

**The North Carolina Science Tests**

**Technical Report**

**End-of-Course Physical Science Test**

**End-of-Course Biology Test**

**End-of-Course Chemistry Test**

**End-of-Course Physics Test**

June 2009

---

---

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

Dr. Rebecca Garland, Chief Academic Officer  
Academic Services and Instructional Support  
6368 Mail Service Center  
Raleigh, NC 27699-6368  
Telephone (919) 807-3200; fax (919) 807-4065

# Table of Contents

<b>Chapter One: Introduction</b>	<b>1</b>
1.1 Universal Participation	1
1.2 The North Carolina Testing Program	2
1.3 The North Carolina Science Tests	4
<b>Chapter Two: Test Development Process</b>	<b>6</b>
2.1 Test Development Process for the North Carolina Testing Program	6
2.2 The Curriculum Connection	8
2.3 Test Specifications	8
2.4 Item Development	9
2.5 Item Format	10
2.6 Selection and Training of Item Writers	11
2.7 Reviewing Items for Field Testing	11
2.8 Assembling Field Test Forms	12
2.9 Sampling Procedures and Field Test Sample Characteristics	13
2.10 Item Analysis	14
2.11 Classical Measurement Analysis	14
2.12 Item Response Theory (IRT) Analysis	14
2.13 Differential Item Functioning Analysis	16
2.14 Expert Review	17
2.15 Criteria for Inclusion in Item Pool	18
2.16 Item Pool Parameter Estimates	19
2.17 Operational Test Construction	19
2.18 Establishing the Target p-value for Operational Tests	19

2.19 Comparison of Item Pool p-Values with Operational p-Values	20
2.20 Review of Assembled Operational Tests	20
2.21 Establishing the Test Administration Time	21
<b>Chapter Three: Test Administration</b>	<b>22</b>
3.1 Test Administration	22
3.2 Training for Test Administrators	22
3.3 Preparation for Test Administration	22
3.4 Test Security and Handling Materials	23
3.5 Student Participation	23
3.6 Alternate and Alternative Assessments	24
3.7 Testing Accommodations	26
3.8 Students with Limited English Proficiency	26
3.9 Medical Exclusions	26
3.10 Reporting Student Scores	27
3.11 Confidentiality of Student Test Scores	27
<b>Chapter Four: Scaling, Equating, and Standard-Setting for the North Carolina Tests of Science</b>	<b>28</b>
4.1 Conversion of Raw Test Scores	28
4.2 Setting the Standards	28
4.3 Score Reporting for the North Carolina Tests	29
4.4 Achievement Level Descriptors	29
4.5 Achievement Level Cut Scores	30
4.6 Achievement Level Trends	30
4.7 Percentile Ranking	33
<b>Chapter Five: Reports</b>	<b>34</b>

5.1 Use of Test Score Reports Provided by the North Carolina Testing Program	32
5.2 Reporting by Student	32
5.3 Reporting by Classroom	33
5.4 Reporting by School	33
5.5 Reporting by District	34
5.6 Reporting by the State	34
<b>Chapter Six: Descriptive Statistics and Reliability</b>	<b>35</b>
6.1 Descriptive Statistics for the First Operational Administration of the Tests	35
6.2 Means and Standard Deviations for the First Year of Operational Administration	35
6.3 Population Demographics for the First Operational Administration	36
6.4 Scale Score Frequency Distributions	37
6.5 Reliability of the North Carolina Science Tests	41
6.6 Internal Consistency of the North Carolina Science Tests	41
6.7 Standard Error of Measurement	43
6.8 Equivalency of Test Forms	48
<b>Chapter Seven: Evidence of Validity</b>	<b>53</b>
7.1 Evidence of Validity	53
7.2 Content Validity	53
7.3 Criterion-Related Validity	54
7.4 Concurrent and Predictive Validity	56
7.5 Alignment	56
<b>Chapter Eight: Quality Control Procedures</b>	<b>58</b>
8.1 Quality Control Prior to Test Administration	58

