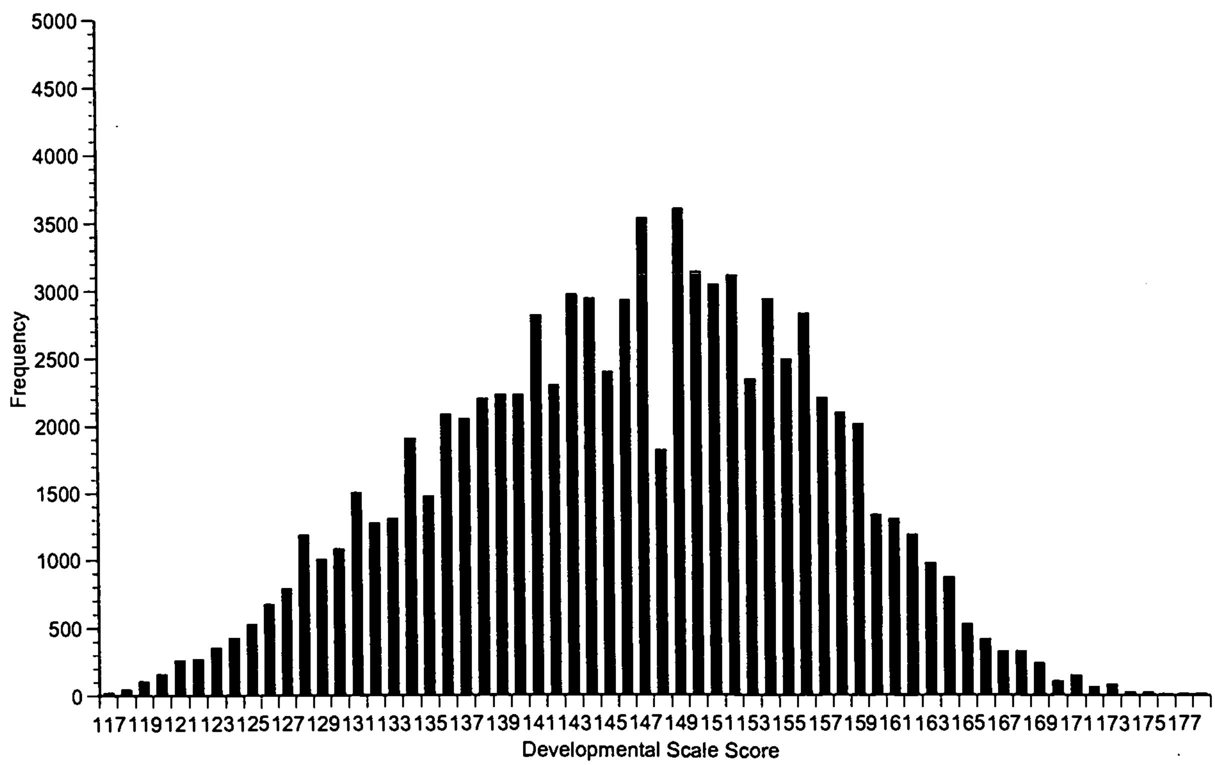


Figure 18. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 4, Forms A, B, and C (N = 84,453).

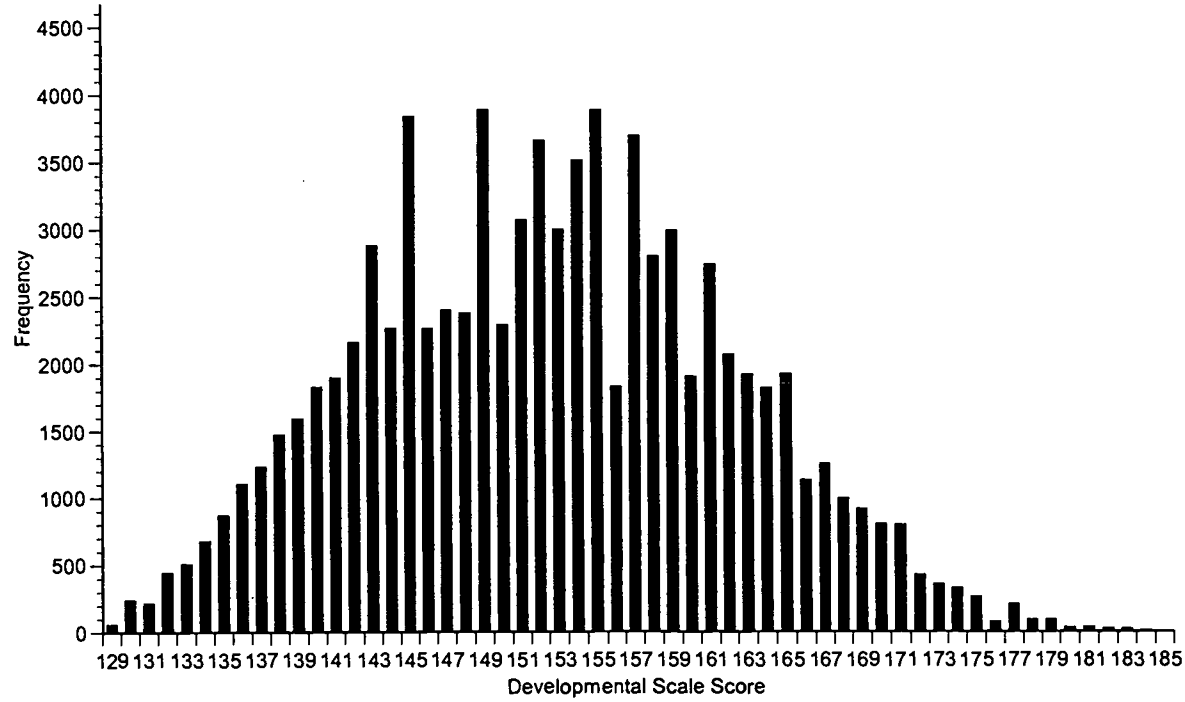


Figure 19. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 5, Forms A, B, and C (N = 84,999).

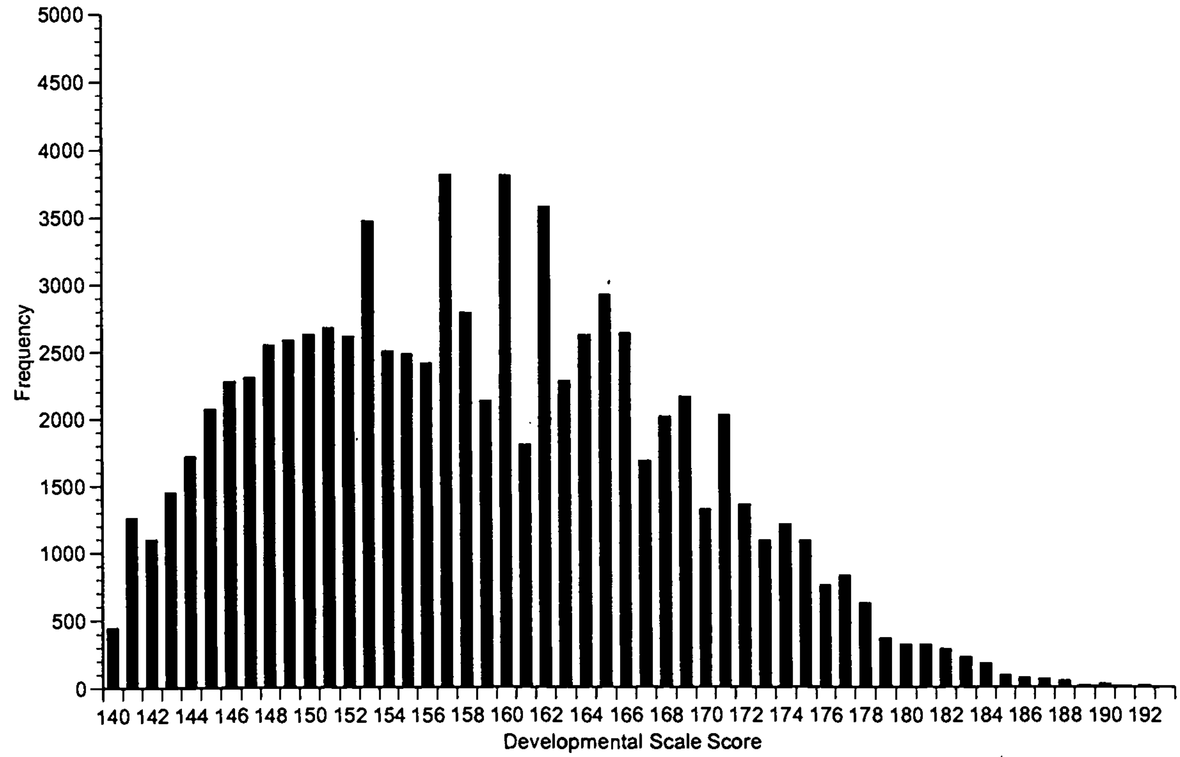


Figure 20. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 6, Forms A, B, and C (N = 83,683).

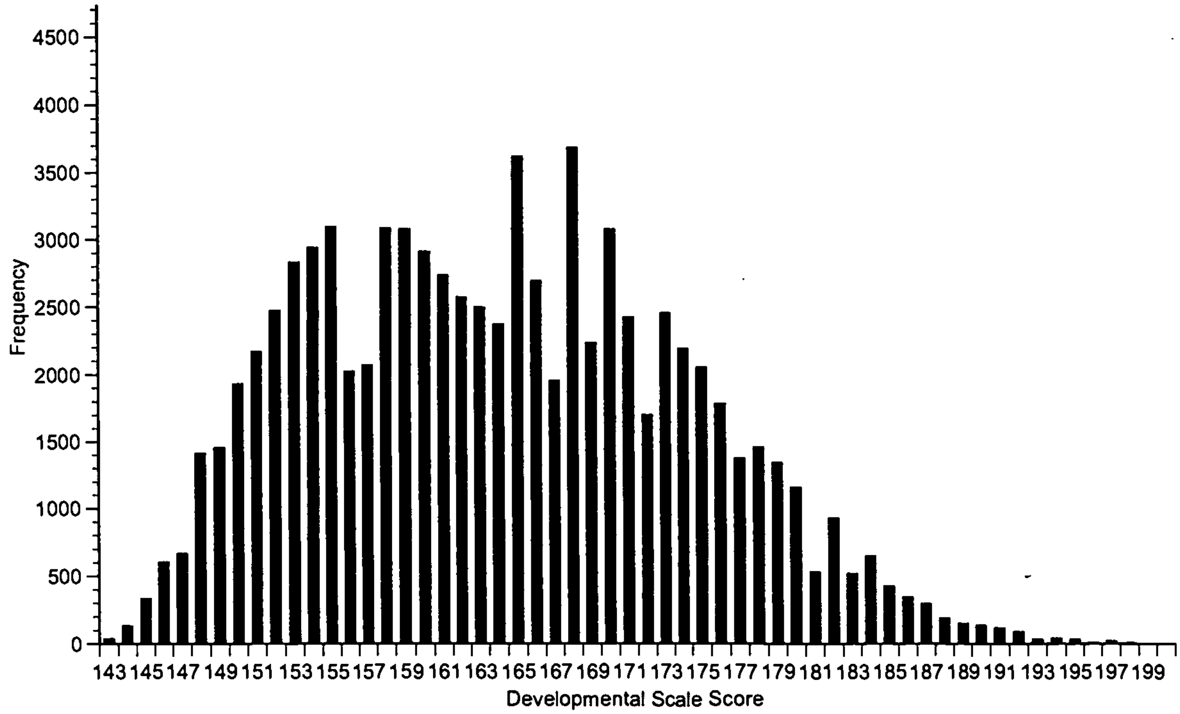


Figure 21. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 7, Forms A, B, and C (N = 83,143).

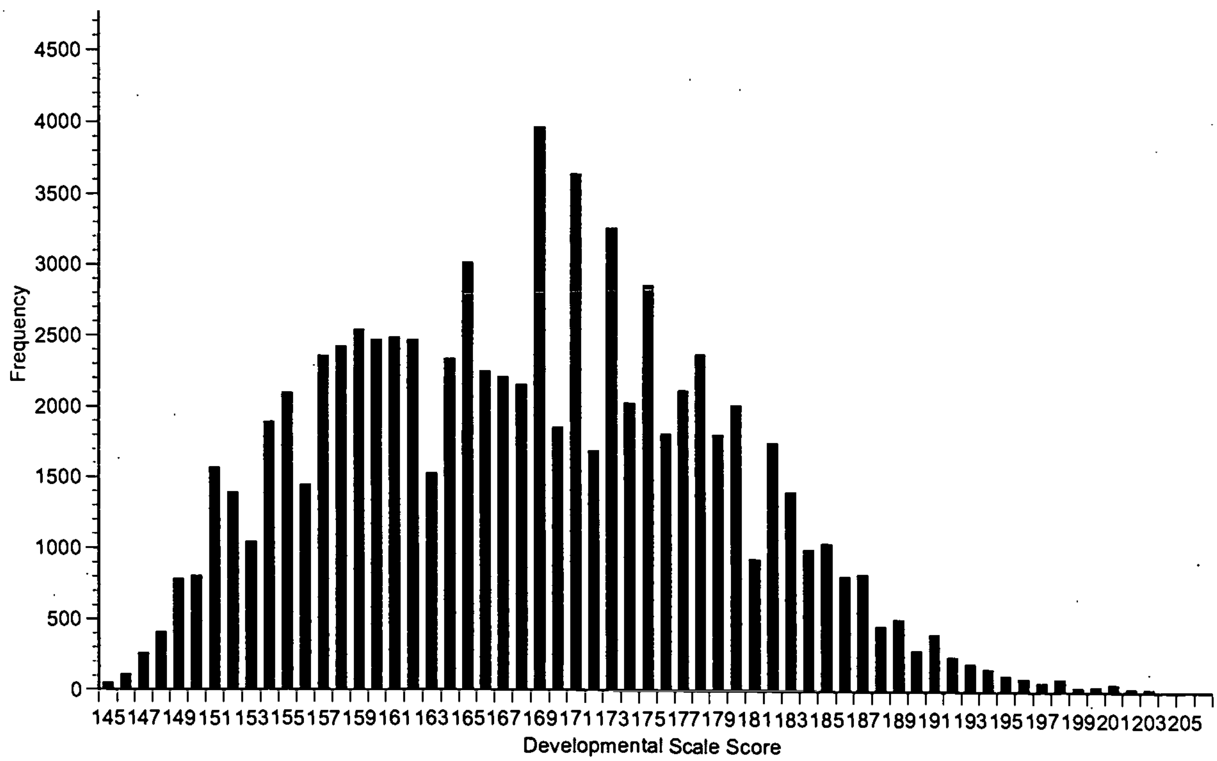


Figure 22. Frequency distribution of scores on the North Carolina End-of-Grade Test of Mathematics—Grade 8, Forms A, B, and C (N = 80,032).

Reliability

Reliability refers to the consistency of scores obtained by the same person when examined with the same test on different occasions or with different sets of equivalent items. If any use is to be made of the information from a test, then it is desirable that the test results be reliable. If decisions about individuals are to be made on the basis of the test data (for example, placement or instructional program decisions), then it is desirable that the test results be reliable and exhibit a reliability coefficient of at least 0.85. In testing, if use is to be made of some piece of information, then the information should be stable, consistent, and dependable.

Alternate-Form/Test-Retest Reliability Alternate-form reliability examines the extent to which two equivalent forms of a test yield the same results (students' scores have the same rank order on both tests). Test-retest reliability examines the extent to which two administrations of the same test yield similar results. In research done in one North Carolina school system, when a second form of the grade 7 reading comprehension test was administered to three classes of students (*n*s of 20, 23, and 27) one week apart, the reliability estimate was 0.86.

Internal-Consistency Reliability Internal-consistency reliability examines the extent to which the test measures a single basic concept. One procedure for determining the internal consistency of a test is coefficient alpha (α). Coefficient alpha sets an upper limit to the reliability of tests constructed in terms of the domain-sampling model. Table 18 presents the item- and passage-level values of coefficient α for the North Carolina Tests of Reading Comprehension and Mathematics. The passage-level coefficient α is slightly lower because this treats each of the passages as an item; therefore, the length of the test is reduced from 56 to 68 items to 10 items, decreasing the reliability.

Table 18. Item- and passage-level values of coefficient α for the 1993 administration of the North Carolina End-of-Grade Tests—Forms A, B, and C.

Grade	Mathematics	Reading Comprehension Item-Level	Reading Comprehension Passage-Level
3	0.94	0.92	0.90
4	0.94	0.94	0.92
5	0.92	0.93	0.91
6	0.92	0.94	0.92
7	0.91	0.93	0.92
8	0.92	0.93	0.92

Note: Passage-level coefficient α from Wainer and Thissen (1996)

Standard Error of Measurement The standard error of measurement of a test is the standard deviation of the error scores of a test. Typically the standard error of measurement is determined by the following formula:

$$\sigma_{\text{meas}} = SD_t \sqrt{1 - r_{tt}} \qquad \text{(Equation 4)}$$

When using item response theory measurement, the test information function, which depends only on the items included in the test, permits the estimation of the error of measurement at each ability level (or score). The standard error of measurement (estimation in IRT methodology) is determined by the following formula:

$$\text{SE}(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \qquad \text{(Equation 5)}$$

The magnitude of the standard error of θ (an examinee's estimated ability level) depends on the following characteristics of the test:

- the number of test items—smaller standard errors are associated with longer tests,
- the quality of the test items—in general, smaller standard errors are associated with highly discriminating items for which the correct answers cannot be obtained by guessing, and
- the match between item difficulty and examinee ability —smaller standard errors are associated with tests composed of items with difficulty parameters approximately equal to the the ability parameter of the examinee, as opposed to tests that are relatively easy or relatively difficult (Hambleton, Swaminathan, and Rogers, 1991).

Tables 19 and 20 show the standard error of measurement ranges for scores on the North Carolina End-of-Grade Tests. For students with scores within two standard deviations of the mean (95% of the students), standard errors are typically 2 to 3 points. As scores become more extreme there is less measurement precision associated with a score.

Table 19. Standard error of measurment for ranges of scores on the North Carolina End-of-Grade Test of Reading Comprehension.

Score	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
190						
180						4
170			3	3	3	2
160	3	2	2	2	2	2
150	3	2	2	2	3	3
140	2	3	4	5	5	
130	4	5				
120						
110						

Note: From Pommerich, Billeaud, Williams, and Thissen (1993)

Table 20. Standard error of measurment for ranges of scores on the North Carolina End-of-Grade Test of Mathematics.

Score	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
190					3	3
180			3	2	2	2
170		4	2	2	2	3
160	4	3	2	3	4	5
150	3	2	3	5	6	
140	2	3	5			
130	3	4				
120	4					
110						

Note: From Pommerich, Billeaud, Williams, and Thissen (1993)

The validity of a test is the degree to which the test actually measures what it purports to measure. Validity provides a direct check on how well the test fulfills its function. For all forms of test development, validity is a predominant theme from the time the idea for the test is conceived until the final test scores have been analyzed and interpreted. For convenience, the various components of test validity—content, criterion-related, and construct—will be described as if they were unique, independent components rather than interrelated parts.

Content Validity

The content validity of a test relates to the adequacy with which important content has been sampled and the adequacy with which the content is evidenced in the test items. Content validity was built into the North Carolina End-of-Grade Tests during the development process. All items are aligned with the North Carolina *Standard Course of Study* for Mathematics and English Language Arts, the basis for instruction in North Carolina schools. The items were written and reviewed by North Carolina teachers who are in contact with the students every day in the classroom.

Criterion-Related Validity

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 the test can be measured at the same time (concurrent validity) or at some later time (predictive validity). The following discussion of the relationship between scores on the North Carolina End-of-Grade Tests and teacher judgments of student achievement is evidence of concurrent validity.

Achievement Levels The North Carolina *Standard Course of Study* outlines the content standards for North Carolina in that it describes the knowledge, skills, and other understanding that schools should teach in order for students to attain high levels of competency in challenging subject matter. Educators in North Carolina felt that performance standards should also be developed which identify levels of competency expected in each content area. Unlike percentiles, which yield only relative comparisons, the performance standards give common meaning throughout the state as to what is expected at various levels of competence in each subject area. Performance standards, called Achievement Levels, are one way that scores on the North Carolina End-of-Grade Tests are reported. These categories are used to better describe the scores on the tests and are based on external evidence about the relative skill of students.

The achievement levels for the North Carolina End-of-Grade Tests are based on the contrasting groups method of standard setting. This method involves having students categorized into the various achievement levels by expert judges who are knowledgeable of the students' achievement in various domains assessed outside of the testing situation. Teachers are able to make informed judgments about students' achievement because the teachers have observed the breadth and depth of the work each student has accomplished during the school year.

During field testing (May 1992), teachers were asked to categorize each of their students on the basis of "absolute" achievement (comparison to an external standard). Each student was categorized into one of four achievement levels based on the teacher's experiences with the student throughout the school year.

Level I	Fails to achieve at a basic level: 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	Achieves at a *basic* level: 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	Achieves at a *proficient* level: 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	Achieves at an *advanced* level: Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient at grade level work.
	or
	Not a clear example of any of these achievement levels.

In all, the judgments of more than 5,000 teachers about the performance of more than 160,000 students were involved in the standard setting process statewide. More than 95% of the students field tested were categorized into one of the four achievement levels, with the remainder categorized as not a clear example of any of the achievement levels. The verbal descriptors "below basic," "basic," "proficient," and "advanced" were dropped after the field testing to avoid confusion with the NAEP achievement levels and to lessen the impact of labeling students, especially at the "below basic" level.

The percentage of students in each achievement level were remarkably similar across subjects and grades. The percentages are presented in Table 21 below.

Table 21. Percent of students assigned to each achievement level by teachers (May 1992).

Subject/Grade		Level I	Level II	Level III	Level IV
Reading	3	14.3%	26.9%	37.8%	21.1%
	4	12.5%	28.5%	39.6%	19.5%
	5	10.7%	28.3%	40.1%	20.9%
	6	11.1%	27.7%	41.2%	19.9%
	7	11.1%	28.7%	38.3%	21.9%
	8	9.0%	26.2%	41.2%	23.6%
Mathematics	3	12.0%	28.1%	40.6%	19.2%
	4	10.3%	27.2%	42.8%	19.6%
	5	13.0%	27.8%	40.8%	18.3%
	6	12.1%	28.1%	40.4%	19.4%
	7	12.4%	27.9%	39.8%	19.9%
	8	11.2%	28.8%	40.4%	19.6%

Figures 23 and 24 show the relationship between students' scores on the field test with the teacher judgements concerning achievement (central two-thirds of scores for each achivement level). As expected, the scaled scores increase over the achievement levels, and also across grades. Students rated by teachers as high achievers (Level IV) scored high on the tests, while students who were rated low by teachers scored low on the test (Level I).

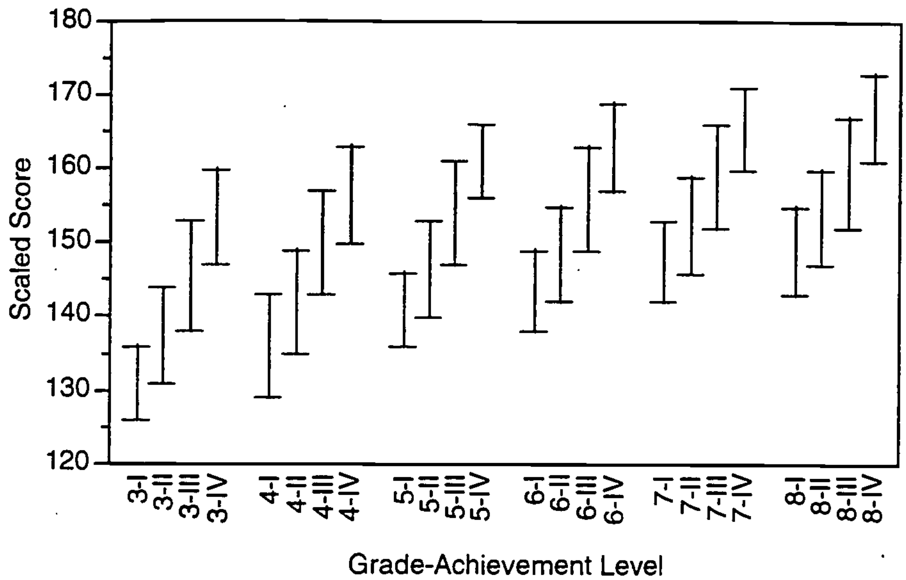


Figure 23.The relationship between teacher judgments of student achievement and scores on the North Carolina End-of-Grade Test of Reading Comprehension field test (May 1992).

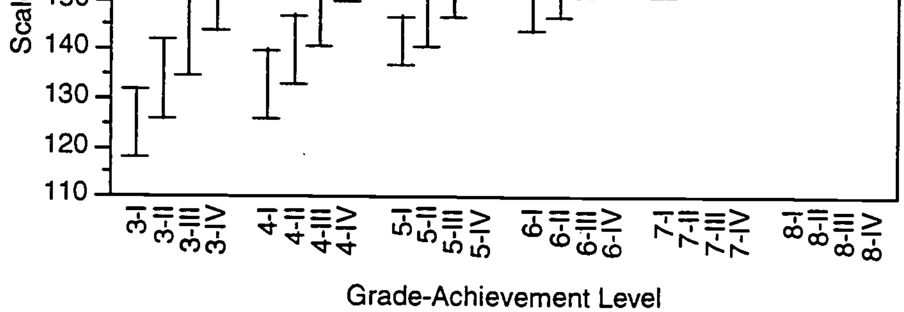


Figure 24. The relationship between teacher judgments of student achievement and scores on the North Carolina End-of-Grade Test of Mathematics field test (May 1992).

The percentages of students shown in Table 21 for each subject and grade were used in conjunction with the frequency distributions of scores from the first administration of the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics administered in May 1993 to determine where the cut points should be for the achievement levels. Table 22 gives the range of scores associated with each achievement level.

Table 22. Range of scores associated with each achievement level for score reporting.

Subject/Grade		Level I	Level II	Level III	Level IV
Reading	3	115-130	131-140	141-150	151-172
	4	119-134	135-144	145-155	156-174
	5	124-138	139-148	149-158	159-178
	6	128-140	141-151	152-161	162-180
	7	130-144	145-154	155-163	164-183
	8	132-144	145-155	156-165	166-184
Mathematics	3	98-124	125-137	138-149	150-171
	4	111-131	132-142	143-155	156-178
	5	117-140	141-149	150-160	161-185
	6	130-145	146-154	155-167	168-193
	7	138-151	152-160	161-172	173-201
	8	140-154	155-164	165-177	178-206

Construct Validity

The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait, such as reading comprehension or mathematics achievement. "Correlations between a new test and other similar tests . . . [are] evidence that the new test measures approximately the same general area of behavior as other tests designated by the same name" (Anastasi, 1982). The following sections provide evidence of the construct validity of the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics.

North Carolina Open-Ended Tests The North Carolina Open-Ended Tests were designed to measure broadly higher level thinking skills by requiring students to apply or demonstrate skills and knowledge beyond the recall level. These items were designed to be open so that the quality of the student's response would determine his or her score. Each form of the test contained items that assessed the reading strand of the English Language Arts *Standard Course of Study* and items that assessed the mathematics *Standard Course of Study*. Some of the mathematics items required the production of a specific answer to a problem, but the student was also asked to explain how he or she arrived at the answer. This explanation helped to determine the student's score.

The open-ended items were written and reviewed by advisory committees composed of testing consultants, teachers, curriculum specialists, and university professors. The items were field tested on approximately 500 students randomly selected from across the state to examine how each item performed (score distribution) and to refine the scoring rubric. Specific scoring rubrics, based on the general scoring rubric for the content area, describe the standards used to judge specific items within the content area. The items were field tested a second time to verify the scoring rubric. Results were analyzed using Samejima's graded item response theory model (Hambleton, Swaminathan, & Rogers, 1991). The field test information was examined for each item and the decision was made to retain the item for future test development or to delete the item at this time.

Tests were developed to reflect the breadth and depth of the curriulum. Ten reading and ten mathematics items were selected for each grade level that coverend the content for the grade level as defined by the North Carolina *Standard Course of Study*. Three test forms were developed with either 3 or 4 reading items, 3 or 4 mathematics items, and 3 or 4 social studies items, for a total of 10 items on each form.

Statewide scoring of the open-ended items was conducted in the summer at a central location with trained readers for each grade level. Inter-rater agreement averaged 0.68 across grades 3 through 8. Using projection, developmental scale scores for the open-ended items were derived from the multiple choice developmental scales. Scores on the open-ended tests were reported at the school and school system level in terms of developmental scale scores and percentiles. The standard errors of measurement for the open-ended tests were 7 to 8 scale score points near the middle of the distribution.

Table 23 shows the correlations between the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics and the North Carolina Open-Ended Test which measure the same content. The correlations range from the mid 50s for reading to the upper 60s for mathematics. Considering the differences in format, this is the level of correlation that would be expected between the two sets of scores in a multi-method design (Campbell and Fiske, 1959).

Table 23. Correlations between the North Carolina Open-Ended Tests and the North Carolina End-of-Grade Multiple-Choice Tests.

Grade	Correlation Between Open-Ended Mathematics and Multiple-Choice Mathematics	Correlation Between Open-Ended Reading and Multiple-Choice Reading
3	0.66	0.58
4	0.65	0.55
5	0.64	0.57
6	0.68	0.53
7	0.67	0.57
8	0.66	0.54

North Carolina End-of-Course Tests of English I and Algebra I The North Carolina Tests of English I and Algebra I are achievement tests developed by the NCDPI for the assessment of achievement at the end of the English I and Algebra I courses (typically grade 9 courses). The English I test assesses the same goals and objectives as the end-of-grade reading test assesses; the only difference between the two tests is that the English I test also assesses the student's ability to edit for grammar, mechanics, usage, and spelling. The Algebra I test assesses goals 3 (pre-algebra), 5 (problem solving), 6 (statistics and data analysis), and 7 (computation) of the mathematics *Standard Course of Study* for grades 3 through 8. For information concerning the psychometric characteristics of the English I and Algebra I tests see *Technical Characteristics of the North Carolina End-of-Course Tests* (NCDPI, 1996).

The scores of students who were administered the 1995 North Carolina End-of-Course Tests of English I and Algebra I were matched with the their scores on the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics in Grade 8. A total of 43,194 students had both English I and end-of-grade reading test scores. The correlation between the two sets of scores was 0.81. A total of 27,076 students had both Algebra I and end-of-grade mathematics test scores. The correlation between the two sets of scores was 0.73.

Iowa Tests of Basic Skills (ITBS) Public School Law 115C-174.11 (a) states that "if the State Board of Education finds that testing in grades other than first and second grade is necessary to allow comparisons with national indicators of student achievement, that testing shall be conducted with the smallest size sample of students necessary to assure valid comparisons with other states." In 1992, after evaluating several nationally-normed tests that had been recently re-normed, the Iowa Tests of Basic Skills (ITBS) was selected to be administered each year to a sample of North Carolina students to provide such comparisons.

The ITBS assesses reading, language, and mathematics on the survey battery. On the reading subtest, students are asked to construct literal meanings from a text, and to go beyond the test to interpret and infer underlying, unstated meanings. Students must also be able to construct evaluative meanings and make judgments about the author's craft. Vocabulary is assessed in context. Like the North Carolina End-of-Grade Tests of Reading, the ITBS reading subtest is aligned with the International Reading Association standards. The mathematics subtest consists of four parts—concepts, estimation, problem solving, and data interpretation—and reflects the general objectives of mathematics instruction relating to understanding quantitative processes and using mathematics in problem solving. Computation is assessed separately. Like the North Carolina End-of-Grade Test of Mathematics, the ITBS mathematics subtest is aligned with the National Council of Teachers of Mathematics standards. While similiar in overall content, the ITBS and the North Carolina End-of-Grade Tests are different in philosophy. The End-of-Grade test for each grade assesses the knowledge and skills that should be taught at *that particular grade level*; the ITBS assesses the knowledge and skills taught across *two or three grade levels* to students in North Carolina.

The ITBS Survey Battery (Form K) is administered each year to a representative sample of approximately 3,000 grade 5 students and approximately 3,000 grade 8 students. The schools were selected on the basis of the gender and ethnicity/racial characteristics of the student population and on the basis of size and geographic location and then verified by comparing the characteristics of the schools selected with those of the state. Approximately 30 students were selected at random from each school. The number of students tested each year at grades 5 and 8 and the correlations with the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics are presented in Table 24. The ITBS Reading Total and the Mathematics Total (without computation) scores are the most similar to the scores on the end-of-grade tests and only those results are presented.

Table 24. Correlations between the Iowa Tests of Basic Skills (ITBS) and the North Carolina End-of-Grade Tests of Reading and Mathematics—Grades 5 and 8.

Grade	Year	Correlation Between ITBS Mathematics and EOG Mathematics	Correlation Between ITBS Reading and EOG Reading
5	1993	0.84	0.82
	1994	0.80	0.79
	1995	0.83	0.81
8	1993	0.78	0.77
	1994	0.79	0.76
	1995	0.78	0.77

Sample Sizes for Grade 5: 1993—2,606: 1994—2,815; 1995—2,823
Sample Sizes for Grade 8: 1993—2,605; 1994—2,709: 1995—2,819

National Assessment of Educational Progress (NAEP)—Grade 8 Mathematics NAEP is a congressionally-mandated survey of the achievement of the nations fourth-, eighth-, and twelfth-grade students. NAEP assessments are administered every two years in areas such as reading, writing, mathematics, science, history, and geography. Like the North Carolina End-of-Grade Tests, NAEP mathematics tests are aligned with the National Council of Teachers of Mathematics standards. The NAEP mathematics assessments are organized according to three mathematical abilities—conceptual understanding, procedural knowledge, and problem solving—and five content areas—numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions.

In 1990, NAEP began a voluntary state-by-state assessment program (Trial State Assessment—TSA) which allows states to compare their achievement with that of other states and with the nation as a whole. North Carolina was among 37 states that participated in the grade 8 mathematics assessment in 1990 and among 42 states that participated in 1992. In 1994 the United States Congress did not fund state-level NAEP assessment.

In a special arrangement with the National Center for Educational Statistics and the Educational Testing Service, North Carolina re-administered two blocks of items from the 1992 NAEP assessment to a sample of North Carolina eighth-grade students in February 1994 (N = 2,824 students in 103 schools in a sample drawn by using the national sampling frame for the 1994 TSA). In addition, a short form of the North Carolina End-of-Grade Tests: Grade 8 Mathematics was administered to the same students. The purpose of this special administration was to link the North Carolina End-of-Grade Mathematics Test to the NAEP scale. In the study, Williams, Billeaud, Davis, Thissen, and Sanford (1995) observed that "there is considerable overlap in the content frameworks of the two tests" (p. 8). They observed a correlation of 0.70 between the North Carolina End-of-Grade Test of Mathematics (Grade 8) and the NAEP Grade 8 special assessment.

Lexile Framework The Lexile Framework, a measure of reading fluency, was developed based on construct generalization. This theory permits the linkage of text environments to a student's ability level. This means that a student's ability to comprehend reading materials can be linked to any text environment, including textbooks, tests, manuals, newspapers, or curriculum materials. The Lexile Framework enables the expression of any measure of reading ability in terms that have concrete meaning. For example, a school system can determine how their text materials relate to their student's ability level based on end-of-grade scores; teachers can evaluate their curriculum based on each student's ability to comprehend the materials; and parents, students, and educators can be provided concrete, real-life information about the level of ability required to comprehend text environments at each grade level.

The Lexile Framework, developed by Metametrics, is based on research investigating how students acquire reading skills. The work of Chall, Flesch, Carroll, and Bromuth concerning readability and of Rasch in measurement were instrumental in developing the Lexile Framework. Rasch calibration of reading material permits the conversion of counts correct on tests to an objective measure of reading.

In order to link the Lexile Framework to the North Carolina End-of-Grade Test of Reading Comprehension, Metametrics conducted a linking study in the Spring of 1995. Because the Lexile theory provides complementary procedures for measuring people and text, the scale was used to match a person's level of comprehension with books that the person is predicted to read with a high comprehension rate. A convienience sample of 250 students at each of grades 3, 4,5, and 8 were administered a Lexile reading inventory and the North Carolina End-of-Grade Test or Reading Comprehension within a two-week interval. Results from the testing were plotted and an overall correlation of .90 was observed between the Lexile Test and the North Carolina End-of-Grade Test of Reaidng Comprehension administeed in grades 3 through 8 (separately, grade 3, $r = 0.90$; grade 4, $r = 0.88$; grade 5, $r = 0.87$; and grade 8, $r = 0.88$). Based on regression analyses, a conversion table was established between the two tests permitting the end-of-grade test score to be expressed as a Lexile measure (standard errors associated with each score were establised through boot-strap procedures).

Table 25. Linear linking of the Lexile Framework with the North Carolina End-of-Grade Test of Reading Comprehension.

Score on the North Carolina End-of-Grade Test	Corresponding Lexile Framework Score	Stanard Error of Lexile Framework Score
130	208	18
134	318	15
138	429	13
142	539	10
146	649	8
150	760	7
154	870	7
158	981	8
162	1091	11
166	1202	13
170	1312	16
174	1423	19

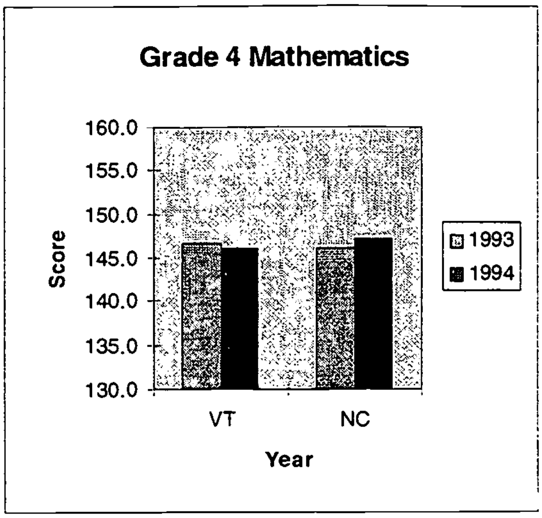
As part of the linkage of the end-of-grade tests to the Lexile Frawework, the Lexile MAP was developed (North Carolina version). The MAP provides a point of reference to the state's standards by correlating well-known material at each grade level with scores on the end-of-grade test. The top of the MAP identifies different categories of material that are familiar to teachers such as the titles of modern classics, everyday world items, periodicals, textbooks, assessment instruments, and workplace examples. Literature titles include works such as *The Last of the Mohicans*, calibrated at 1340 Lexiles (1340L), *Little Women* (1100L), and *Sarah Plain and Tall* (540L). The titles were drawn from the North Carolina reading list along with nationally recommended reading lists representing the range of material from children's first books to adult documents. The "real-world" items on the Lexile MAP exemplify material that adults encounter in their day-to-day lives. For example, a standard credit card application has a calibration of 1400L while a board game instruction, such as *CLUE*, has a calibration of 820L. The periodicals on the map are widely read magazines and most of them fall within an adult reading level that ranges from 1200L to 1400L. Such periodicals include *Newsweek* (1270L), *Fortune* (1300L), and the *Wall Street Journal* (1400L). In addition, the major newspapers from across the state will be included in order to provide a local reference point for the students. The textbooks column will cover a range of subject areas including elementary series such as DC Heath's *Come Back Here Crocodile* (180L) to McGraw-Hill's college level text, *Human Anatomy and Physiology* (1450L).

Vermont Uniform Assessment The Vermont Uniform Assessment is administered at grades 4 and 8 in writing and mathematics. The Uniform Assessment is one part of the voluntary assessment program in Vermont that includes the development of a student portfolio for each content area. The mathematics curriculum in Vermont, like North Carolina's, is based on the National Council of Teachers of Mathematics standards.

The Vermont Uniform Assessment of mathematics is composed of items developed in North Carolina as part of the End-of-Grade Testing Program at grades 4 and 8. In an agreement with Vermont, the raw data from the assessment is returned to North Carolina each year. Based on the items selected for inclusion in the assessment (and the associated item parameters), the program EOG_SCAL.LSP was used to generate raw-score-to-scale-score conversion tables for the grades 4 and 8 Vermont Uniform Assessments on the North Carolina mathematics developmental scale.

Figure 25 shows the performance of North Carolina students compared to Vermont students in 1993 and 1994. Based upon the information in the graphs, North Carolina students and Vermont students are achieving at a similar level in both grades 4 and 8. Based on the actual scores, the following conclusions can be drawn:

- in 1993, North Carolina grade 4 students did not score significantly different from Vermont grade 4 students (difference of $^{-}0.48$);
- in 1993, North Carolina grade 8 students scored significantly lower than Vermont grade 8 students (difference of $^{-}1.71$ between means, significant at $\alpha = 0.05$);
- in 1994, North Carolina grade 4 students scored significantly higher than Vermont grade 4 students (difference of 0.8 between means, significant at $\alpha = 0.05$); and
- in 1994, North Carolina grade 8 students scored significantly lower than Vermont grade 8 students (difference of $^{-}1.05$ between means, significant at $\alpha = 0.05$).



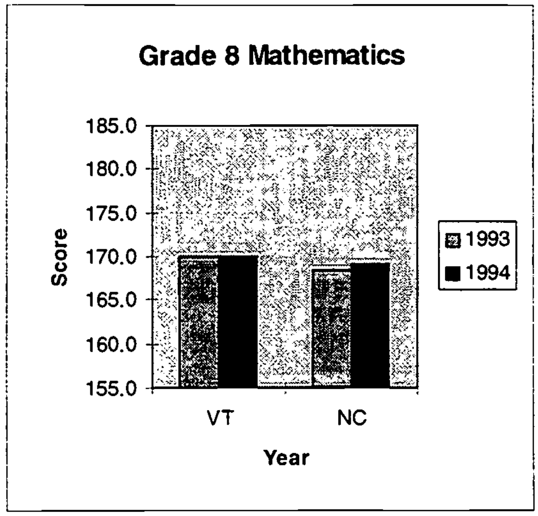


Figure 25. Comparison of North Carolina and Vermont students on the North Carolina developmental scale for mathematics in 1993 and 1994.

Growth Study This study was undertaken to examine the growth of reading comprehension and mathematics as measured by the end-of-grade tests. In March 1995, samples of students and adults were administered the North Carolina End-of-Grade Test of Reading Comprehension (Grade 8) or the North Carolina End-of-Grade Test of Mathematics (Grade 8). Students and adults were selected as follows:

- grade 8 sample—based on the grade 7 achievement level (reading or mathematics): 25% in Level I, 25% in Level II, 25% in Level III, and 25% in Level IV (104 students);
- grade 12—50% who planned to go on to college and 50% who planned to enter the work force after graduation (44 students); and
- adults—25% who had been in the work force for more than 10 years and only had a high school education, 25% who were attending technical college or had completed technical college, 25% who were employed with a four-year degree, and 25% who had just graduated from college (42 adults).

Half of each sample took the reading comprehension test and half took the mathematics tests. The results are presented in Table 26.

Table 26. Mean developmental scale scores on the North Carolina End-of-Grade Tests of Reading Comprehension and Mathematics.

Group	Mean EOG Reading Score	Mean EOG Mathematics Score
Grade 8 (Grade 7 Score)	159.6	169.3
Grade 8 (March 1995/7.5)	159.8	172.6
Grade 12	165.9	182.4
Adults	171.6	177.5

Consistent with the previous work with the reading developmental scale in grades 3 through 8, reading grows at a slower rate in middle school and high school (about 1 point per year) than in elementary school (see Table 14), but it does continue growing into adulthood. Mathematics, on the other hand, continued to grow at a steady rate from middle school through high school (about 2 to 3 points per year). There was a sharp decrease in the mean mathematics scores in adulthood. This decrease may be explained by the changes in the mathematics curriculum—the shift to more data analysis, "real world" interpretations of mathematical concepts, and explaining how to solve a problem (not just solving it)—or because mathematics skills, when not used routinely, are lost.

Anastasi, A. (1982). *Psychological Testing*. New York: Macmillan Publishing Company, Inc.

Averett, C.P. (1994). North Carolina End-of-Grade Tests: Setting standards for the achievement levels. Unpublished manuscript.

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.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full information factor analysis. *Applied Psychological Measurement, 12*, 261-280.

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

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cattell, R.B. (1956). Validation and intensification of the Sixteen Personality Factor Questionnaire. *Journal of Clinical Psychology, 12*, 105-214.

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.

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.

Michigan Curriculum Review Committee and Michigan Reading Association. (1985). *New Dimensions in Reading: Instruction.*

Mislevy, R.J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Muraki, E., Mislevy, R.J., & Bock, R.D. *PC-Bimain Manual.* (1991). Chicago, IL: Scientific Software, Inc.

National Council of Teachers of Mathematics. *Assessment Standards for School Mathematics.* (1995). Reston, VA: Author.

North Carolina Department of Public Instruction. (1992). *Teacher Handbook for English Language Arts.* Raleigh, NC: Author.

North Carolina Department of Public Instruction. (1992). *Teacher Handbook for Mathematics*. Raleigh, NC: Author.

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.

North Carolina State Board of Education. (1993). *Public School Laws of North Carolina 1994.* Raleigh, NC: The Michie Company.

North Carolina State Board of Education. (March 1996). *Examinig the Structure and Functions of the Public School System of North Carolina*. Raleigh, NC: Author.