8.2 Quality Control in Data Preparation and Test Administration	58
8.3 Quality Control in Data Input	59
8.4 Quality Control of Test Scores and Data Merging	59
8.5 Quality Control in Reporting	59
<b>Glossary of Key Terms</b>	<b>60</b>
<b>References</b>	<b>63</b>
<b>Additional Resources</b>	<b>64</b>
<b>Appendix A: Test Specification Summaries</b>	<b>66</b>
<b>Appendix B: Item Development Guidelines</b>	<b>94</b>
<b>Appendix C: Testing Code of Ethics</b>	<b>96</b>

## List of Tables

Table 1:	Number of items field tested for North Carolina EOC Tests of Science	12
Table 2:	Field test population (2006-07) for EOC Tests of Science	13
Table 3:	Average item pool parameter estimates for EOC Tests of Science by course	19
Table 4:	Comparison of p-value of item pool with p-values of assembled forms and operational p-values	20
Table 5:	Number of items per test and time allotted by subject	21
Table 6:	Available assessments in the North Carolina EOC Science Testing Program	24
Table 7:	EOC Tests of Science achievement levels and corresponding scale scores	30
Table 8:	Achievement level trends for EOC Physical Science	31
Table 9:	Achievement level trends for EOC Biology	31
Table 10:	Achievement level trends for EOC Chemistry	31
Table 11:	Achievement level trends for EOC Physics	31
Table 12:	Descriptive statistics by course for the first administration of the North Carolina EOC Tests of Science	35
Table 13:	Population demographics for the first administration of the North Carolina EOC Tests of Science	36
Table 14:	Reliability indices averaged across North Carolina EOC Tests of Science Forms	41
Table 15:	Reliability indices averaged across North Carolina EOC Tests of Science Forms (Gender)	41
Table 16:	Reliability indices averaged across North Carolina EOC Tests of Science Forms (Ethnicity)	42
Table 17:	Reliability indices averaged across North Carolina EOC Tests of Science Forms (Other Characteristics)	42

Table 18:	Ranges of standard error of measurement for scale scores by subject	43
Table 19:	Pearson correlation coefficient table for variables used to establish criterion-related validity for the North Carolina EOC Tests of Science	55
Table 20:	Tetrachoric correlation coefficient table for additional, presumably uncorrelated variables used to establish criterion-related validity for the North Carolina EOC Tests of Science	56

## List of Figures

Figure 1:	Flow chart of the test development process used in development of North Carolina Tests	7
Figure 2:	Thinking skills framework used to develop the North Carolina Science Tests	10
Figure 3:	Typical item characteristic curve (ICC) for a 4-option multiple-choice item	15
Figure 4:	Physical Science Scale Score Frequency Distribution	37
Figure 5:	Biology Scale Score Frequency Distribution	38
Figure 6:	Chemistry Scale Score Frequency Distribution	39
Figure 7:	Physics Scale Score Frequency Distribution	40
Figure 8:	Standard Errors of Measurement on the Physical Science Test Forms	44
Figure 9:	Standard Errors of Measurement on the Biology Test Forms	45
Figure 10:	Standard Errors of Measurement on the Chemistry Test Forms	46
Figure 11:	Standard Errors of Measurement on the Physics Test Forms	47
Figure 12:	Test Characteristic Curves for the Physical Science Test Forms	49
Figure 13:	Test Characteristic Curves for the Biology Test Forms	50
Figure 14:	Test Characteristic Curves for the Chemistry Test Forms	51
Figure 15:	Test Characteristic Curves for the Physics Test Forms	52

# Chapter One: Introduction

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

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

## 1.1 Universal Participation

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

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

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

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

In 2001, the reauthorization of the Elementary and Secondary Education Act (ESEA) ushered in a new era of accountability at the federal level as well. Popularly referred to as No Child Left Behind (NCLB), this law was designed to improve American education by ensuring that even the neediest students receive a sound basic education and that no child is trapped in a failing school. The cornerstones of NCLB include annual testing of all students in language and

mathematics in grades 3 through 8; annual testing of all students in language and math once in high school; and annual testing of all students in science in each grade span 3–5, 6–9, and 10–12. These assessment results are to be broken out (disaggregated) by ethnic, disability, poverty, and English proficiency. The end goal of NCLB is to have all students performing at a level deemed proficient, by 2014. A major provision of the act focuses on accountability for results.