Nunnally, J. (1978). *Psychometric Theory.* New York: McGraw-Hill Book Company.

Pommerich, M., Billeaud, K., Williams, V.S.L., & Thissen, D. (1993). *User's Guide for the North Carolina End of Grade Tests*. Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill Book Company.

Sanford, E.E., Ballator, N., Kramer, L., & Morgan, J. (1995) Individual differences in performance on items with similar content and different format. Paper presented at the annual convention of the National Council on Measurement in Education (San Francisco).

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

Schmid, J. & Leiman, J.M. (1957). The devlopment of heirarchical factor solutions. *Psychometrika, 22,* 53-61.

Traub, R.E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage Publciations, Inc.

Wainer, H. and Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Pactice,* 15(1), 22-29.

Williams, V.S.L., Billeaud, K., Davis, L.A., Thissen, D., & Sanford, E.E. (1995). *Projecting to the NAEP scale: Results from the North Carolina End-of-Grade testing program.* Technical Report #34. Research Triangle Park, NC: National Institute of Statistical Sciences.

Williams, V.S.L., Pommerich, M., & Thissen, D. (1996). A comparison of developmental scales based on Thurstone methods and item response theory. Unpublished manuscript.

Appendices

Sample Items and Passages with Curricular and Psychometric Information

4507 Cons Info

Does your bicycle need some repairs? The following sign will tell you how much money you will need for this project.

Bike Repair Shop

Replacements		Repairs	
Tires	$3.95	Tires	$2
Tubes	$1.95	Gear adjustment	$4
Seats	$4.95 – $7.95	Brake repair	$6
Pedals (set)	$3.25 – $4.95	Seat adjustment	$1

40110 How much does it cost to repair one tire?

A $1.50

B $1.95

C $2.00

D $3.95

40823 Which item could be the most to replace?

A pedals

B tubes

C tires

D seats

40230 Which of the following statements is true?

A It costs less to buy a new tire than to buy a new tube.

B It costs less to repair an item than to replace it.

C It costs less to repair brakes than to adjust gears.

D It costs less to buy a seat than to buy a set of pedals.

40310 What would be the cost of one new tire and one new tube?

A $5.90

B $4.95

C $3.95

D $1.95

Reading—Grade 5 Content-Based Passage

To Reach The Promised Land

by Stephen Ray Lilley

Today, public schools in the United States are free and open to everyone. There was a time, however, when going to school was not a simple matter. In the following passage, read about the sacrifices one famous American educator had to make in order to go to school.

Nine-year-old Booker, his sister Amanda, and older brother John stood close to their mother. Excitement filled the air as the Yankee army moved through Virginia in the spring of 1865.

For months Booker had heard his mother praying at night as he drifted off to sleep by the fire: "Lord, let the Yankees win this war, and let them make me and my children free." Now they watched a blue-uniformed soldier standing on the "big house" porch unfold a piece of paper and begin reading.

"All persons held as slaves... henceforward shall be free," he proclaimed.

Life suddenly became very difficult for Booker's family. They had always been owned, like land or livestock. Now free, they had no home, no jobs, no money, only each other. Booker's stepfather worked at a salt furnace near Malden, West Virginia. Putting their belongings in a small cart, the family walked hundreds of miles through the Appalachian Mountains to join him.

In Malden, Booker and John went to work with their stepfather. Work began before daylight and ended after dark. As he shoveled salt into huge wooden barrels, Booker saw children walking to school. "I had the feeling that to get into a schoolhouse and study...would be about the same as getting into paradise," he later said.

But the family needed Booker's income. Booker's stepfather, a tough and practical man, told him attending school was impossible. Knowing how much her son wanted to learn to read, Booker's mother saved every spare penny and bought him a well-used copy of Webster's "Blue-Backed Speller." For weeks he pored over the book, memorizing the alphabet and letter sounds.

Booker convinced his parents he should take lessons at night from a black teacher. Then he told them he wished to attend day school. His stepfather finally accepted the idea, on condition that Booker work at the salt furnace before and after school. Overjoyed, Booker quickly agreed.

Each day Booker faced new obstacles. For a time he worked in a coal mine deep underground in terrifying conditions. Sometimes his candle blew out, and he wandered helplessly in total darkness. Still, he studied at night. Then one day he heard some miners speaking of a school called the Hampton Institute where poor students could work to pay their expenses. "I resolved at once to go to that school, although I had no idea where it was...or how I was going to reach it," he later wrote.

Booker T. Washington became Hampton's most famous graduate and devoted his life to teaching. He taught the first classes at the Tuskegee Institute in Alabama and then built it into one of the most important schools for blacks in the United States. Today, millions of people admire this man who struggled to reach "the promised land."

Released Items

Reading Grade 5 Goal 2

Objective 2.2— The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.

What would be the ***best*** description of Booker T. Washington's attitude toward attending school?

A determined

B hopeless

C practical

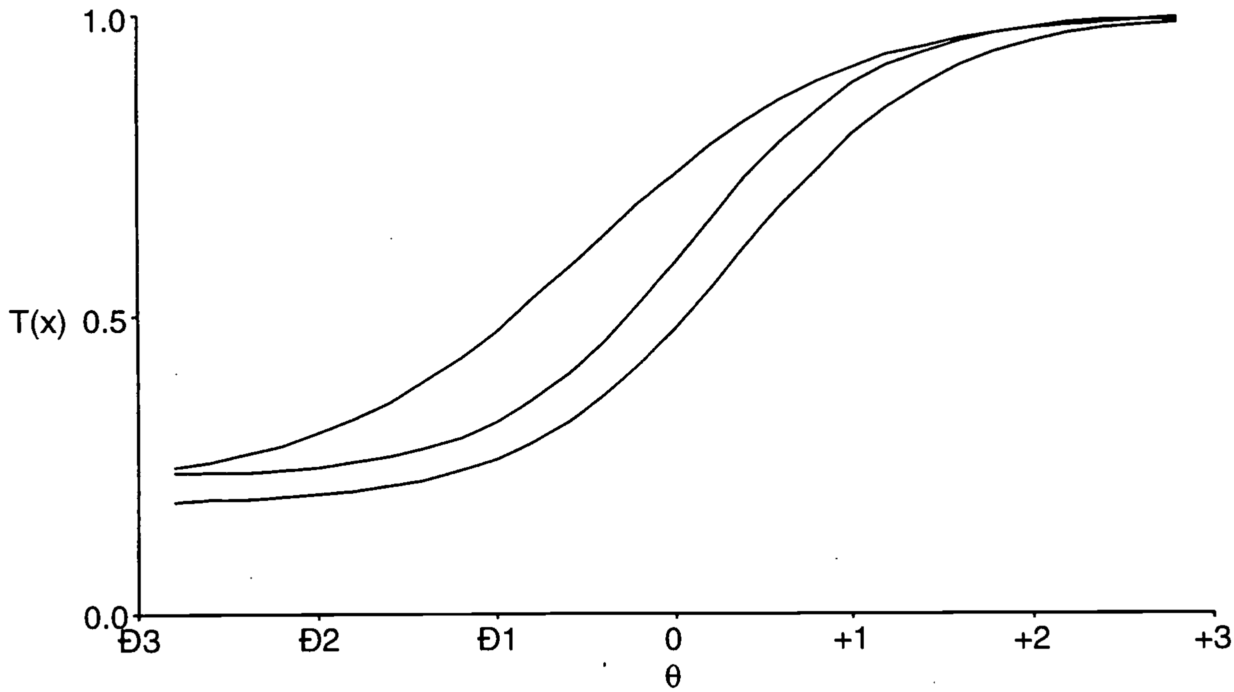
D anxious

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
500R2	7	8	2.2	57R1	1	577	95	122	171

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.60	0.584	.49	0.970	0.904	0.078	1.096	.230

NCTests


1.0
T(x) 0.5
0.0
Đ3
Đ2
Đ1
0
+1
+2
+3
θ

Released Items

end of grade testing

Reading Grade 5 Goal 2

Objective 2.2— The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.

To Booker, what is "the promised land"?

A a faraway country

B a good education

C a well-paying job

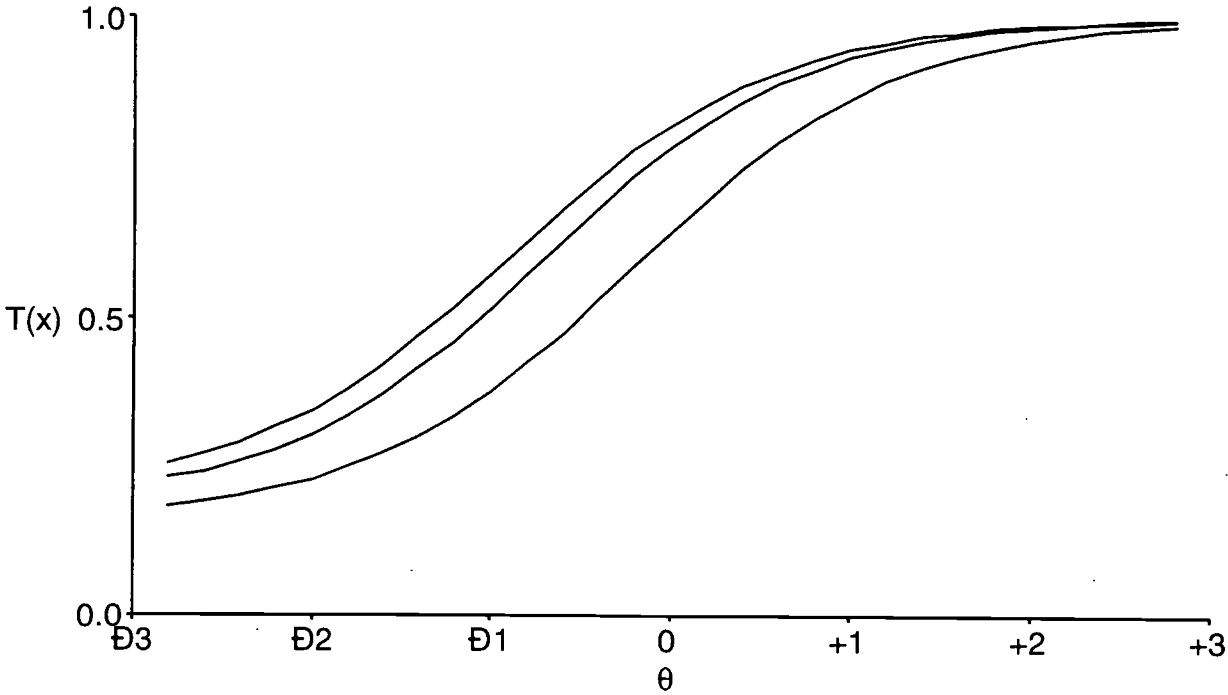
D a guaranteed place

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
51296	7	11	2.2	57R1	2	73	707	60	124

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.73	0.571	.44	0.965	1.021	-0.711	0.826	.188

NCTests


1.0
T(x) 0.5
0.0
Đ3
Đ2
Đ1
0
+1
+2
+3
θ

Released Items

end of grade testing

Reading Grade 5 Goal 3

Objective 3.1— The learner will assess the validity and accuracy of information and ideas.

What would be the ***best*** way to check to see if the information in this passage is accurate?

A Ask a neighbor who lived during that time.

B Read a biography about Booker T. Washington.

C Watch a movie about the Civil War.

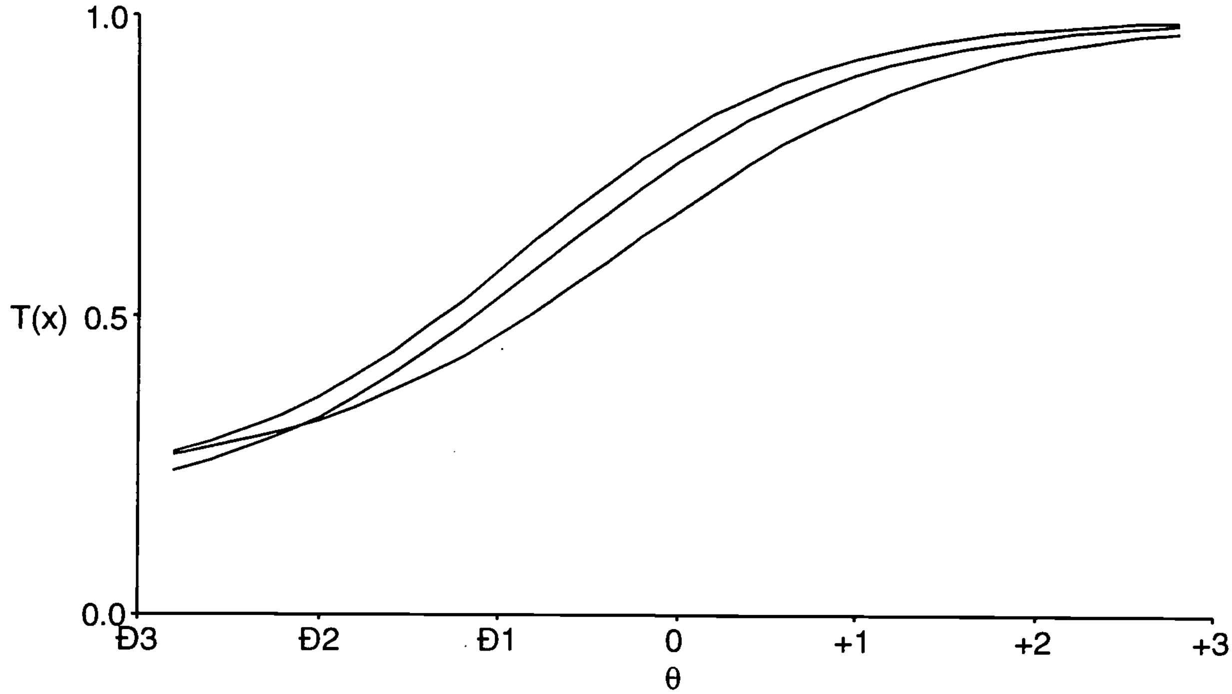
D Reread the passage.

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
500R6	7	14	3.1	57R1	2	42	692	52	178

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.72	0.494	.45	0.945	1.309	-0.769	0.666	.165

NCTests



1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

Reading Grade 5 Goal 3

Objective 3.2— The learner will determine the value of information and ideas.

end of grade testing

What might be the ***best*** reason for recommending this passage to a friend?

A It quotes Booker T. Washington.

B It describes working in a coal mine.

C It sets a good example for other people to follow.

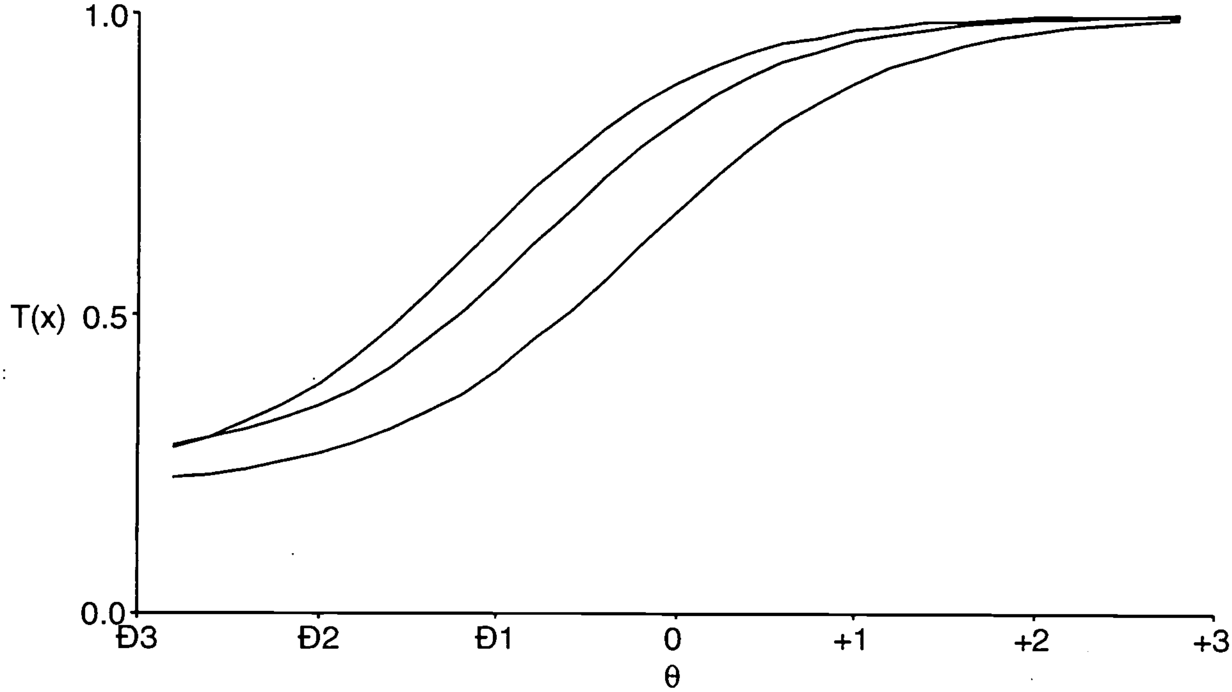
D It describes the southern plantations.

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
500R4	7	15	3.2	57R1	3	119	50	743	52

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Threshold	IRT Slope	IRT Asymptote
.77	0.573	.42	1.219	1.081	-0.759	0.915	.251

NCTests


1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

Reading Grade 5 Goal 2

Objective 2.1— The learner will identify, collect, or select information and ideas.

In the third paragraph, what does "henceforward" mean?

A in front of

B up until now

C from now on

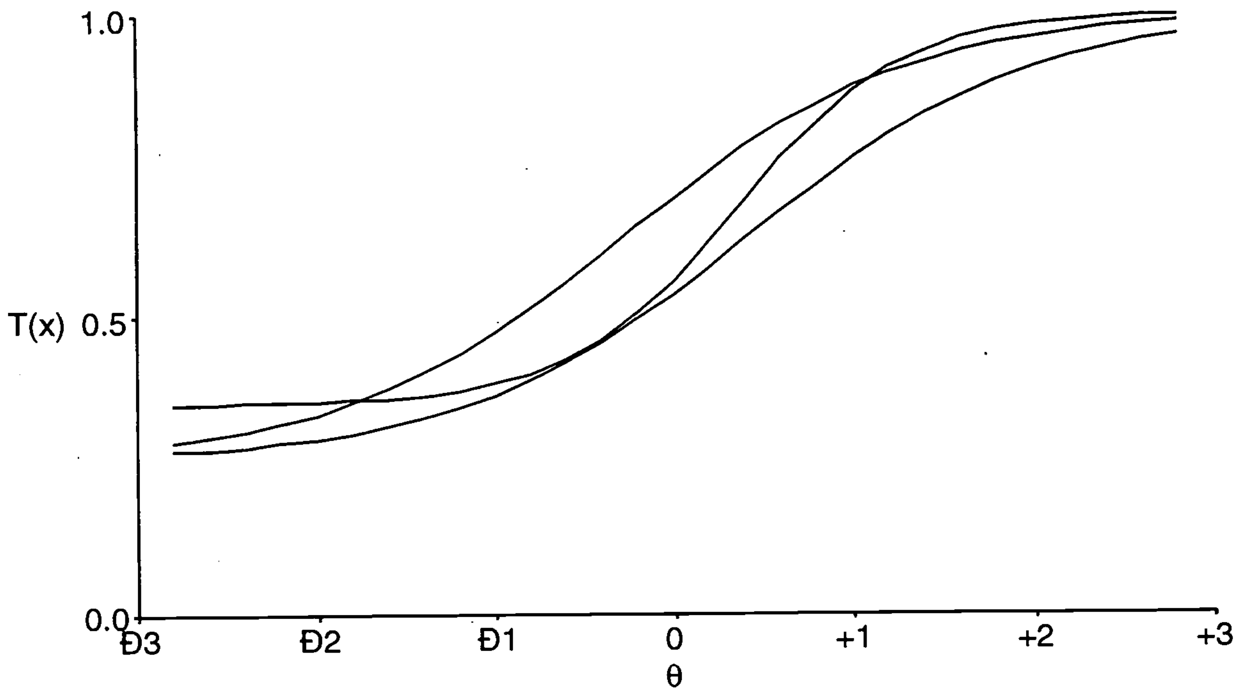
D on the porch

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
51295	7	13	2.1	57R1	3	131	184	587	60

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.61	0.477	.49	0.822	1.119	0.360	1.269	.352

NCTests



1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

end of grade testing

Reading Grade 5 Goal 1

Objective 1.0— The learner will use appropriate preparation strategies to comprehend or convey experiences and information.

After you have read the passage, which of the following is the ***best*** thing to do to help you understand it better?

A Read the passage again to make sure you did not miss any words.

B Retell the main events of the passage to see if you understood it.

C Count how many paragraphs you read with no mistakes.

D Reread the passage aloud.

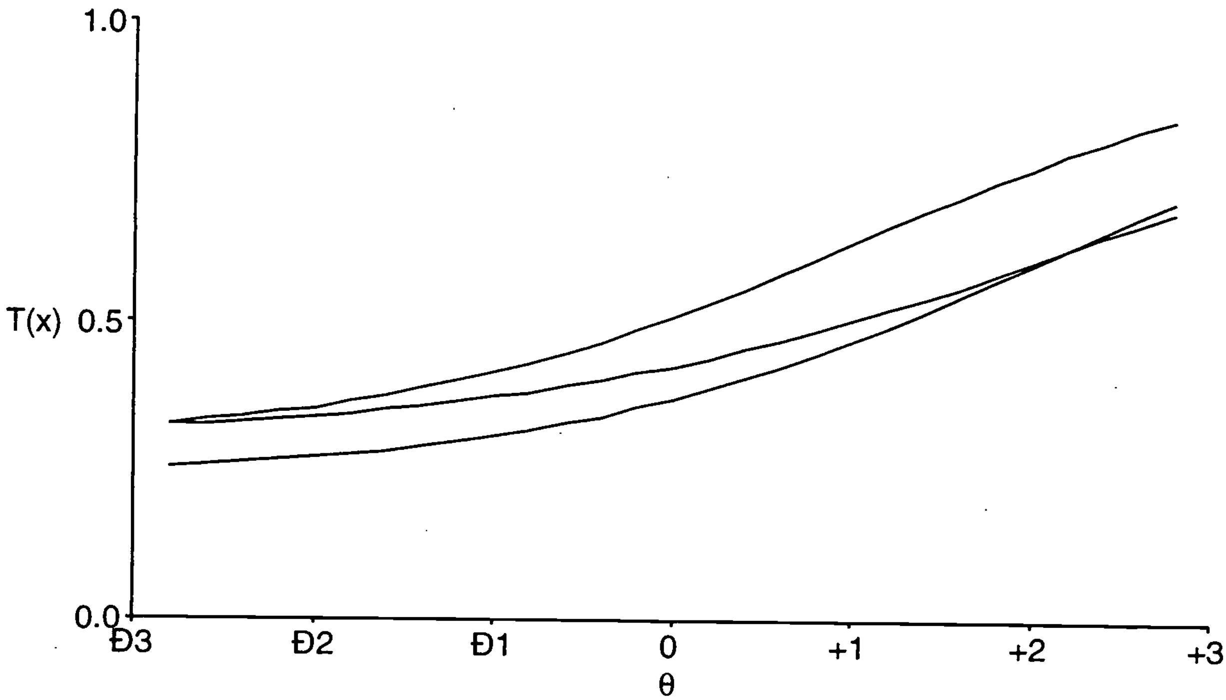
Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
500R9	7	7	1.0	57R1	2	327	417	28	190

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Threshold	IRT Slope	IRT Asymptote
.43	0.114	.50	0.997	0.944	2.452	0.362	.296

NCTests

BEST COPY AVAILABLE

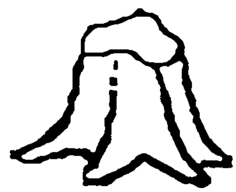

1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

71R2

Read the poem below by Donald Hall and answer the questions that follow.

The Stump

Today they cut down the oak.
Strong men climbed with ropes
in the brittle tree.
The exhaust of a gasoline saw
was blue in the branches.



It is February. The oak has been dead a year.
I remember the great sails of its branches
rolling out greenly, a hundred and twenty feet up,
and acorns thick on the lawn.
Nine cities of squirrels lived in that tree.
Today they run over the snow
squeaking their lamentation.

Yet I was happy that it was coming down
"Let it come down!" I kept saying to myself
with a joy that was strange to me.
Though the oak was the shade of old summers,
I loved the guttural saw.

Released Items

end of grade testing

Reading Grade 7 Goal 3

Objective 3.3— The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.

Which line from the poem helps the reader see the oak swaying in the wind?

A "The exhaust of a a gasoline saw was blue in its branches."

B "I remember the great sails of its branches"

C "Nine cities of squirrels lived in that tree."

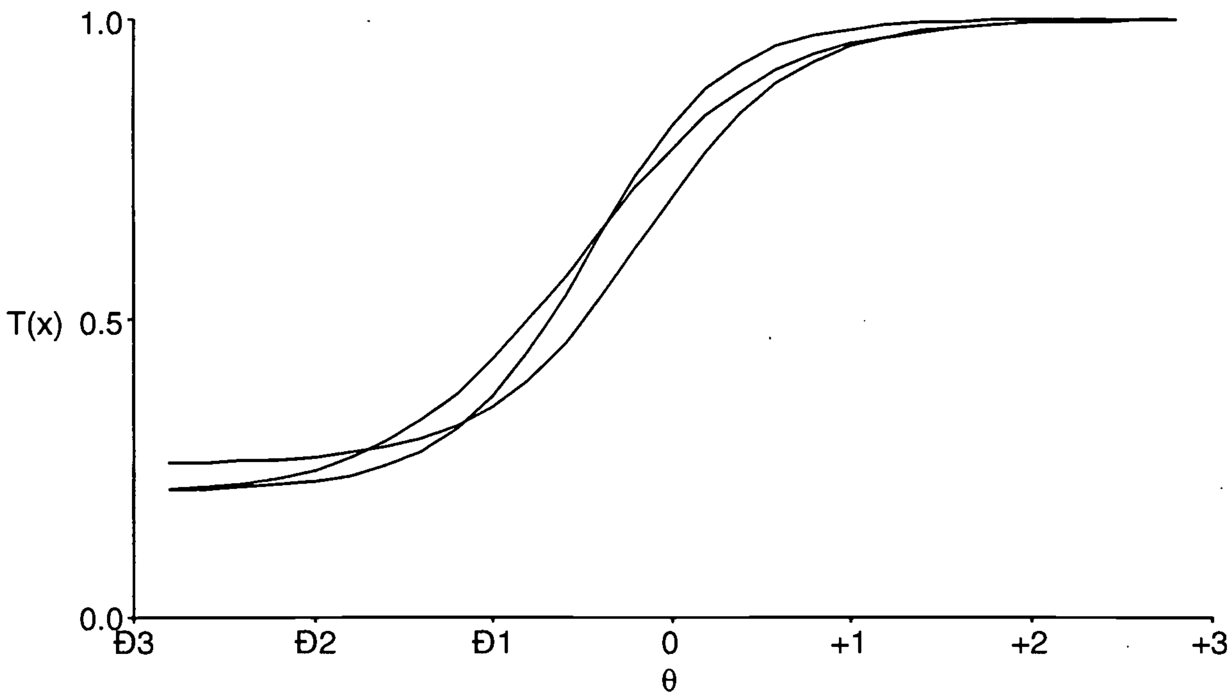
D "Though the oak was the shade of old summers,"

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
7R030	4	60	3.3	71R2	2	109	652	74	60

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.72	0.645	.45	1.110	1.500	-0.520	1.121	.201

NCTests


1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

end of grade testing

Reading Grade 7 Goal 2

Objective 2.2— The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.

Which of the following lines from the poem does ***not*** show an important contrast?

A "Strong men climbed with ropes in the brittle tree."

B "It is February. The oak has been dead a year."

C "Nine cities of squirrels lived in that tree. / Today they run over the snow"

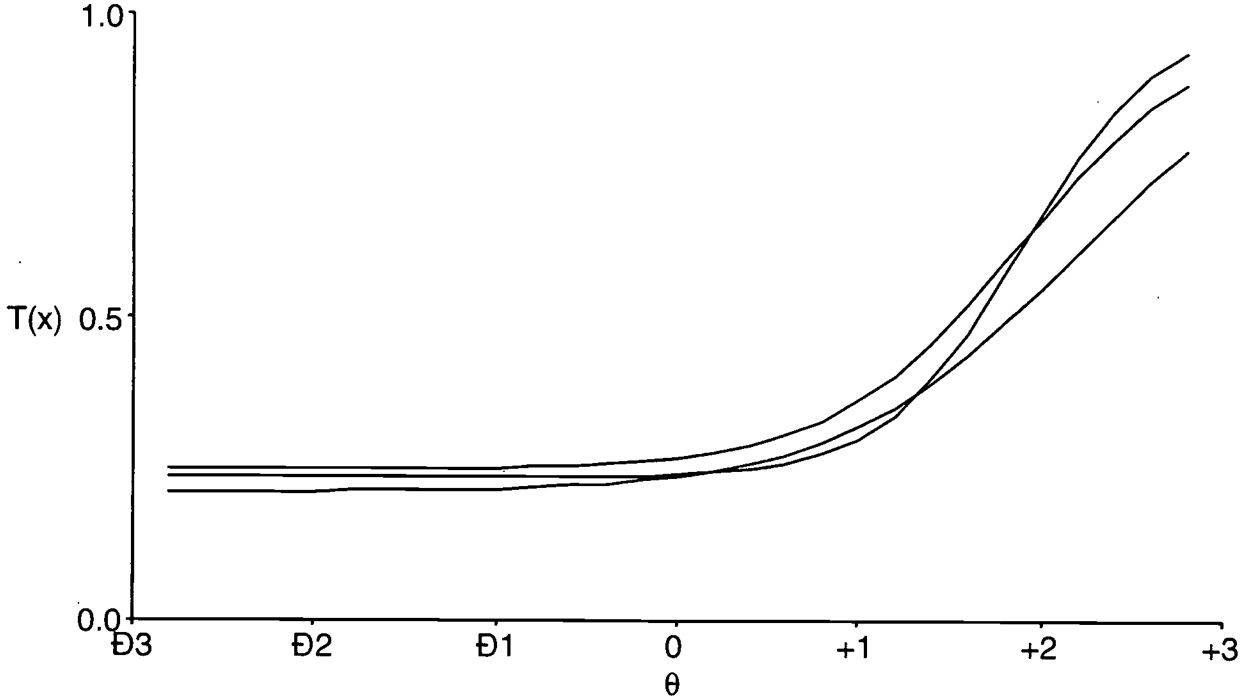
D " 'Let it come down!' I kept saying to myself / with a joy that was strange to me."

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
7R031	4	61	2.2	71R2	2	318	248	190	137

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.28	0.148	.45	0.990	0.982	1.894	1.559	.235

NCTests


1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

end of grade testing

Reading Grade 7 Goal 1

Objective 1.0— The learner will use appropriate preparation strategies to comprehend or convey experiences and information.

If you did not know the word "lamentation" (line 12), what is the ***first*** thing you should do?

A Go back to the beginning of the poem and reread up to line 12.

B Look the word up in the dictionary or ask someone for the meaning.

C Imagine how the squirrels must feel without their tree.

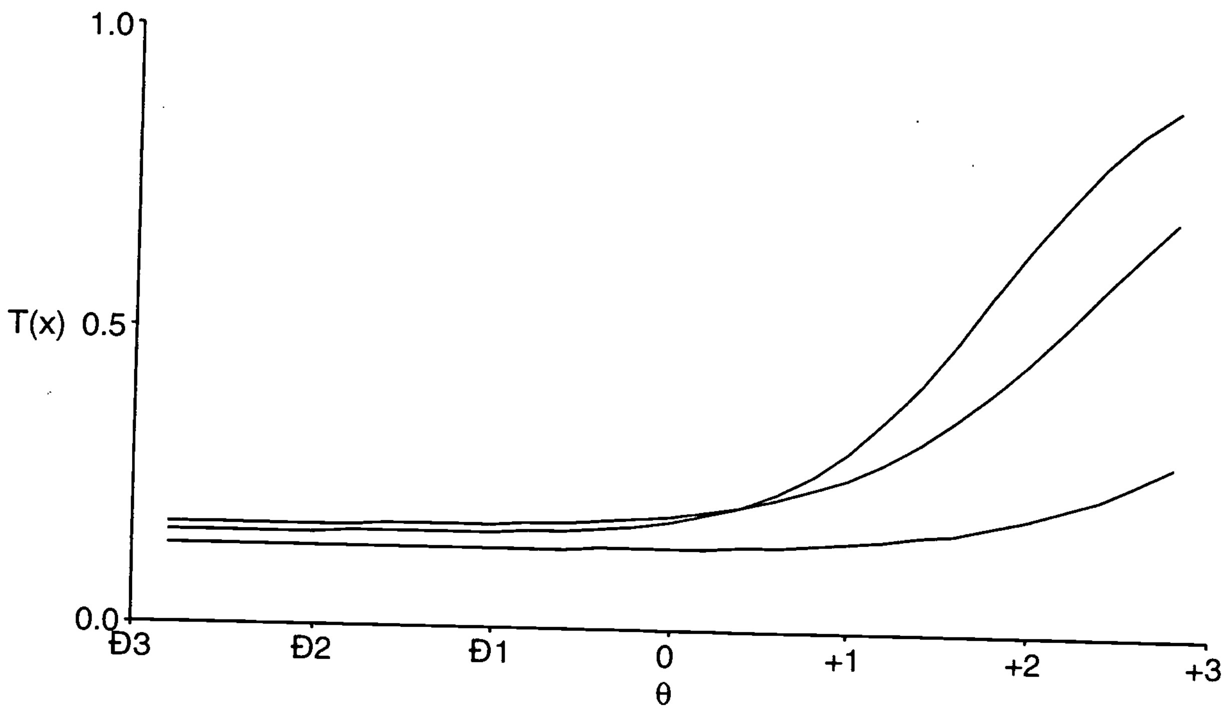
D Examine the word closely, paying attention to its roots, prefixes, and suffixes.

Item Statistics

Origno	Form	Item	Obj	Psg	Key	Choice A	Choice B	Choice C	Choice D
7R029	4	62	1.0	71R2	3	186	311	171	227

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.19	-.008	.39	0.698	0.703	2.428	0.868	.169

NCTests


1.0
T(x) 0.5
0.0
Ð3
Ð2
Ð1
0
+1
+2
+3
θ

Released Items

end of grade testing

Mathematics Grade 3 Goal 4

Objective 4.1— Estimate length and height; measure with appropriate tools using inches, feet, yards, centimeters and meters.

Which of these would be a fairly good estimate for the height of a classroom door?

A 4 feet

B 7 feet

C 25 feet

D 300 feet

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
3R5	7	36	4.1	2	161	580	243	66

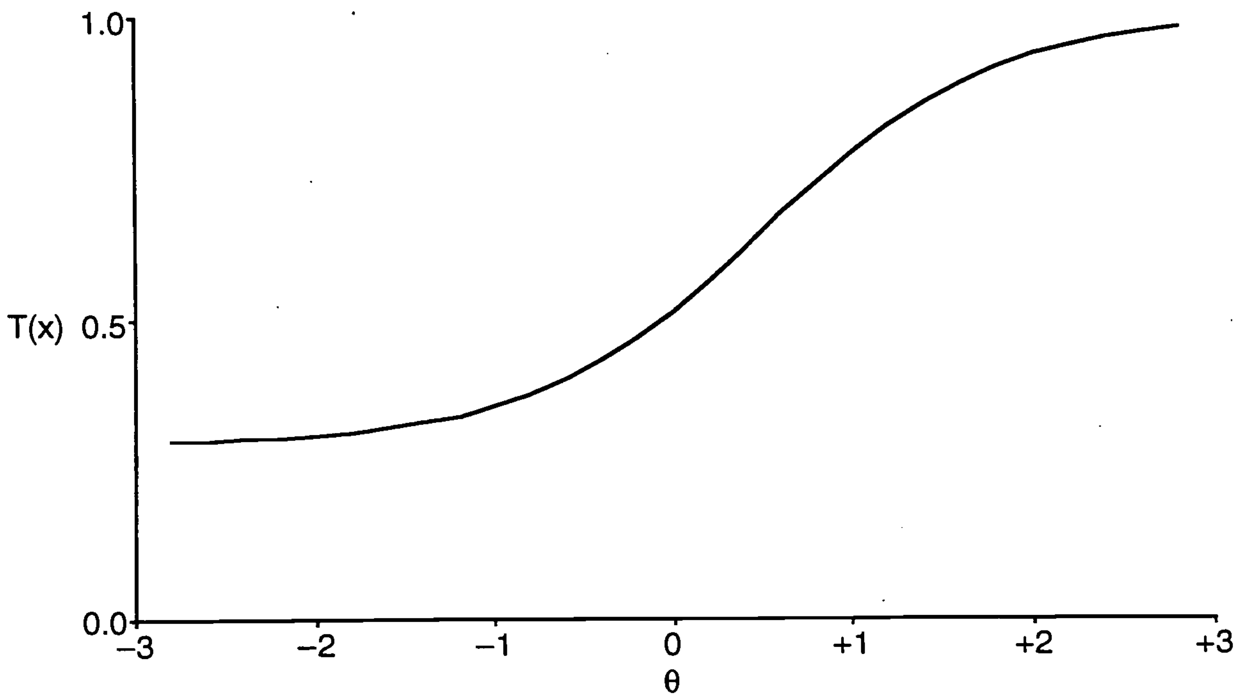
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.55	0.429	.50	1.777	0.686	0.520	0.922	.294

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.285	.434	.637	.872

NCTests

Mathematics- Released Items



Math-Grade 3 3R5 Form 7,Item 36

Released Items

end of grade testing

Mathematics Grade 3 Goal 5

Objective 5.1— Identify and describe problems in given situations.

Shira and Donna ride their bikes to school. They can ride their bikes 5 miles in 1 hour. What other information is needed to determine how long it takes to get to school?

A the name of the school

B the time they left home

C the kind of bikes they were riding

D the distance to the school

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
3R7	4	49	5.1	4	49	271	102	648

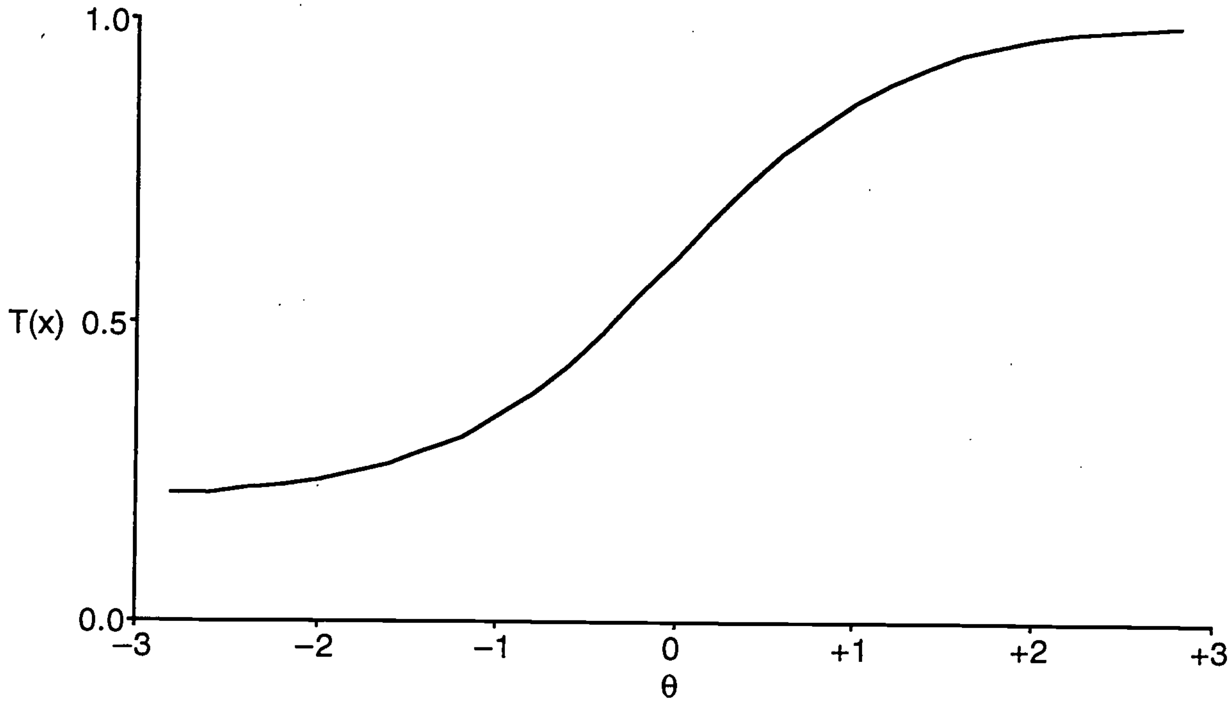
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Threshold	IRT Slope	IRT Asymptote
.60	0.542	.49	1.216	1.089	-0.005	0.924	.202

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.263	.441	.748	.902

NCTests

Mathematics- Released Items



Math-Grade 3 3R7 Form 4, Item 49

Released Items

Mathematics Grade 3 Goal 5

Objective 5.2 —Develop stories to illustrate problem situations and number sentences.

end of grade testing

Which of the following stories explains the problem $5 \times 2 = 10$?

A Betty is going to bake 5 batches of chocolate chip cookies. Her recipe uses 2 eggs. How many eggs will Betty need?

B There are 5 boys in the club. There are 2 girls in the club. How many students are there in the club?

C The snack bar has 5 candy bars. Mary and Joey each buy a candy bar. How many candy bars are left?

D Sally has 5 cupcakes. She wants to give half of them to her friend Suzy. How many cupcakes will Suzy have?

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
3R2	4	52	5.2	1	558	241	143	123

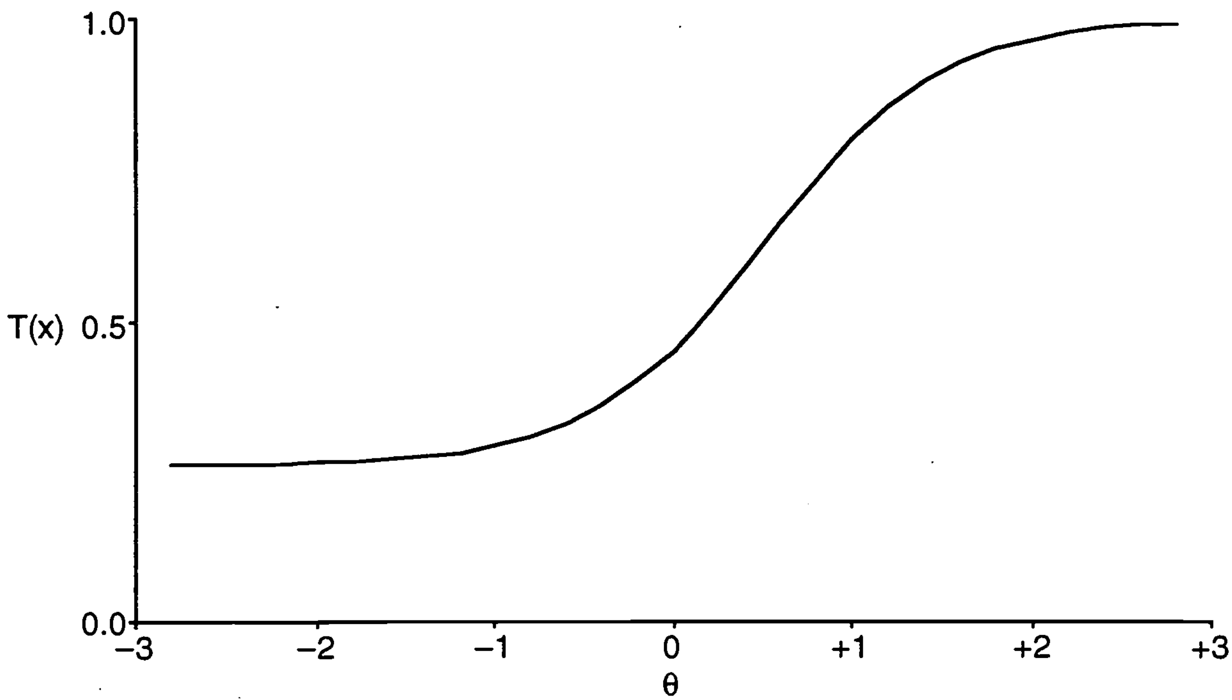
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Threshold	IRT Slope	IRT Asymptote
.52	0.492	.50	0.790	0.971	0.519	1.211	.261

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.254	.358	.616	.910

NCTests

Mathematics- Released Items



Math-Grade 3 3R2 Form 4,Item 52

Released Items

Mathematics Grade 3 Goal 6

Objective 6.6 —Locate points on a coordinate grid; name with ordered pairs.

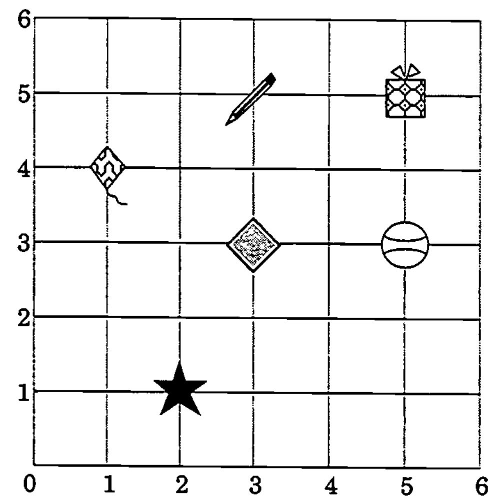
The pencil is found at which ordered pair?

A (3, 3)

B (3, 5)

C (5, 3)

D (5, 5)



Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
3R3	1	62	6.6	2	107	578	277	76

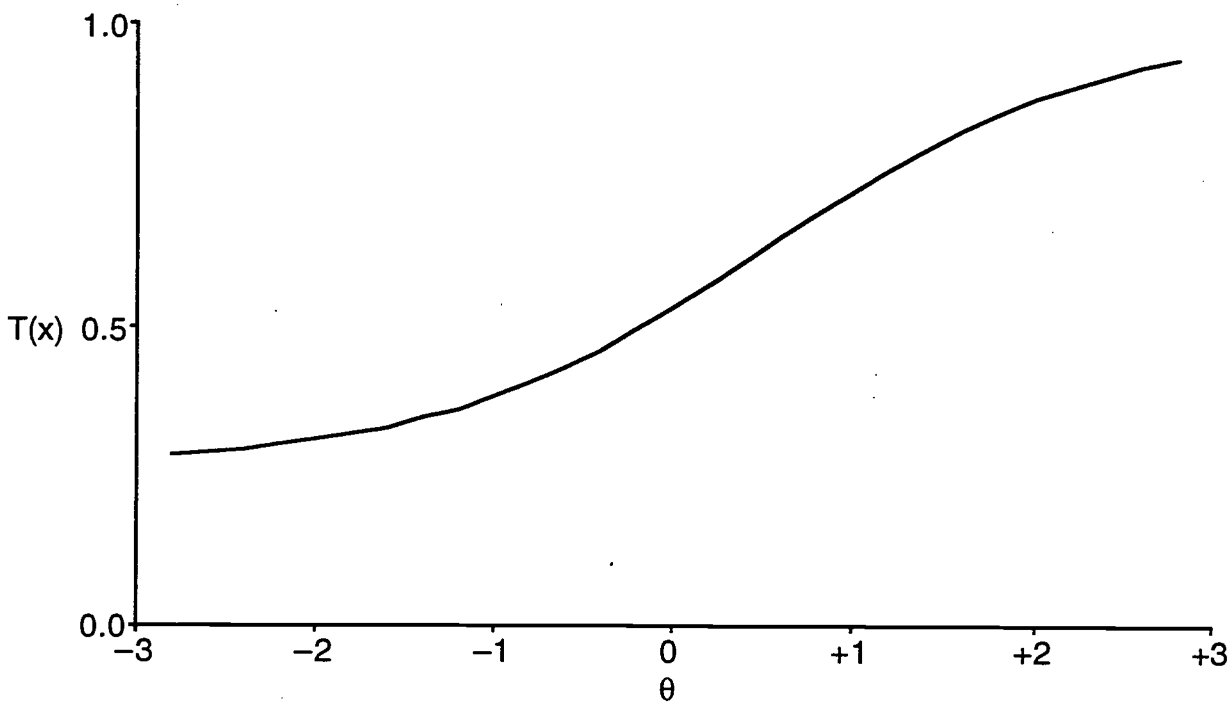
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.55	0.375	.50	0.977	1.274	0.549	0.638	.266

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.266	.484	.647	.908

NCTests

Mathematics- Released Items



Math-Grade 3 3R3 Form 1,Item 62

Released Items

end of grade testing

Mathematics Grade 6 Goal 1

Objective 1.5— Use prime factorization to investigate common factors and common multiples using a calculator when appropriate.

There are 50 people in a 10k roadrace. Every 6th finisher in the race receives a T-shirt. Every 8th finisher in the race receives a hat. How many people will receive a T-shirt *and* a hat?

A 2

B 6

C 8

D 10

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
6R2	3	18	1.5	1	357	148	192	224

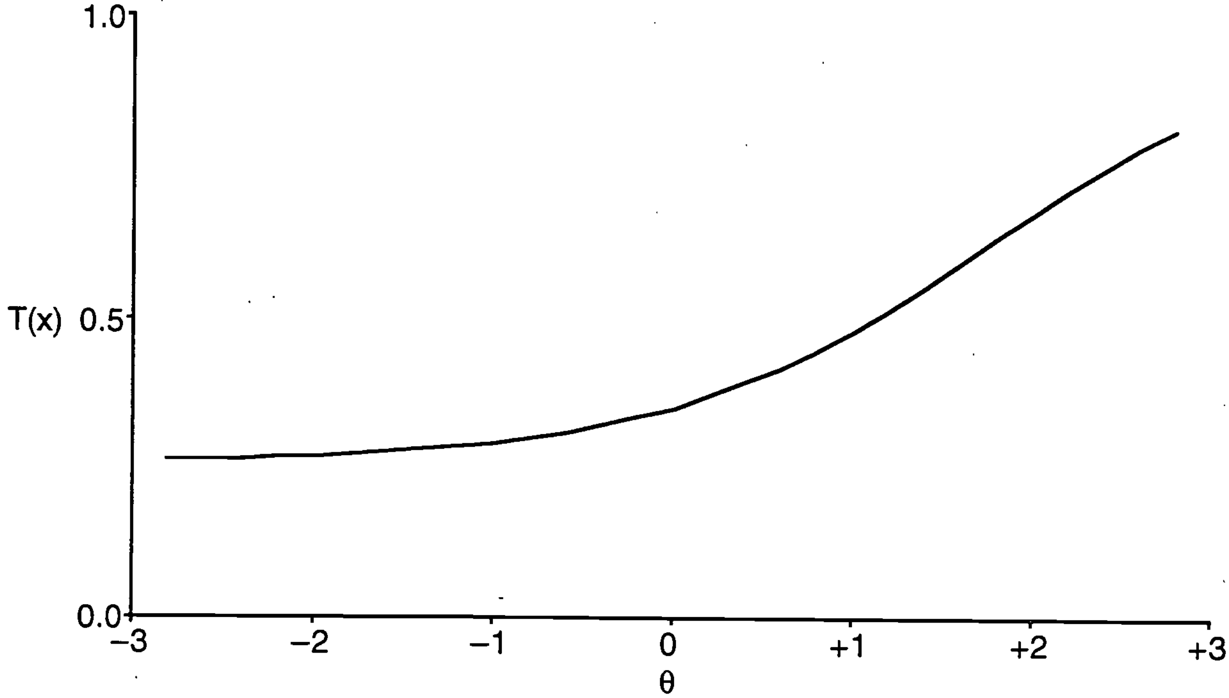
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.38	0.251	.49	0.959	1.049	1.790	0.643	.256

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.199	.328	.426	.604

NCTests

Mathematics- Released Items



Math-Grade 6 6R2 Form 3,Item 18

Released Items

end of grade testing

Mathematics Grade 6 Goal 6

Objective 6.2—Use measures of central tendency (mean, median, and mode) and range to describe meaningful data; compare two sets of unequal data.

Mrs. Larkin asked her students the following question:

If each number in a list is increased by 4, how does the mean of the new list compare with the mean of the old list?

Andy said, "The mean of the new list will be four times the mean of the old list."

Betty said, "The mean of the new list will be four points higher than the mean of the old list."

Carl said, "The mean of the new list will be four points lower than the mean of the old list."

Denise said, "There is no way to find out what the mean of the new list would be."

Which student answered correctly?

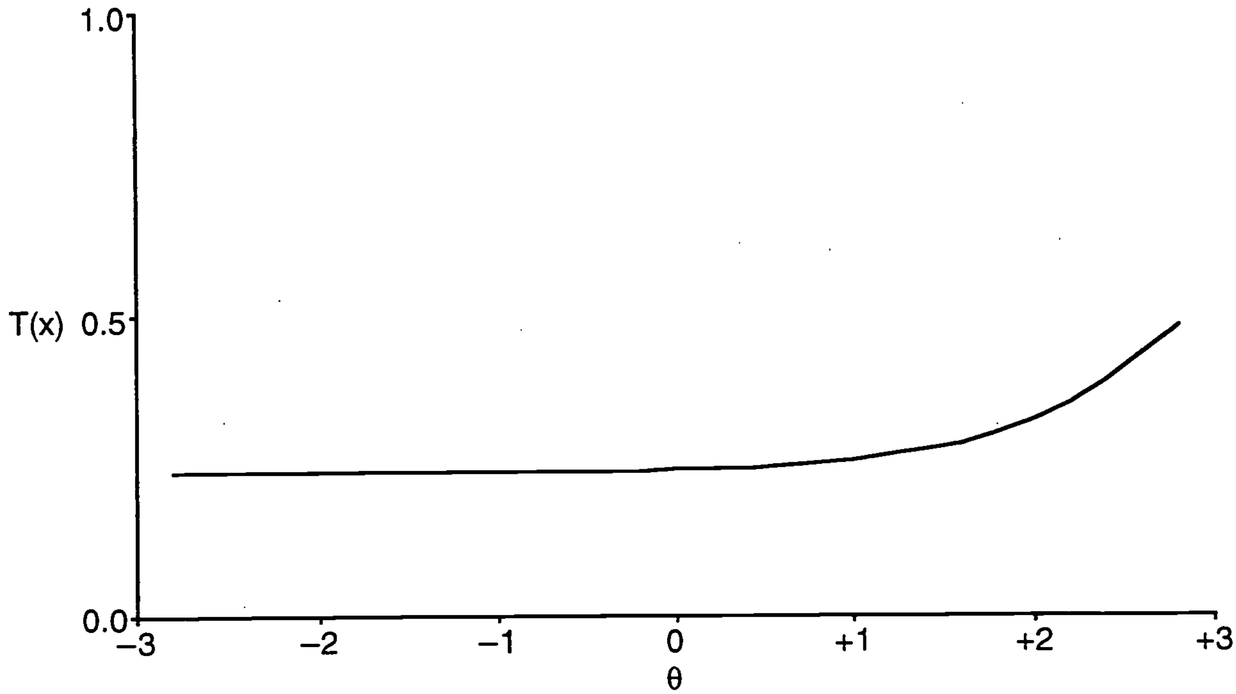
A Andy

B Betty

C Carl

D Denise

NCTests



Math-Grade 6 6R1 Form 10,Item 61

Released Items

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
6R1	10	61	6.2	2	284	227	126	236

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.26	0.012	.44	0.773	0.834	3.277	0.950	.239

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.206	.283	.242	.305

Released Items

Mathematics Grade 7 Goal 4

Objective 4.6— Estimate answers; solve problems related to volume.

end of grade testing

One way to earn money during the summer is to grow and sell vegetables. One person can easily take care of a vegetable bed that is six feet by eight feet. If the bed needs to be six inches deep, how much topsoil will be needed to fill the bed?

A 24.0 cubic feet

B 28.8 cubic feet

C 48.0 cubic feet

D 288 cubic feet

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
7R2	1	44	4.6	1	223	208	315	167

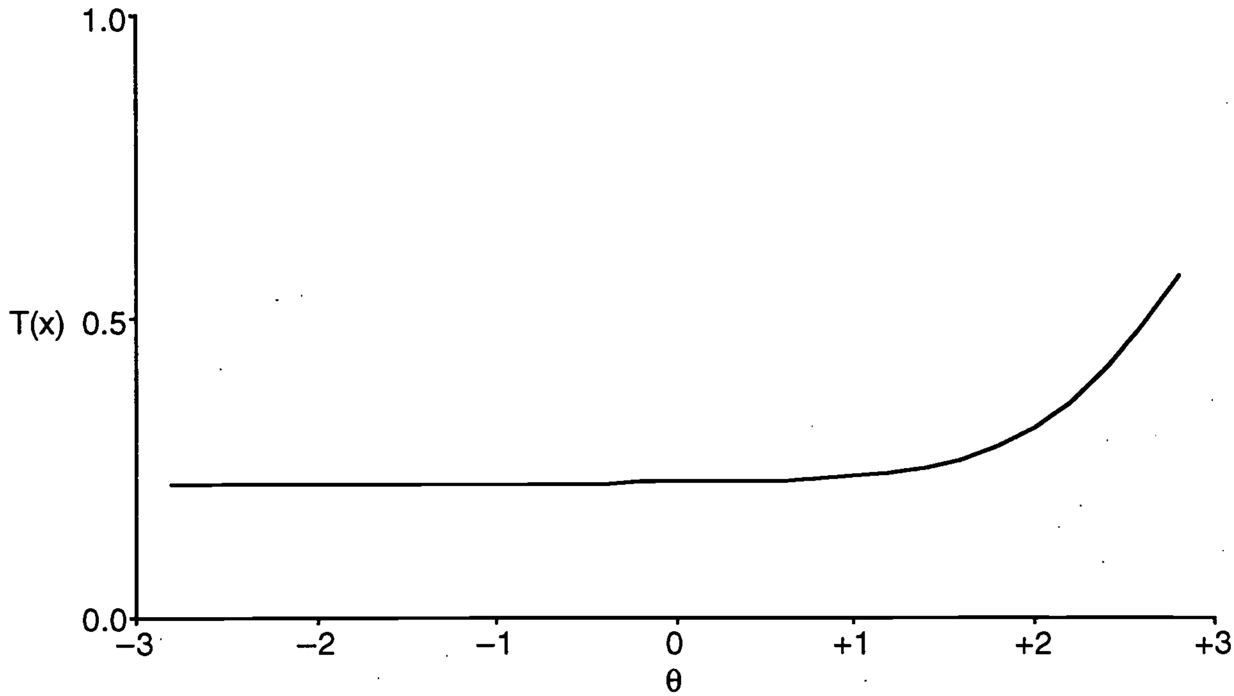
P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.24	-.019	.43	1.954	0.998	2.896	1.334	.224

Achievement Levels

Percent Correct	Level 1	Level 2	Level 3	Level 4
	.194	.278	.217	.238

NCTests

Mathematics- Released Items



Math-Grade 7 7R2 Form 1,Item 44

Released Items

Mathematics Grade 8 Goal 3

Objective 3.4— Using patterns and algebraic methods, solve problems, including those with integers.

Anne sold 5 pairs of earrings she had made. After deducting the $6 she had spent on materials, Anne donated her $9 profit. If each pair cost the same amount, how much did she charge for each pair of earrings?

A $.60

B $3.00

C $4.00

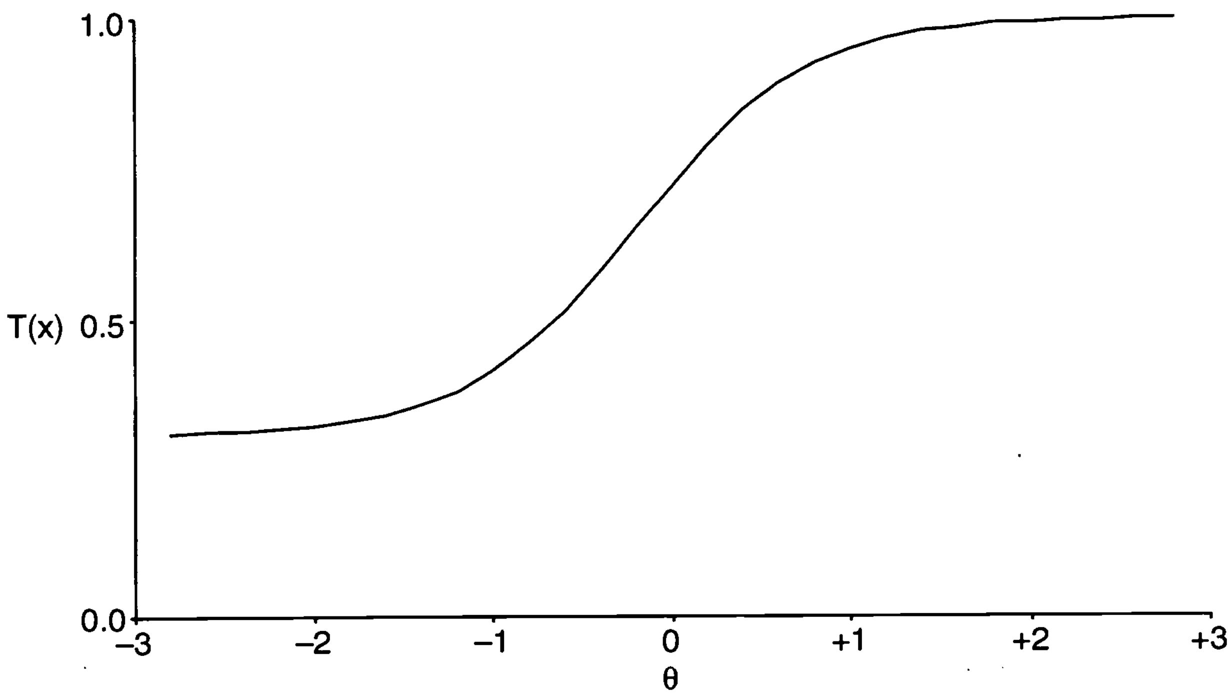
D $4.50

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
80270	9	31	3.4	2	118	641	83	68

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.70	0.544	.46	1.001	2.034	-0.191	1.239	.305

NCTests

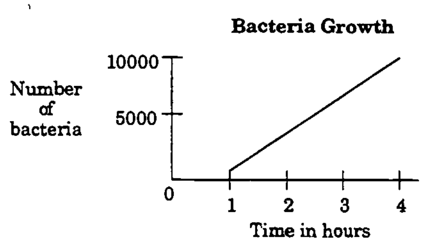

1.0
T(x) 0.5
0.0
−3
−2
−1
0
+1
+2
+3
θ

You are growing bacteria in a test tube to determine the rate of growth. Every hour you remove a sample and count the number of bacteria. You obtain the following information:

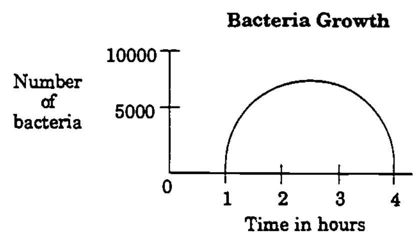
Hour	Number of Bacteria
1	10
2	100
3	1000
4	10000

Which of the following graphs accurately shows the growth of the bacteria?

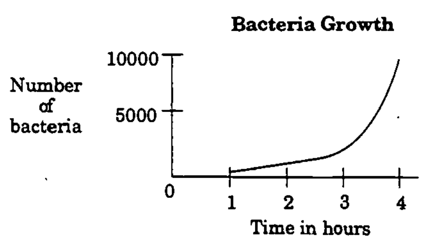
A



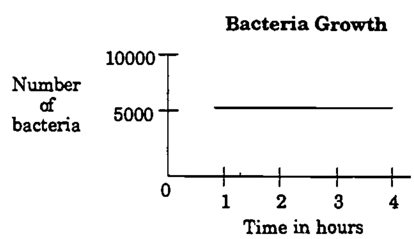
B



C



D



Released Items

Mathematics Grade 8 Goal 3

Objective 3.6— Investigate non-linear equations and inequalities informally.

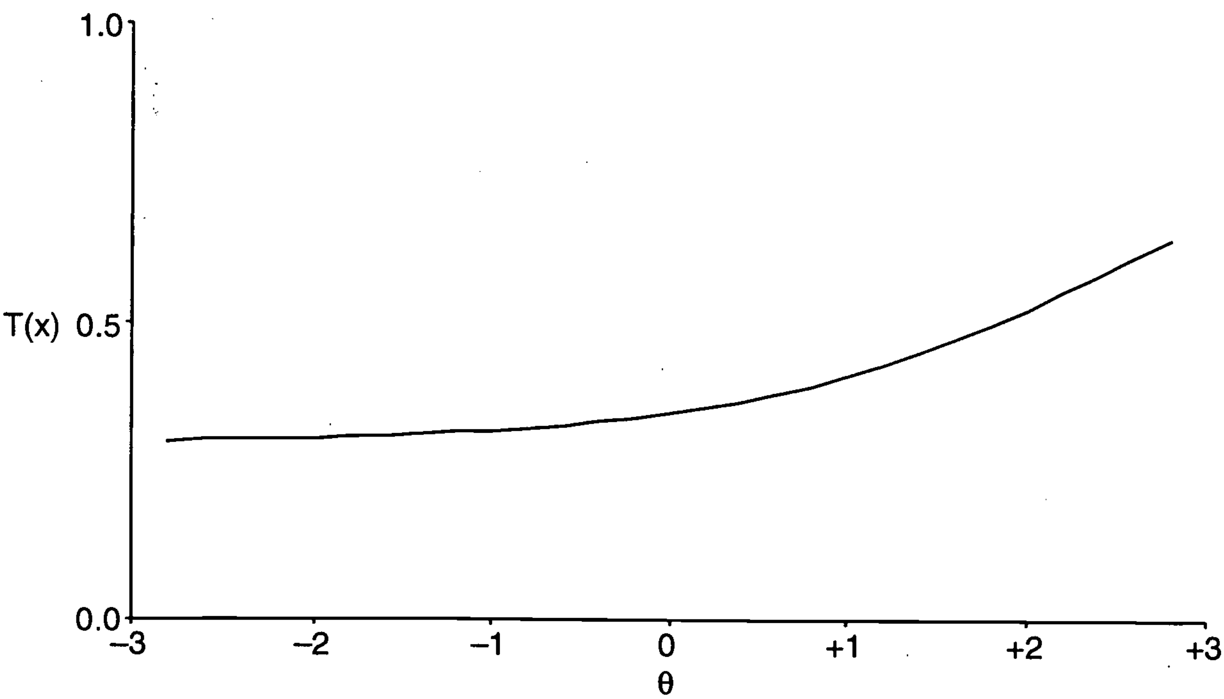
end of grade testing

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
80311	2	35	3.6	3	470	83	347	50

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.36	0.124	.48	1.059	0.713	2.837	0.527	.295

NCTests



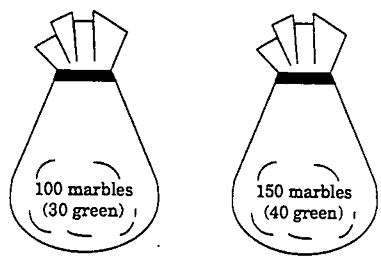
Released Items

end of grade testing

Mathematics Grade 8 Goal 6

Objective 6.6— Find the probability of simple and compound events using experiments, computer simulations, random number generation, and theoretical methods.

What is the probability of reaching into a bag without looking and pulling out a green marble?



BAG 1 BAG 2

A greater for Bag 1 than Bag 2

B greater for Bag 2 than Bag 1

C the same for both bags

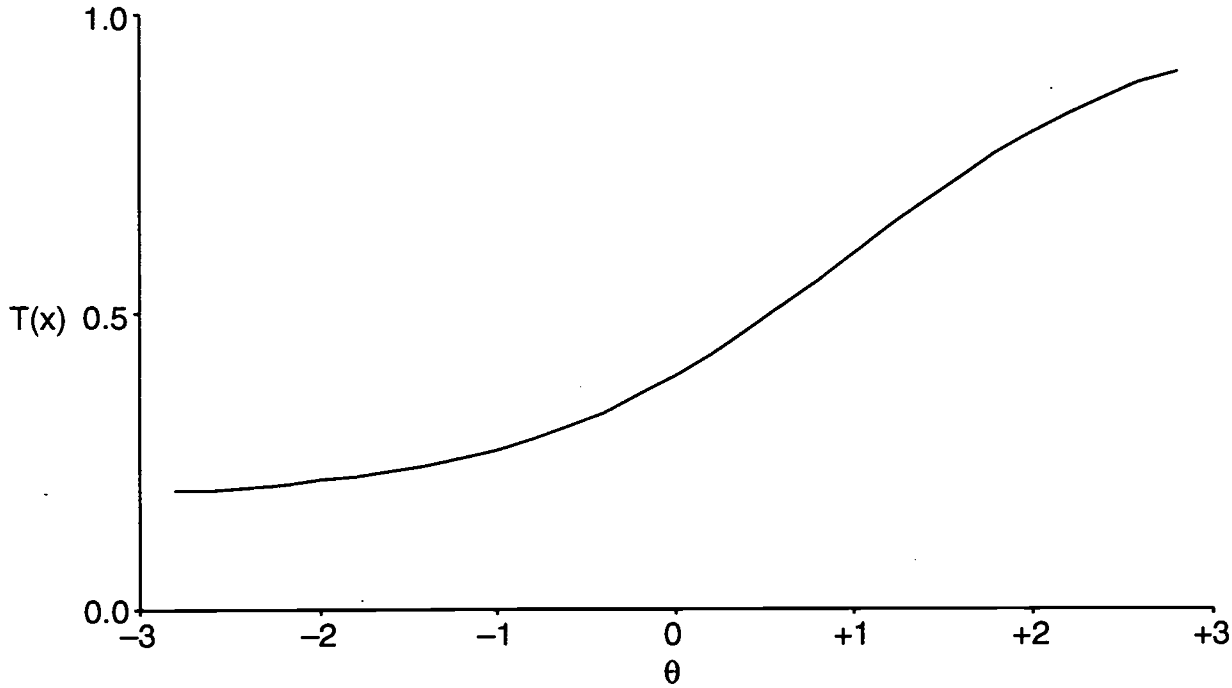
D cannot be determined from the information given

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
80671	10	66	6.6	1	374	249	141	105

P	Bis	Psd	Bias Ethnic	Bias Gender	IRT Parameters Threshold	IRT Parameters Slope	IRT Parameters Asymptote
.42	0.383	.49	1.504	0.842	0.991	0.651	.187

NCTests


1.0
T(x) 0.5
0.0
−3
−2
−1
0
+1
+2
+3
θ

Released Items

Mathematics Grade 8 Goal 7

Objective 7.2— In meaningful contexts, develop the laws of exponents; solve problems involving exponentiation.

end of grade testing

Light travels at 186,000 miles per second, or $1.86 \times 100,000$ miles per second. If this were expressed as 1.86×10^x, what would be the value of x?

A 6

B 5

C 4

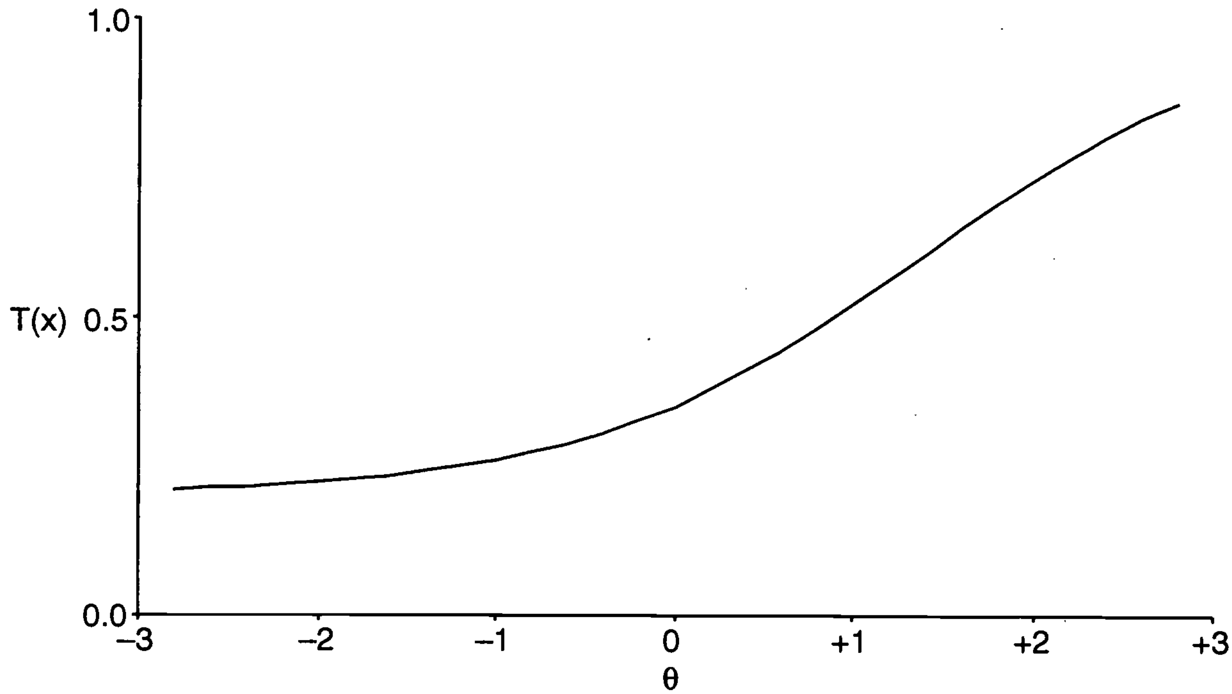
D 3

Item Statistics

Origno	Form	Item	Obj	Key	Choice A	Choice B	Choice C	Choice D
80741	7	79	7.2	2	166	357	254	122

P	Bis	Psd	Bias: Ethnic	Bias: Gender	IRT Parameters: Threshold	IRT Parameters: Slope	IRT Parameters: Asymptote
.39	0.331	.49	1.005	0.884	1.386	0.634	.202

NCTests



1.0
T(x) 0.5
0.0
−3
−2
−1
0
+1
+2
+3
θ

Reading and Mathematics Curricula, Test Specifications, and Average Difficulty of Item Pools

Notes:

Number of Items per Form: This is the typical number of items per form. Some forms may have one more or one less because the objective may be slightly more important than another objective, therefore additional items are needed for curriculum coverage.

Number of Items per Class: This is the number of items that are administered in each class in order to evaluate the implementation of the curriculum.

Difficulty of Pool: This is the average percent correct across all items that measure the goal or objective. (NT = Not Tested)

Reading—Grade 3

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
2.0	The learner will use language for the acquisition, interpretation, and application of information.	44	130-135	0.622
2.1	The learner will identify, collect, or select information and ideas.	18	52-58	0.667
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	18	48-59	0.582
2.3	The learner will apply, extend, and expand on information and concepts.	8	24	0.609
3.0	The learner will use language for critical analysis and evaluation.	12	33-38	0.498
3.1	The learner will assess the validity and accuracy of information and ideas.			0.506
3.2	The learner will determine the value of information and ideas.			0.514
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.482

Reading—Grade 4

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
2.0	The learner will use language for the acquisition, interpretation, and application of information.	50	150	0.599
2.1	The learner will identify, collect, or select information and ideas.	23	67-72	0.631
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	18	52-53	0.562
2.3	The learner will apply, extend, and expand on information and concepts.	9	26-30	0.588
3.0	The learner will use language for critical analysis and evaluation.	15	45	0.534
3.1	The learner will assess the validity and accuracy of information and ideas.			0.522
3.2	The learner will determine the value of information and ideas.			0.560
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.516

Reading—Grade 5

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use strategies and processes that enhance control of communication skills development.	4	12-14	0.546
2.0	The learner will use language for the acquisition, interpretation, and application of information.	44	133	0.598
2.1	The learner will identify, collect, or select information and ideas.	19	58-59	0.634
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	19	56-57	0.567
2.3	The learner will apply, extend, and expand on information and concepts.	6	18	0.571
3.0	The learner will use language for critical analysis and evaluation.	16	48-50	0.522
3.1	The learner will assess the validity and accuracy of information and ideas.			0.673
3.2	The learner will determine the value of information and ideas.			0.541
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.488

Reading—Grade 6

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use strategies and processes that enhance control of communication skills development.	4	8-14	0.554
2.0	The learner will use language for the acquisition, interpretation, and application of information.	48	143	0.594
2.1	The learner will identify, collect, or select information and ideas.	19	58	0.622
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	22	63-72	0.569
2.3	The learner will apply, extend, and expand on information and concepts.	6	13-22	0.589
3.0	The learner will use language for critical analysis and evaluation.	14	38-44	0.523
3.1	The learner will assess the validity and accuracy of information and ideas.			0.490
3.2	The learner will determine the value of information and ideas.			0.572
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.504

Reading—Grade 7

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use strategies and processes that enhance control of communication skills development.	4	9-17	0.538
2.0	The learner will use language for the acquisition, interpretation, and application of information.	47	139-145	0.579
2.1	The learner will identify, collect, or select information and ideas.	16	47-48	0.637
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	26	73-80	0.547
2.3	The learner will apply, extend, and expand on information and concepts.	6	17-19	0.563
3.0	The learner will use language for critical analysis and evaluation.	14	42-44	0.549
3.1	The learner will assess the validity and accuracy of information and ideas.			0.566
3.2	The learner will determine the value of information and ideas.			0.602
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.508

Reading—Grade 8

Goal/ Objective	Description of Goal/Objective	Avg # Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will use strategies and processes that enhance control of communication skills development.	4	9-15	0.519
2.0	The learner will use language for the acquisition, interpretation, and application of information.	52	153-157	0.603
2.1	The learner will identify, collect, or select information and ideas.	16	46-49	0.642
2.2	The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.	31	92-94	0.584
2.3	The learner will apply, extend, and expand on information and concepts.	5	14-15	0.585
3.0	The learner will use language for critical analysis and evaluation.	12	32-42	0.582
3.1	The learner will assess the validity and accuracy of information and ideas.			0.491
3.2	The learner will determine the value of information and ideas.			0.669
3.3	The learner will develop criteria and evaluate the quality, relevance, and importance of the information and ideas.			0.555

Mathematics—Grade 3

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will identify and use numbers to 1,000 and beyond.	8	24	0.644
1.1	Group objects/model 3-digit numbers; relate models to standard and expanded notations.	1	4	0.735
1.2	Compare and order numbers less than 1000.	1	3	0.683
1.3	Read, write, and use whole numbers appropriately in a variety of ways.	1	4	0.536
1.4	Estimate; approximate multiples of 10 or 100.	1	3	0.538
1.5	Model odd and even numbers; generalize ways to determine odd or even.	1	3	0.596
1.6	Model fractions and mixed numbers; describe relationships of parts to whole.	NT	NT	NT
1.7	Relate fractions and mixed numbers to models and pictures for both regions and sets.	1	4	0.555
1.8	Compare fraction models; describe comparisons and explain different names for the same fractional parts.	1	3	0.714
2.0	The learner will demonstrate an understanding and use of geometry.	8	24	0.674
2.1	Classify plane and solid figures; describe rules for grouping.	1	3	0.657
2.2	Construct with cubes a solid to match a given model or picture.	1	3	0.328
2.3	Describe a 3-dimensional object from different perspectives.	1	3	0.559
2.4	Identify and model symmetry with concrete materials, drawings, and computer graphics.	2	6	0.696
2.5	Investigate congruence with concrete materials, drawings, and computer graphics.	2	6	0.751
2.6	Observe and describe geometry in the environment.	1	3	0.708
3.0	The learner will demonstrate an understanding of classification, pattern, and seriation.	8	24	0.619
3.1	Organize objects or ideas into groups; describe attributes of groups and rules for sorting.	1	3	0.414

Mathematics—Grade 3 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
3.2	Describe (demonstrate) patterns in skip counting and multiplication; continue sequences beyond memorized/modeled numbers.	2	6	0.662
3.3	Extend/create geometric and numerical sequences; describe patterns.	2	6	0.630
3.4	Observe/analyze patterns; describe pattern properties and given examples of similar patterns in varied forms.	1	3	0.556
3.5	Use patterns to make predictions and solve problems.	1	3	0.519
3.6	Use understanding of seriation in real life situations.	1	3	0.568
3.7	Explore number patterns with calculators.	NT	NT	NT
4.0	The learner will understand and use standard units of metric and customary measure.	12	36	0.625
4.1	Estimate length and height; measure with appropriate tools using inches, feet, yards, centimeters and meters.	1	4	0.589
4.2	Estimate weight in ounces, pounds, grams and kilograms; measure and describe results.	1	2	0.533
4.3	Estimate capacity; measure with appropriate units (teaspoons, tablespoons, cups, pints, quarts, liters).	1	1	0.478
4.4	Tell/write time to nearest minute with digital and traditional clocks.	1	4	0.692
4.5	Use calendar and appropriate vocabulary to describe time and to solve problems.	1	4	0.662
4.6	Read Celsius and Fahrenheit thermometers; relate temperatures to everyday situations.	1	4	0.648
4.7	Model/compare units within the same measurement system.	1	3	0.521
4.8	Evaluate sets of coins; create equivalent amounts with different coins.	1	4	0.621
4.9	Estimate costs of items; identify coins/bills for purchase; make change less than $5.00.	1	4	0.573
4.10	Read/write given amounts of money in decimal form up to $5.00.	1	3	0.736

Mathematics—Grade 3 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.11	Explore concept of area by covering figures with concrete materials; describe results of experiments.	NT	NT	NT
4.12	Explore concept of perimeter with nonstandard and standard units; explain results.	NT	NT	NT
4.13	Estimate results; solve non-routine and real life problems using measurement concepts and procedures.	1	3	0.474
5.0	The learner will use mathematics reasoning and solve problems.	12	36	0.512
5.1	Identify and describe problems in given situations.	3	9	0.533
5.2	Develop stories to illustrate problem situations and number sentences.	3	9	0.547
5.3	Solve routine and non-routine problems using a variety of strategies, such as use models and "act out", use drawings, diagrams, and organized lists, use spatial visualization, logical thinking, estimation, guess and check and patterns.	3	10	0.435
5.4	Explore different methods of solving problems, including using manipulatives, pencil and paper, mental computation, calculators, and computers.	NT	NT	NT
5.5	Describe processes used in finding solutions; suggest alternate strategies/methods.	1	5	0.484
5.6	Discuss reasonableness of solutions and completeness of answers.	1	3	0.366
6.0	The learner will demonstrate an understanding of data collection, display, and interpretation.	8	24	0.618
6.1	Gather and organize data from surveys and classroom experiments, including data collected over a period of time.	1	3	0.529
6.2	Display data on charts and graphs; summarize and explain information.	1	4	0.716

Mathematics—Grade 3 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
6.3	Interpret/make pictographs and bar graphs where each symbol/block represents multiple units.	1	4	0.510
6.4	Use charts and graphs as sources of information; identify main idea, draw conclusions, and make predictions.	1	4	0.597
6.5	Locate a designated position using ordered pairs named by letters and numbers.	1	3	0.801
6.6	Locate points on a coordinate grid; name with ordered pairs.	1	3	0.602
6.7	Use a time line to display a sequence of events.	1	3	0.421
7.0	The learner will compute with whole numbers.	24	72	0.695
7.1	Describe and illustrate the connection between models used to demonstrate multiple-digit addition and subtraction and the algorithms.	1	3	0.644
7.2	Model subtraction with zeros; estimate results and demonstrate proficiency with 2-digit and 3-digit addition and subtraction.	8	24	0.799
7.3	Solve meaningful problems using addition and subtraction facts and algorithms; use a calculator in situations involving large numbers and many addends.	3	9	0.625
7.4	Compute total costs of items up to $5.00 and change from up to $5.00.	2	6	0.505
7.5	Demonstrate with a variety of concrete models multiplication and division, including properties of multiplication (identity, commutative, associative).	2	6	0.450
7.6	Memorize multiplication facts/tables: 2s, 5s, 1s, 10s, 9s; explore commutativity and all other facts with concrete materials.	4	12	0.918
7.7	Model division with 1-digit divisor as sharing equally and as repeated subtraction; record results.	1	3	0.494
7.8	Use models to solve real life problems involving multiplication/division.	3	9	0.597

Mathematics—Grade 4

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will identify and use rational numbers.	12	36	0.632
1.1	Within meaningful contexts express numbers (up to 6 digits) in a variety of ways, including oral and written forms using standard and expanded notation.	1	4	0.652
1.2	Use models to explain how the number system is based on 10 and identify the place value of each digit in a multi-digit numeral.	1	4	0.689
1.3	Compare and order numbers less than one million.	1	4	0.873
1.4	In real world situations, discuss when it is appropriate to round numbers; round numbers to an appropriate place.	1	4	0.653
1.5	Use regions, sets, number lines and other concrete and pictorial models to represent fractions and mixed numbers; relate symbols to the models.	1	4	0.693
1.6	Use models and pictures to compare fractions including equivalent fractions and mixed numbers; explain the comparison.	1	4	0.432
1.7	Use models and pictures to demonstrate the value of decimal numerals with tenths and hundredths; show decimals as an extension of the base 10 system.	1	4	0.540
1.8	Use models and pictures to compare decimals (wholes, tenths, hundredths) which relate to real world situations; record and real results.	1	4	0.570
1.9	Use models and pictures to establish the relationship between whole numbers, decimals, and fractions; describe using appropriate language.	1	4	0.527
2.0	The learner will demonstrate an understanding and use properties and relationships of geometry.	7	21	0.465
2.1	Use manipulatives, pictorial representations, and appropriate geometric vocabulary (e.g. sides, angles, and vertices) to identify properties of polygons and other two-dimensional figures.	1	4	0.384

Mathematics—Grade 4 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
2.2	Use manipulatives and appropriate geometric vocabulary (e.g. edges, faces, and vertices) to identify properties of polyhedra and other three-dimensional figures.	1	4	0.584
2.3	Explore turns, flips, and slides with figures.	0	1	0.250
2.4	Make models of line segments and their midpoints, intersecting lines, parallel lines, and perpendicular lines, using materials such as geoboards, paper-folding, straws, and computer graphics.	1	4	0.477
2.5	Use a variety of models to illustrate acute, right, and obtuse angles.	1	4	0.430
2.6	Relate concrete models of lines and angles to pictorial representations and to examples in the environment.	1	4	0.514
3.0	The learner will demonstrate an understanding of patterns and relationships.	7	21	0.532
3.1	Identify and describe mathematical patterns and relationships that occur in the real world.	1	3	0.599
3.2	Demonstrate or describe patterns in geometry, data collection, and arithmetic operations.	1	3	0.500
3.3	Identify patterns as they occur in mathematical sequences.	1	3	0.513
3.4	Extend and make geometric patterns.	1	3	0.559
3.5	Given a table of number pairs, find a pattern and extend the table.	1	3	0.562
3.6	Use patterns to make predictions and solve problems; use calculators when appropriate.	1	3	0.485
3.7	Use intuitive methods, inverse operations, and other mathematical relationships to find solutions to open sentences.	1	3	0.509
4.0	The learner will understand and use standard units of metric and customary measure.	12	36	0.538
4.1	Select an appropriate unit and measure length (inches, feet, yards, centimeters and meters).	1	4	0.630
4.2	Weigh objects using appropriate units and tools (ounces, pounds, grams, kilograms).	1	4	0.648

Mathematics—Grade 4 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.3	Measure capacity with appropriate units (milliliters, teaspoons, tablespoons, cups, pints).	1	3	0.590
4.4	Identify a model that approximates a given capacity unit (cup, quart, gallon, milliliter, and liter).	1	3	0.610
4.5	Estimate the number of units of capacity in a given container and check the estimate by actual measurement.	NT	NT	NT
4.6	Compare units of length, capacity, and weight within the same system.	1	4	0.396
4.7	Explore elapsed time problems using clocks and calendars.	1	3	0.536
4.8	Use appropriate language and proper notation to express and compare money amounts.	1	4	0.609
4.9	Use models to develop the relationship between the total number of square units and the length and width of rectangles. Measure perimeter and determine area of rectangles using grids.	1	4	0.533
4.10	Find the approximate area of regular and irregular figures using grids.	1	3	0.449
4.11	Formulate and solve meaningful problems involving length, weight, time, capacity, and temperature; and verify reasonableness of answers.	1	4	0.450
5.0	The student will solve problems and reason mathematically.	12	36	0.484
5.1	Develop an organized approach to solving problems involving patterns, relations, computation, measurement, geometry, numeration, graphing, probability and statistics.	2	6	0.494
5.2	Communicate an understanding of a problem through oral and written discussion.	NT	NT	NT
5.3	Determine if there is sufficient data to solve a problem.	2	6	0.564

Mathematics—Grade 4 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
5.4	In solving problems, select appropriate strategies such as: act it out, make a model, draw a picture, make a chart or graph, look for patterns, make a simpler problem, use logic, work backwards, guess and check, break into parts.	2	6	0.466
5.5	Estimate solutions to problems and justify.	1	3	0.449
5.6	Solve problems by observation and/or computation, using calculators and computers when appropriate.	3	9	0.469
5.7	Verify and interpret results with respect to the original problem. Discuss alternate methods for solutions.	1	3	0.434
5.8	Formulate engaging problems including ones from every day situations.	1	3	0.478
6.0	The learner will demonstrate an understanding and use of graphing, probability, and statistics.	7	21	0.538
6.1	Collect, organize, and display data from surveys, research, and classroom experiments, including data collected over a period of time. Include data from other disciplines such as science, physical education, and social studies.	1	3	0.706
6.2	Formulate questions and interpret information orally and in writing including main idea, from charts, tables, tallies and graphs (bar, line, stem and leaf, pictographs, circle).	1	3	0.580
6.3	As a group, display the same data in a variety of ways; discuss advantages and disadvantages of each form, including ease of creation and purpose of graph.	1	3	0.551
6.4	Explore range, median, and mode as ways of describing a set of data.	1	3	0.508
6.5	Name the ordered pair of a point on a grid; plot positions named by ordered pairs on a coordinate grid.	1	3	0.460
6.6	Use ordered pairs in a variety of engaging situations (e.g. map reading, treasure hunts, games, and designs).	1	3	0.526

Mathematics—Grade 4 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
6.7	Show all possible ways to sequence a given set of objects; list and explain all possible outcomes in a given situation.	1	3	0.393
7.0	The learner will compute with rational numbers.	23	69	0.668
7.1	Estimate results and solve meaningful problems involving addition and subtraction of multi-digit numbers, including those with two or three zeros. Use a calculator in situations involving large numbers (more than 4 digits) or more than 3 addends.	1	4	0.502
7.2	Use mental math skills to approximate answers and to solve problems, using strategies such as estimation and clustering.	1	3	0.507
7.3	Explain multiplication through the use of various models or by giving realistic examples.	1	3	0.643
7.4	Model and explain division in a variety of ways such as sharing equally, repeated subtraction, and rectangular arrays.	1	3	0.536
7.5	Memorize multiplication facts and relate to division facts.	1	3	0.666
7.6	Demonstrate with models special properties of multiplication: commutative, associative, and identity; and the relationship of multiplication and division.	1	3	0.496
7.7	Estimate results; then solve meaningful problems using the multiplication algorithm with 1-digit times 1- to 3-digit and two 2-digit numbers where one is a multiple of 10.	1	4	0.561
7.8	Solve division problems with single-digit divisors and no renaming.	5	16	0.697
7.9	Estimate results; then use calculators and computers to solve problems involving multiple-digit numbers.	1	3	0.383
7.10	Estimate and use models and pictures to add and subtract decimals, explaining the processes and recording results.	1	3	0.503
7.11	Add/subtract whole numbers.	4	12	0.840
7.12	Multiply 1-digit times 1- to 3-digits and two 2-digit numbers where one is a multiple of 10.	4	12	0.813

Mathematics—Grade 5

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will identify and use rational numbers.	12	36	0.449
1.1	Apply place value skills through millions in real world situations including reading, writing, approximating, and comparing numbers in a variety of forms.	1	4	0.628
1.2	Demonstrate and explain the relationship among whole numbers, decimals, and fractions using various models and other representations, choosing the most appropriate form for the task.	1	3	0.474
1.3	Find multiples and factors of a number, explain the process.	1	3	0.424
1.4	Relate exponential notation to repeated multiplication.	2	6	0.411
1.5	Decide whether a given number less than 100 is prime or composite; explain.	1	3	0.303
1.6	In meaningful contexts, name equivalent fractions at the symbolic level. Explain the equivalence.	1	4	0.451
1.7	In realistic situations use symbols to compare decimals (wholes, tenths, hundredths, and thousandths); explain the comparison.	1	4	0.369
1.8	Read, write, and use decimals and fractions in various forms.	1	3	0.530
1.9	Tell whether a fraction is closer to 0, 1/2, or 1; round a mixed fraction or decimal to the nearest whole number.	1	3	0.521
1.10	In meaningful contexts compare fractions, explaining the rationale and using common denominators when appropriate.	1	3	0.207
2.0	The learner will demonstrate an understanding and use properties and relationships of geometry.	10	30	0.465
2.1	Use concrete and pictorial representations, and appropriate vocabulary to compare and classify polygons and polyhedra.	1	3	0.487
2.2	Create models of polyhedra (cubes, cylinders, rectangles, prisms, pyramids) using a variety of materials.	1	3	0.591

Mathematics—Grade 5 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
2.3	Use designs, concrete models, and computer graphics to illustrate reflections, rotations, and translations of plane figures and record your observations.	1	3	0.335
2.4	Draw circles with a compass and identify radius, diameter, chord, center and circumference.	1	3	0.425
2.5	Explore the relationship between radius and diameter; circumference and diameter.	1	3	0.324
2.6	Use a protractor to draw and measure acute, right, and obtuse angles.	2	6	0.502
2.7	Identify and label the vertex, rays, interior and exterior of an angle.	1	3	0.504
2.8	Use a variety of quadrilaterals and triangles to draw a conclusion about the angles' measures.	1	3	0.408
2.9	Use geometric concepts and spatial visualization to estimate results and solve problems.	2	6	0.488
2.10	Explore topics which relate geometry to other strands of mathematics.	NT	NT	NT
3.0	The learner will demonstrate an understanding of patterns and relationships.	8	24	0.478
3.1	Identify and describe patterns as they occur in numeration, computation, geometry, graphs and other applications.	1	4	0.609
3.2	Investigate patterns that occur when changing numerators and denominators of fractions beginning with concrete models and extending to calculator investigations.	1	4	0.497
3.3	Use patterns to solve problems, make generalizations, and predict results.	1	4	0.355
3.4	Create a set of ordered pairs by using a given rule.	1	4	0.458
3.5	Given a group of ordered pairs, identify a rule to generate them or new pairs in the group, using calculators or computers where appropriate.	1	4	0.424
3.6	Model the concept of a variable using realistic situations.	1	4	0.569

Mathematics—Grade 5 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.0	The learner will understand and use standard units of metric and customary measure.	8	24	0.335
4.1	Use and make models to demonstrate formulas for areas and perimeters of squares and rectangles.	1	3	0.336
4.2	Use models to compare units of area within the same system.	1	3	0.249
4.3	Use models to explore and compare given units of volume (cubic inch, cubic foot, cubic yard, cubic centimeter, and cubic meter).	1	3	0.410
4.4	Describe and record the relationships between perimeter and area, and area and volume.	1	4	0.190
4.5	Identify and demonstrate specific relationships of units within the same measurement system.	1	4	0.306
4.6	Solve problems involving applications of length, weight, time, capacity, temperature, perimeter, and area. Check reasonableness of answer.	2	7	0.414
5.0	The student will solve problems and reason mathematically.	12	36	0.432
5.1	Use an organized approach to solve multi-step problems involving numeration, geometry, measurement, patterns, relations, graphing, computation, probability and statistics.	3	6	0.409
5.2	Communicate an understanding of a problem using models, known facts, properties, and relationships.	1	5	0.421
5.3	Determine if there is sufficient information to solve a problem; identify missing and extraneous data.	1	5	0.535
5.4	Use appropriate strategies to solve problems such as restate problems, use models, patterns, classify, sketches, simpler problem, lists, number sentences, guess and check.	1	5	0.379
5.5	In problem solving situations, use calculators and computers as appropriate.	1	5	0.401
5.6	Verify and interpret the results with respect to the original problem. Identify several strategies for solving a problem.	1	5	0.539

Mathematics—Grade 5 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
5.7	Make generalizations and apply them to new problem situations.	1	2	0.330
6.0	The learner will demonstrate an understanding and use of graphing, probability, and statistics.	8	24	0.473
6.1	Explain the kinds of decisions that need to be made in constructing graphs.	1	4	0.446
6.2	Systematically collect, organize, appropriately display and interpret data both orally and in writing using information from many content areas.	1	4	0.562
6.3	Explore increasingly complex displays of data, including multiple sets of data on the same graph, computer applications, and Venn diagrams.	1	4	0.571
6.4	Use range, median and mode as ways of describing a set of data and explore the use of statistics in science, social studies, and the media.	1	2	0.343
6.5	Explore proportions by reducing or enlarging drawings using grids.	1	2	0.556
6.6	Plot points that represent ordered pairs of data from many different sources such as economics, science experiments, and recreational activities.	1	2	0.408
6.7	Investigate probabilities by experimenting with devices that generate random outcomes (i.e. coins, number cubes, spinners), discussing probable outcomes.	1	2	0.444
6.8	Use a fraction to describe the probability of an event.	1	2	0.421
6.9	In a group compare experimental results with (theoretical) expected results for increasingly larger sample sizes.	1	2	0.350
7.0	The learner will compute with rational numbers.	22	66	0.550
7.1	Estimate products and multiply 2-digit numbers.	3	8	0.728

Mathematics—Grade 5 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.2	Explain the division process with 1- and 2-digit divisors.	1	2	0.620
7.3	Justify, estimate, and solve division problems with divisors that are less than 10 or multiples of 10.	3	8	0.671
7.4	Explain what happens when zeros are involved in computation.	1	2	0.427
7.5	Use models to add and subtract fractions with like denominators.	1	2	0.588
7.6	Estimate results; add and subtract fractions with like denominators in the context of problem solving situations.	1	2	0.328
7.7	Use models and pictures to find a fraction of a whole number; explain and record results.	1	2	0.268
7.8	Estimate results and compute sums and differences, with decimal numbers.	5	14	0.547
7.9	Use models and pictures to multiply a whole number times a decimal number; record and explain results.	1	2	0.434
7.10	Estimate and compute products of decimal numbers with 2-digit factors.	1	2	0.406
7.11	Estimate products of multi-digit decimal numbers; find results with a calculator if exact answer is required.	1	2	0.404
7.12	Compare whole number remainders in division to decimal remainders when using a calculator.	1	2	0.203
7.13	Compute averages within a context; use calculator if appropriate.	1	3	0.344
7.14	Within the context of problem solving situations, add, subtract, and multiply decimal numbers.	1	3	0.493
7.15	Add/subtract fractions with like denominators.	4	12	0.576

Mathematics—Grade 6

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will demonstrate an understanding and use of rational numbers.	9	27	0.467
1.1	Use models to relate percent to fractions and decimals; record, read, and explain.	1	4	0.470
1.2	Use models and pictures to demonstrate ratios, proportions and percents; explain relationships.	1	4	0.502
1.3	Read, write, and use numbers in various forms, including fractions, decimals, percents, and exponential notations, choosing the appropriate form for a given task.	1	4	0.408
1.4	Find the prime factorization of a number less than 100.	1	4	0.286
1.5	Use prime factorization to investigate common factors and common multiples using a calculator when appropriate.	1	4	0.402
1.6	Explore relationships among whole numbers, fractions, decimals, and percents using money, concrete models, or a calculator.	1	4	0.475
1.7	Explore other numeration systems, including ancient number systems and alternate bases.	NT	NT	NT
1.8	Explore the meaning of integers in real-life situations.	1	3	0.763
2.0	The learner will demonstrate an understanding and use properties and relationships of geometry.	9	27	0.446
2.1	Build models of 3-dimensional figures (prisms, pyramids, cones, and other solids); describe and record their properties.	1	5	0.462
2.2	Classify angles (interior, exterior, complementary, supplementary) and pairs of lines including skew lines.	2	7	0.393
2.3	Construct congruent segments and congruent angles. Construct bisectors of line segments; using a straight edge and compass.	NT	NT	NT
2.4	Identify and distinguish among similar, congruent, and symmetric figures; name corresponding parts.	1	5	0.636

Mathematics—Grade 6 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
2.5	Recognize the results of translations, reflections, and rotations using technology when appropriate.	1	5	0.319
2.6	Explore changes in shape through stretching, shrinking and twisting.	NT	NT	NT
2.7	Recognize geometry in the environment (e.g. art, nature, architecture).	1	5	0.383
3.0	The learner will demonstrate an understanding of patterns, relationships and pre-algebra.	8	24	0.486
3.1	Represent number patterns in a variety of ways including the use of calculators and computers.	1	5	0.482
3.2	Use patterns to explore the rules for divisibility.	NT	NT	NT
3.3	Use graphs and tables to represent relations of ordered pairs, using a calculator or a computer where appropriate; describe the relationships.	1	5	0.503
3.4	Identify and use patterning as a strategy to solve problems.	1	4	0.426
3.5	Use realistic examples or models to represent concepts and properties of variables, expressions, and equations. (Identity property of zero, Identity property of one.)	1	5	0.561
3.6	Use the order of operations to simplify numerical expressions, verifying the results with a calculator or computer.	1	5	0.421
4.0	The learner will demonstrate an understanding and use of measurement.	8	24	0.320
4.1	Convert measures of length, area, volume, capacity and weight expressed in a given unit to other units in the same measurement system.	1	4	0.352
4.2	Determine whether a given measurement is precise enough for the specific situation; determine when estimates are sufficient for the measurement situation.	1	4	0.306

Mathematics—Grade 6 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
4.3	Explore the relationship of areas of triangles and rectangles with the same base and height. Use models to demonstrate formulas for finding areas of triangles, parallelograms, and circles.	1	4	0.318
4.4	Explore the effect on area and perimeter when changing one or two of the dimensions of a rectangle.	1	3	0.258
4.5	Develop the concept of volume for rectangular solids as the product of area of base and height using models.	1	4	0.347
4.6	Estimate solutions and solve problems related to volumes of rectangular solids.	1	5	0.291
5.0	The student will solve problems and reason mathematically.	12	36	0.418
5.1	Use an organized approach to solve non-routine and increasingly complex problems involving numeration, geometry, pre-algebra, measurement, graphing, computation, probability and statistics.	3	10	0.363
5.2	Analyze problem situations and apply appropriate strategies for solving them.	3	10	0.518
5.3	Use inductive and deductive reasoning to solve problems.	1	3	0.379
5.4	Select an appropriate method for solving problems including estimation, observation, formulas, mental math, paper and pencil calculation, calculator and computers.	3	10	0.375
5.5	Make conjectures and arguments and identify various points of view.	1	3	0.345
6.0	The learner will demonstrate an understanding and use of graphing, probability, and statistics.	12	36	0.399
6.1	Create and evaluate graphic representations of data, including circle graphs.	3	9	0.410
6.2	Use measures of central tendency (mean, median, and mode) and range to describe meaningful data; compare two sets of unequal data.	3	9	0.350

Mathematics—Grade 6 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
6.3	Display data using computer software and explore the use of spreadsheets.	NT	NT	NT
6.4	Locate ordered pairs in meaningful situations using whole numbers, fractions, and decimals in the coordinate plane.	1	3	0.441
6.5	Estimate the likelihood of certain events from experiments or graphical data.	1	3	0.374
6.6	Interpret a statistical statement and discuss the extent to which the results of a sample can be generalized.	1	3	0.333
6.7	Find probabilities of simple events and discuss the implications.	2	6	0.443
6.8	Design an experiment to test a theoretical probability; record and explain results.	1	3	0.451
7.0	The learner will compute with rational numbers.	22	66	0.471
7.1	Use whole number operations to solve real world applications, demonstrating competence with and without calculators (multiplication and division up to 3-digits by 2-digits).	1	4	0.472
7.2	Select appropriate strategies, solve a variety of application problems, and justify the selection.	1	3	0.417
7.3	Divide decimal numbers, record results and explain procedure (1- and 2-digit divisors).	2	7	0.414
7.4	Within a context, estimate results and apply appropriate operations with decimals.	1	4	0.467
7.5	Use models and pictures to demonstrate multiplication and division of fractions and mixed numbers, record and explain results.	1	2	0.295
7.6	Within a meaningful context, use estimation and operations with fractions less than one.	1	2	0.419
7.7	In problem situations, use estimation and operations with fractions and mixed numbers.	1	2	0.401
7.8	In meaningful contexts develop the concept of adding and subtracting integers; record results.	1	2	0.490

Mathematics—Grade 6 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.9	Translate word problems into number sentences that use integers.	1	2	0.550
7.10	Estimate percents in real world situations and justify the estimate.	1	2	0.400
7.11	Use mental math to solve problems involving simple fractions, decimals, and percents.	1	2	0.498
7.12	Relate common fractions to frequently used percents; estimate and calculate using these percents (multiples of 10, 25, 33-1/3, 66-2/3, 75).	1	2	0.389
7.13	Use ratios and proportions to explore probability and other interesting problems, discussing reasonableness of results.	1	2	0.396
7.14	Add/subtract fractions with unlike denominators.	4	12	0.503
7.15	Multiply/divide fractions with unlike denominators.	4	12	0.452
7.16	Multiply decimal numbers (up to 3-digits by 2-digits).	2	6	0.649

Mathematics—Grade 7

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will demonstrate an understanding and use of real numbers.	8	24	0.469
1.1	Use models to represent positive and negative rational numbers.	1	4	0.614
1.2	Compare and order rational numbers in meaningful contexts.	1	4	0.561
1.3	Express whole numbers in scientific notation; convert scientific notation to standard form; explore the use of scientific notation.	1	3	0.382
1.4	Use exponential notation to express prime factorization of numbers less than 100.	1	3	0.518
1.5	Within meaningful contexts use estimation techniques with rational numbers; justify the strategy chosen.	1	3	0.609
1.6	Use geometric models to develop the meaning of the square and the positive square root of a number; estimate square root and find square roots on the calculator.	1	3	0.338
1.7	In meaningful contexts, relate concepts of ratio, proportion, and percent.	1	4	0.365
2.0	The learner will demonstrate an understanding and use properties and relationships of geometry.	8	24	0.344
2.1	Make constructions of perpendicular and parallel lines using straight edge and compass.	1	1	0.300
2.2	Use the concepts and relationships of geometry to solve problems.	1	5	0.309
2.3	Use models to develop the concept of the Pythagorean Theorem.	1	4	0.277
2.4	Identify applications of geometry in the environment.	1	5	0.316
2.5	Given models of 3-dimensional figures, draw representations.	NT	NT	NT
2.6	Given the end, side, and top views of 3-dimensional figures, build models.	1	4	0.567
2.7	Graph on a coordinate plane geometric shapes and congruent figures.	1	5	0.331

Mathematics—Grade 7 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
3.0	The learner will demonstrate an understanding of pre-algebra.	12	36	0.432
3.1	Describe, extend, analyze and create a wide variety of patterns to investigate relationships and solve problems.	2	7	0.332
3.2	Use concrete materials as models to develop the concept of operations with variables.	2	7	0.515
3.3	Use concrete, informal and formal methods to model and solve simple equations.	2	8	0.515
3.4	Investigate and evaluate algebraic expressions using mental calculations, pencil and paper and calculators where appropriate.	2	7	0.427
3.5	Given a simple equation, formulate a problem.	2	7	0.388
4.0	The learner will demonstrate an understanding and use of measurement.	10	30	0.315
4.1	Apply measurement concepts and skills as needed in problem solving situations.	3	9	0.310
4.2	Make judgments about degree of precision needed and reasonableness of results in measurement situations.	1	3	0.398
4.3	Use models to develop the concept and formula for surface area for rectangular solids and cylinders.	2	6	0.277
4.4	Use models to develop the concept of volume for prisms/cylinders as the product of area of the base and height.	1	6	0.358
4.5	Use models to explore the relationship of the volume of a cone to a cylinder, and a pyramid to a prism, with the same base and height.	NT	NT	NT
4.6	Estimate answers; solve problems related to volume.	2	6	0.311
5.0	The student will solve problems and reason mathematically.	14	42	0.353
5.1	Use an organized approach and a variety of strategies to solve increasingly complex non-routine problems.	3	9	0.313
5.2	Use calculators and computers in problem solving situations as appropriate.	5	15	0.367

Mathematics—Grade 7 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
5.3	Discuss alternate strategies, evaluate outcomes, and make conjectures and generalizations based on problem situations.	2	6	0.375
5.4	Use concrete or pictorial models involving spatial visualization to solve problems.	2	6	0.273
5.5	Solve problems involving interpretation of graphs, including inferences and conjectures.	2	6	0.449
6.0	The learner will demonstrate an understanding and use of probability and statistics.	8	24	0.404
6.1	Create, compare, and evaluate both orally and in writing different graphic representations of the same data.	1	3	0.572
6.2	Construct a box plot (box and whiskers) by ordering data, identifying the median, quartiles, and extremes.	1	3	0.283
6.3	Evaluate appropriate uses of different measures of central tendency.	1	3	0.280
6.4	Draw inferences and construct convincing arguments based on analysis of data.	1	3	0.431
6.5	Investigate and recognize misuses of statistical or numerical information.	1	3	0.496
6.6	Show all possible outcomes by making lists, tree diagrams, and frequency distribution tables.	1	3	0.420
6.7	Explain the relationship between experimental results and mathematical expectations.	1	3	0.359
6.8	Find the probability of simple events using experiments, random number generation, computer simulation, and theoretical methods.	1	3	0.322
6.9	Explore permutations and combinations in applications.	NT	NT	NT
7.0	The learner will compute with real numbers.	20	60	0.408
7.1	Select appropriate operations, strategies, and methods of solving a variety of application problems using positive rational numbers, and justify the selection.	2	7	0.385

Mathematics—Grade 7 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.2	Estimate and solve problems using ratio, proportion, and percent; select and use appropriate methods; explain the process used.	5	19	0.439
7.3	Apply concepts of ratio, proportion, and percent to real life situations such as consumer applications, science and social studies.	2	7	0.353
7.4	Use real world examples and models to represent multiplication and division of integers; record and explain procedures used.	2	7	0.361
7.5	Use operations with integers in relevant problem situations.	2	8	0.372
7.6	Use operations with integers.	4	12	0.443

Mathematics—Grade 8

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
1.0	The learner will demonstrate an understanding and use of real numbers.	11	33	0.425
1.1	Explore the real number system by describing and using various forms of numbers in realistic situations.	2	5	0.386
1.2	Use appropriate estimation techniques in meaningful situations; justify the technique.	2	6	0.465
1.3	Use and explain definitions and laws of exponents to write expressions in equivalent forms.	2	6	0.413
1.4	Use scientific notation to express whole numbers and numbers less than one, using a calculator when appropriate.	2	6	0.305
1.5	Investigate irrational numbers and their representations on a calculator as they arise from problem situations.	NT	NT	NT
1.6	Describe the properties of terminating, repeating, and non-repeating decimals and be able to convert fractions to decimals and decimals to fractions.	2	5	0.457
1.7	Explore the absolute value of a number using the number line.	2	5	0.513
2.0	The learner will demonstrate an understanding and use properties and relationships of geometry.	8	24	0.332
2.1	Use the Pythagorean Theorem to find the missing side of a right triangle; use calculator when appropriate.	2	6	0.302
2.2	Solve problems related to similar figures using indirect measures to determine missing sides.	2	6	0.370
2.3	Draw 3-dimensional figures from different perspectives (top, side, front).	1	3	0.610
2.4	Graph on a coordinate plane similar figures, reflections, and translations.	1	3	0.346
2.5	Explore the triangle congruency relationships: ASA, SSS, SAS.	NT	NT	NT
2.6	Explore the relationships of the angles formed by cutting parallel lines by a transversal.	1	3	0.575
2.7	Solve problems that relate geometric concepts to real world situations.	1	3	0.228

Mathematics—Grade 8 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
3.0	The learner will demonstrate an understanding of pre-algebra.	14	42	0.383
3.1	Describe, extend, analyze and create a wide variety of geometric and numerical patterns, such as Pascal's triangle or the Fibonacci sequence.	2	6	0.310
3.2	Identify and define the commutative, associative and distributive properties; give examples and explain their meanings.	2	6	0.356
3.3	Analyze representations of data with tables, graphs, verbal rules and equations to explore properties and relationships.	2	6	0.361
3.4	Using patterns and algebraic methods, solve problems, including those with integers.	2	6	0.424
3.5	Generate ordered pairs with and without a calculator and graph the linear equation.	2	6	0.340
3.6	Investigate non-linear equations and inequalities informally.	2	6	0.479
3.7	Given a formula make appropriate substitutions and solve for one unknown.	2	6	0.371
4.0	The learner will demonstrate an understanding and use of measurement.	8	24	0.344
4.1	Estimate the answer; then solve complex problems that include application of measurement; determine precision and check for reasonableness of results.	3	9	0.373
4.2	Determine the number of significant digits, the greatest possible error and relative error in measurement situations.	NT	NT	NT
4.3	Select an appropriate unit and tool to find a measurement based upon the degree of accuracy required and the nature of the problem situation.	1	3	0.512
4.4	Find the surface area and volume of pyramids, prisms, cylinders, and cones with and without models.	3	9	0.261
4.5	Explore the effect on plane and solid figures when a dimension of the figure is changed.	1	3	0.257

Mathematics—Grade 8 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
5.0	The student will solve problems and reason mathematically.	12	36	0.399
5.1	Use an organized approach and a variety of strategies to solve increasingly complex non-routine problems.	3	9	0.341
5.2	Use calculators and computers in problem solving situations as appropriate.	2	6	0.431
5.3	Make and evaluate conjectures and arguments, using deductive and inductive reasoning.	1	3	0.376
5.4	Investigate open-ended problems, formulate questions, and extend problem solving situations.	1	3	0.417
5.5	Represent situations verbally, numerically, graphically, geometrically, or symbolically.	2	6	0.439
5.6	Use proportional reasoning to solve problems.	3	9	0.399
6.0	The learner will demonstrate an understanding and use of probability and statistics.	10	30	0.394
6.1	Collect data involving two variables and display on a scatter plot; interpret results.	1	3	0.526
6.2	Compute the mean, interpret it, explain its sensitivity to extremes, and explain its use in comparison with the median.	2	7	0.403
6.3	Apply knowledge of statistics in problem solving situations, selecting an appropriate format for presenting data.	2	7	0.442
6.4	Use mathematical probabilities and experimental results for making predictions and decisions.	1	3	0.370
6.5	Evaluate arguments based on data and investigate reasons why an inference made from a set of data can be invalid (biased vs. unbiased).	1	3	0.336
6.6	Find the probability of simple and compound events using experiments, computer simulations, random number generation, and theoretical methods.	2	7	0.343

Mathematics—Grade 8 (continued)

Goal/ Objective	Description of Goal/Objective	No. of Items per Form	No. of Items per Class	Difficulty of Pool
7.0	The learner will compute with real numbers.	17	51	0.437
7.1	Select appropriate operations, strategies, and methods of solving a variety of application problems using real numbers, justifying the selection.	13	42	0.434
7.2	In meaningful contexts, develop the laws of exponents; solve problems involving exponentiation.	3	9	0.452

North Carolina Technical Advisory Group

Dr. John Fremer (Chairperson)
Educational Testing Service

Dr. Mark Applebaum
Vanderbilt University

Dr. Lloyd Bond
University of North Carolina at Greensboro

Dr. Joy McLarty
American College Testing

Dr. David Thissen
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Dr. Gary Williamson
Educational Consultant, North Carolina

Sample Item with Development, Review, and Psychometric Information

ITEM SPECIFICATIONS SHEET

080547

CURRICULUM OBJECTIVE:

5.6 Use proportional reasoning to solve problems.

SUBTOPIC:

CURRICULUM SOURCE:

DIFFICULTY LEVEL: 1 = EASY; (2) = MEDIUM; 3 = HARD	LEVEL OF THINKING: Thinking Skill: Applying(5)		PAGEMAKER: 1 = YES; (2) = NO
ITEM WRITER NUMBER: 822	ARTWORK REQUIRED: (IF YES, PLEASE ATTACH)	1 = YES; (2) = NO	MACEQN: 1 = YES; (2) = NO

MATHEMATICS TEST ITEM (*FINAL DRAFT*) GRADE: 8

A model for a newly designed airplane is being built to a scale of 1" = 10'. If the airplane is 75' long, what is the length of the model?

a) ~~75 in~~ 7.25 in
b) ~~7.25 in~~ 7.5 in
c) ~~7.5 in~~ 75 in
d) 750 in

CORRECT ANSWER ~~C~~ B

EDIT cm ____ PC

Did You...

1. focus directly on the objective?
2. write stem as a complete statement of question?
3. write foils of equal length with *only* one correct answer?
4. use same context and similar ideas in foils?
5. avoid using negatives in the foils?
6. arrange continuous foils in logical order?
7. make each foil credible?
8. check punctuation, spelling, and grammatical structure of item?
9. use artwork *only* when necessary?
10. practice fair representation in sex and race, avoiding culture specific references?

10-56
NCTests

Numbers for Keypunch only	
1,2	I.D.
3,4	Goal
5,6	Obj
7	Think
8	Diff
9	Diagram
10-13	Item No.

Mathematics Item Review — Grade 8

5.6 Use proportional reasoning to solve problems.

Difficulty Level: Medium

80547 A model for a newly designed airplane is being built to a scale of 1" = 10'. If the airplane is 75' long, what is the length of the model?

A 7.25 in

B 7.5 in

C 75 in

D 750 in

14 **Correct Answer:** B

15 **Thinking Skill Level:**
- ☐ Knowledge (1)
- ☐ Organizing (4)
- ☐ Applying (5)
- ☐ Analyzing (6)
- ☐ Generating (7)
- ☐ Integrating (8)
- ☐ Evaluating (9)

CONCEPTUAL
Yes Marginal No
16 ☐ ☐ ☐
- 17 ☐ Objective match
- 18 ☐ Fair representation
- 19 ☐ No cultural bias
- 20 ☐ Clear statement
- 21 ☐ Single problem
- 22 ☐ One best answer
- 23 ☐ Common context in foils
- 24 ☐ Each foil credible
- 25 ☐ Other

LANGUAGE
Yes Marginal No
26 ☐ ☐ ☐
- 27 ☐ Appropriate for age
- 28 ☐ Punctuation, spelling, grammar
- 29 ☐ No excess words
- 30 ☐ No stem/foil clues
- 31 ☐ No negatives in foils
- 32 ☐ Other

FORMAT
Yes Marginal No
33 ☐ ☐ ☐
- 34 ☐ Logical order of foils
- 35 ☐ Familiar presentation style
- 36 ☐ Print size and type
- 37 ☐ Mechanics and appearance
- 38 ☐ Equal length foils
- 39 ☐ Other

DIAGRAM
Yes Marginal No
40 ☐ ☐ ☐
- 41 ☐ Necessary
- 42 ☐ Clear
- 43 ☐ Relevant
- 44 ☐ Unbiased
- 45 ☐ Other

46 **OVERALL RATING:** ☒ Acceptable ☐ Acceptable with modifications ☐ Discard Item

,ɘ/91

Report By Item
MATH ITEM REVIEW - GRADE 8
Item number: 80547

Goal: 5.6

THINKING SKILL LEVEL		ANSWER	
Knowledge:	2	A:	0
Organizing:	0	B:	3
Applying:	2	C:	1
Analyzing:	0	D:	0
Generating:	0	E:	0
Integrating:	0		
Evaluating:	0		

	No	Marginal	Yes
Conceptual :	0	0	4
Language :	0	0	4
Format :	0	0	4
Diagram :	0	0	2

OVERALL RATING
Acceptable : 4
Modify : 0
Discard : 0

TOTAL REVIEWERS: 4

COMMENTS

807 DIFFICULTY = EASY

815 DIFFICULTY LEVEL = EASY

BEST COPY AVAILABLE

Mathematics — Grade 8

Objective: 5.6 Use proportional reasoning to solve problems.

On Grade

Mathematics - Grade 8

Origno	Form	Item	Obj	Key	A	B	C	D
80547	10	56	5.6	2	100	400	195	176

P	Bis	Psd	Bias: Ethnic	Bias: Gender	IRT Parameters: Threshold	IRT Parameters: Slope	IRT Parameters: Asymptote
.45	0.462	.50	1.374	1.043	0.874	0.999	.251

Above Grade

56. A model for a newly designed airplane is being built to a scale of 1" = 10'. If the airplane is 75' long, what is the length of the model?

A 7.25 in.

B 7.5 in.

C 75 in.

D 750 in.

Psychometric Approval	Curriculum Approval
Yes No	(Yes) No

Excerpt from

"Item Response Theory for Scores on Tests including Polychotomous Items with Ordered Responses"

by David Thissen, Mary Pommerich, Kathleen Billeaud, and Valerie S.L. Williams
L.L. Thurstone Psychometric Laboratory of the University of North Carolina at Chapel Hill.
Research Report Number 94-2, Published May, 1994.

Item Response Theory for Scores on Tests including Polychotomous Items with Ordered Responses

David Thissen
Mary Pommerich
Kathleen Billeaud
Valerie S.L. Williams
L.L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Abstract

Item response theory (IRT) provides a ready mechanism for scoring tests including any combination of rated constructed-response and keyed multiple-choice items, in that each response pattern is associated with some modal or expected *a posteriori* estimate of proficiency. However, various considerations that frequently arise in large-scale testing make response-pattern scoring an undesirable solution. In this paper, we describe methods based on IRT that provide scaled scores, or estimates of proficiency, for each summed score for rated responses, or for combinations of rated responses and multiple-choice items. These methods may be used to combine the useful scale properties of IRT-based scores with the practical virtues of a scale based on the summed score for each examinee.

The research reported here was supported by the North Carolina Department of Public Instruction, in conjunction with the development of the North Carolina End of Grade Testing Program. We thank Richard Luecht, Robert McKinley, Robert Mislevy, James Ramsay, and Linda Wightman for their help in the course of this work.

One feature of item response theory (IRT) is that it provides a score scale that is more useful than the summed score, percentage-correct, or percentile scales for many purposes, e.g., for construction of developmental scales or for calibration of tests comprising different types of items or exercises. With the exception of the so-called Rasch-family of models for which the summed score is a sufficient statistic for the characterization of the latent variable (θ) (Rasch, 1960; Masters and Wright, 1984), under IRT models each response pattern is associated with a unique estimate of θ. These estimates of θ may be used as scaled response pattern scores; they have the advantage that they extract all information available in the item responses. In addition, the IRT model produces estimates of the probability that each response pattern will be observed in a sample from a specified population.

However, in applied measurement contexts, it is often desirable for various reasons to consider the implications of the IRT analysis for summed scores, rather than response patterns, even if the IRT model used is not part of the Rasch family. For various practical reasons it may be desirable to report IRT scaled scores based on the summed score, rather than the scaled scores that are associated with each response pattern. In addition, it may be useful to compute model-based estimates of the summed score distribution, e.g., to create percentile tables for use as an interpretive aid for score reporting. Model-based estimates of the summed score distribution may also have value as a statistical diagnostic of the goodness of fit of the IRT model, including the validity of the assumed underlying population distribution.

Many contemporary tests include extended constructed-response items, for which the item scores are ordered categorical ratings provided by judges. In some cases, the constructed-response items comprise the entire test; in other cases, there are multiple-choice items as well. In either case, some total score is often required, combining the judged ratings of the constructed-response items (and the item scores on the multiple-choice items, if any are present). Simple summed scores may not be very useful in this context, because of the problems associated with the selection of relative weights for the different items and item types, and because the constructed-response items are often on forms of widely varying difficulty.

Item Response Theory for Summed Scores

For any IRT model for items indexed by i with ordered item scores $k = 0, \ldots K_i$, the likelihood for any summed score $j = \sum k_i$ is

$$L_j(\theta) = \sum_{\in j = \sum k_i}^{patterns} L(\mathbf{k}|\theta) \ ,$$

where the summation is over the response patterns with total score j. The likelihood for each response pattern is

$$L(\mathbf{k}|\theta) = \prod_i T_{k_i}(\theta)\phi(\theta) \ ,$$

where $T_{k_i}(\theta)$ is the trace line for category k of item i—the conditional probability of response k to item i given θ—and $\phi(\theta)$ is the population density. Thus, the likelihood for each score is

$$L_j(\theta) = \sum_{\in j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta) \ ,$$

and so the probability of each score j is

$$P_j = \int L_j(\theta)\, d\theta \ ,$$

or

$$P_j = \int \sum_{\in j = \sum k_i}^{patterns} L(\mathbf{k}|\theta)\, d\theta \ ,$$

or, most intimidatingly,

$$P_j = \int \sum_{\in j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta)\phi(\theta)\, d\theta\ . \qquad (1)$$

Given an algorithm to compute the integrand in Equation 1, it is straightforward to compute the average (or expected *a posteriori*, or *EAP*) scaled score (Bock and Mislevy, 1982) associated with each score,

$$EAP(\theta | j = \textstyle\sum k_i) = \frac{\int \theta \; L_j(\theta) d\theta}{P_j}\ , \qquad (2)$$

and the corresponding standard deviation,

$$SD(\theta | j = \textstyle\sum k_i) = \sqrt{\frac{\int [\theta - EAP(\theta | \sum k_i)]^2 \; L_j(\theta)\, d\theta}{P_j}}\ . \qquad (3)$$

The values computed using Equation 2 may be tabulated and used as the IRT scaled-score transformation of the raw scores, and the values of Equation 3 may be used as a standard description of the uncertainty associated with those scaled scores.

The score-histogram created using the values of Equation 1 may be used to construct summed-score percentile tables; if the IRT model fits the data, this can be done accurately using only the item parameters, for any group with a known population density. Thus, percentile tables for summed scores can be constructed using item tryout data, before the operational test is administered. This same histogram may also prove useful as a diagnostic statistic for the goodness of fit of the model, by comparing the modeled representation of the score distribution to the observed data.

Algorithms for Computing $L_j(\theta)$

Lord (1953) used heuristic procedures to describe the difference between the distribution of summed scores, $L_j(\theta)$, and the underlying distribution of θ, $\phi(\theta)$ [see also Lord and Novick (1968, pp. 387-392)]. However, practical calculation of the summed score distribution implied by an IRT model has awaited both contemporary computational power and solutions to the apparently intractable computational problem.

Brute force evaluation of Equation 1, requiring the computation of $\prod(K_i + 1)$ likelihoods, is easy for a few items; but it is inconceivable for many items. Brute force may be extended to moderate numbers of items (i.e., about 20) by using an algorithm involving the computation of each pattern likelihood from some other previously-computed pattern likelihood by a single (list) multiplication; this approach is used in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1991). For binary items, by carefully ordering the computation of the likelihoods for the 2^I patterns (where I is the number of items), such an algorithm can compute all 2^I likelihoods at a computational cost of only a single (list) multiplication for each (Thissen, Pommerich, and Williams, 1993). Nevertheless, due to the exponential computational complexity of the brute force approach, this algorithm cannot be extended to more items regardless of improvements in computational speed.

Lord and Novick (1968, p. 525) stated that "approximations appear inevitable," and suggested the use of an approximation to the compound binomial, attributed to Walsh (1963), to compute the likelihood of a summed score for binary items as a function of θ. For I items, this Taylor-series expansion has I terms; however, in practice the first two terms suffice for acceptable accuracy. The two-term version of the approximation is:

$$\sum_{\in j = \sum k_i}^{patterns} \prod_i T_{k_i}(\theta) \cong p_I(j) + \frac{I}{2} VC(j)\ , \qquad (4)$$

where

$$p_I(j) = \begin{cases} \binom{I}{j} M^j (1-M)^{I-j} & \text{for } j = 0, 1, \ldots, I; \\ 0 & \text{otherwise} \end{cases} ,$$

$$C(j) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} p_{I-2}(j-v) ,$$

$$V = \frac{1}{I} \sum_{i} [T_{1_i}(\theta) - M]^2 ,$$

and

$$M = \frac{1}{I} \sum_{i} T_{1_i}(\theta) .$$

Yen (1984) used this approximation to develop an algorithm to compute the mode of

$$\sum_{\in\, j = \sum k_i}^{patterns} \prod_{i} T_{k_i}(\theta)$$

for use as a scaled score for examinees with summed score j on a binary test using the three-parameter logistic model. She reported that the two-term Taylor expansion produced noticeably better results than the one-term solution, which is simply an inverse transformation of the test characteristic curve; but the three- and four-term solutions appeared not to add useful precision.

The approximation in Equation 4 may also be substituted for the sum of products in Equations 1, 2, and 3 to compute P_j, $EAP(\theta | j = \sum k_i)$, and $SD(\theta | j = \sum k_i)$. When the results for the two-term approximation were compared to the correct (brute force) results for one of the 20-item examples used by Yen (1984), the error of approximation was usually less than 0.001 for $EAP(\theta | j = \sum k_i)$, and $SD(\theta | j = \sum k_i)$ (Thissen, et al., 1993). Exceptions tended to be the perfect scores, for which the second term of the two-term approximation is zero; there, the approximation could be off by as much as 0.05. The error of approximation for P_j tended to be of the order of 0.0001. For practical use in constructing score-reporting tables, which usually use no greater precision than tenths of a standard deviation for the scores and their standard errors, and integral values for percentile tables, this degree of precision appears to be sufficient. However, the approximation in Equation 4 is still somewhat computationally burdensome, and no generalization has been offered for items with more than two response categories.

The problem of the computational burden is solved by an alternative procedure briefly described by Lord and Wingersky (1984). Abandoning the contention of Lord and Novick (1968, p. 525) that "approximation is inevitable," Lord and Wingersky described a simple recursive algorithm for the computation of

$$L_j^I(\theta) = \sum_{\in\, j = \sum k_i}^{patterns} \prod_{i} T_{k_i}(\theta)$$

for binary items. The algorithm is based on the distributive law, and generalizes readily to items with any number of response categories.

The generalization follows, using the notation $i = 0, 1, \ldots I$ for the items, $k = 0, 1, \ldots K_i$ for the response categories for item i, and $T_{k_i}(\theta)$ for the trace line for category k of item i. In addition, the summed scores for a set of items $[0 \ldots I^*]$ are $j = 0, 1, \ldots \sum_{I^*} K_i$ and the likelihood for summed score j for a set of items $[0 \ldots I^*]$ is $L_j^{I^*}(\theta)$; the population distribution is $\phi(\theta)$.

The generalized recursive algorithm is:

Set $I^* = 0$

$$L_j^{I^*}(\theta) = T_{j_{I^*}}(\theta) , \text{ for } j = 0, 1, \ldots K_{I^*}$$

Repeat:

For item $I^* + 1$ and scores $j = 0, 1, \ldots \sum_{I^*} K_i$

$$L_{j+k}^{I^*+1}(\theta) = \sum_{k_{I^*+1}} L_j^{I^*}(\theta)\, T_{k_{I^*+1}}(\theta)$$

Set $I^* = I^*+1$

Until $I^* = I$.

For a sample from a population with distribution $\phi(\theta)$, the likelihood for score j is

$$L_j(\theta) = L_j^I(\theta)\,\phi(\theta)$$

and $EAP(\theta|j = \sum k_i)$, $SD(\theta|j = \sum k_i)$, and $P_j(\theta)$ can be computed by integrating $L_j(\theta)$.

No particular parametric form for the trace lines is assumed in the formulation of the recursive algorithm. We have used the three-parameter logistic in work with binary-scored multiple-choice items, and Samejima's (1969) graded model for multiple-category rated items. However, in principle, any trace lines could be used, such as the nonparametric kernel smooths described by Ramsay (1991). The algorithm would produce perfectly accurate, if silly, results if it were used with items for which the responses are *not* ordered. The results would be silly because the response patterns included in any particular summed score would not tend to have likelihoods concentrated near the same values of θ, and so such summed-score likelihoods would tend to be very flat with very large standard deviations.

Nevertheless, the algorithm is completely general. An implementation for the LISP-STAT computing environment (Tierney, 1990) is given in the appendix.

Bias Review Materials

Item Bias Review

Background

To develop achievement tests that are valid, reliable, and educationally appropriate, the North Carolina Department of Public Instruction carries out a series of operations that take several years. In a broad overview, the procedures involve curriculum definition, test design, item writing, item editing, item review, field testing, analyses of field test data, further item editing and review, selection of items for tests, review of tests "as tests", final editing, and then test administration. All item reviews are accomplished by North Carolina teachers and other professional educators. One of the purposes of the item review is to ensure that the test items do not reflect any cultural bias or stereotyping.

All test items have already been reviewed for cultural bias. However, at this time statistical analyses have been performed on the field test data to detect gender and ethnic bias. Items are flagged as "biased" by the statistical techniques if an identified group (males or females, blacks or whites) performs better than would be expected from their overall proficiency in the content area measured by the test. These items require a closer look: judgments must be made about whether the difference in performance on the item is relevant to what the test is intending to measure. In other words, is what is measured by the test item, and the context in which it is measured, something that should be taught as part of the curriculum? If not, then the item is biased and should not be used on the tests. It is these judgments which will determine if an item is eliminated from the item pool due to bias.

Instructions

Enclosed are test items that have been flagged as potentially biased. Each test item is on a separate sheet. Passages and/or supporting materials are included for some questions. At the top of the sheet is the curricular objective the item is intending to measure, followed by statistics reflecting item performance on the field test. Of particular interest to your item review are the following:

- **Origno** - a five-digit item number that uniquely identifies the item
- **P** - the proportion of test-takers that got the item correct
- **Key** - the correct answer choice (1=A, 2=B, 3=C, 4=D)
- **Ethnic/Gender Bias**
  –if greater than 1.5, then the item favors females or whites
  –if less than .66, then the item favors males or blacks
  (Note that the group favored by the item is written in the bottom left-hand corner of the page. If an item favors females, then it is biased against males, etc.)
- **Choice** - the number of test-takers selecting each answer choice

The other numbers on the form reflect other characteristics of the item and can be ignored for this purpose. If the item was tested in the next grade level as well, statistics for that grade level are presented in the "Above Grade" box.

When reviewing the test items, keep in mind the following five questions:

1. Does the item contain any offensive gender, ethnic, and/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 area?

If your answer is *Yes* to any of the five questions, record the five-digit "Origno" and check the appropriate column(s) on the Item Bias Review Sheet. You should comment on all rejected items on your copy of the item sheets. **If the item is acceptable as is, do *not* record its "Origno" on the Item Bias Review Sheet. Only items that should be revised or discarded should be recorded on the Item Bias Review Sheet.**

After you have completed your review, return *all* materials in the enclosed self-addressed envelope.

A Note about Test Security
It is important to note that these achievement test items are the property of the North Carolina Department of Public Instruction. If the items are not securely held, they will be useless to us. Do not copy the items; do not show them to anyone else; do not discuss their content with other people; and do keep them in a secure place when you are not reviewing them. Your help with security is essential to producing a test that is fair to all students.

Item Bias Review Sheet:Rejected Items

Name of Reviewer ______________________________

Directions When reviewing the test items, keep in mind the following five questions:

1. Does the item contain any offensive gender, ethnic, regional, and/or ethnic 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 area?

If the answer is *Yes* to any of the questions, record the five-digit "Origno" below and check the appropriate column(s). You do not need to record items that are acceptable.

	Standard Failed						
Origno (Item #)	**1**	**2**	**3**	**4**	**5**	**Revise**	**Reject**

Published 1996