*H.R. 1 will result in the creation of assessments in each state that measure what children know and learn in reading and math in grades 3-8. Student progress and achievement will be measured according to tests that will be given to every child, every year. ...*

*Statewide reports will include performance data disaggregated according to race, gender, and other criteria to demonstrate not only how well students are achieving overall but also progress in closing the achievement gap between disadvantaged students and other groups of students.*

From: Fact Sheet on the Major Provisions of the Conference Report to H.R. 1, the No Child Left Behind Act

## **1.2 The North Carolina Testing Program**

The North Carolina Testing Program was designed to measure the extent to which students satisfy academic performance requirements. Tests developed by the North Carolina Department of Public Instruction's Test Development Section, when properly administered and interpreted, provide reliable and valid information that enables

- *students to know the extent to which they have mastered expected knowledge and skills and how they compare to others;*
- *parents to know if their children are acquiring the knowledge and skills needed to succeed in a highly competitive job market;*
- *teachers to know if their students have mastered grade-level knowledge and skills in the curriculum and, if not, what weaknesses need to be addressed;*
- *community leaders and lawmakers to know if students in North Carolina schools are improving their performance over time and how our students compare with students from other states or the nation; and*
- *citizens to assess the performance of the public schools (North Carolina Testing Code of Ethics, 1997, revised 2000).*

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

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

---

---

Tests included in the North Carolina Testing Program are designed for use as federal, state, and local indicators of student performance. Interpretation of test scores in the North Carolina Testing Program provides information about a student's performance on the test in percentiles, scale scores, and achievement levels. Percentiles provide an indicator of how a child performs relative to other children who took the test in the norming year, or the first year the test was administered. Percentiles range from 1 to 99. A percentile rank of 65 indicates that a child performed equal to or better than 65% of the children who took the test during the norming year.

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

Science End-of-Course tests are administered in Physical Science, Biology, Chemistry, and Physics. The policy-level generic achievement level descriptors for End-of-Course tests are given below:

**Level I:** Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area.

**Level II:** Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area.

**Level III:** Students performing at this level consistently demonstrate mastery of the course subject matter and are well prepared for a more advanced level in the content area.

**Level IV:** Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area.

The content-specific performance-level descriptors are provided for each assessment as a part of the test specifications in Appendix A.

The North Carolina End-of-Grade (EOG) Tests include multiple-choice assessments of reading comprehension in grades 3 through 8; mathematics in grades 3 through 8 and 10 (the grade 10 assessment is only for students in Title I schools who have not fulfilled the Algebra I requirement by the 10<sup>th</sup> grade); and science in grades 5 and 8. There is also a pretest administered at the beginning of the 3<sup>rd</sup> grade to measure baseline performance in reading comprehension and mathematics. The North Carolina End-of-Course (EOC) Tests include multiple-choice assessments of composition and literary analysis in English I; and mathematics and mathematical reasoning in Algebra I, Geometry, and Algebra II. In addition to the English and mathematics tests, the North Carolina Testing Program includes science EOC tests in

Biology, Chemistry, Physical Science, and Physics; Social Studies EOC tests in Civics and Economics and U.S. History; writing assessments in grades 4, 7, and 10; the North Carolina Tests of Computer Skills; and alternate and alternative assessments developed to validly measure student abilities in populations who are not able to access the general assessments even with accommodations.

The End-of-Grade tests in grades 3 through 8 mathematics, 3 through 8 reading comprehension, and 5 and 8 science are used for determining AYP at the elementary and middle school levels. At the high school level, the End-of-Course tests in English I, Algebra I, and Biology, and the grade 10 Writing assessment, are used for determining AYP. For students who are not able to access the general assessments, the corresponding alternate or alternative assessment is used.

In 2006, the North Carolina State Board of Education approved new graduation standards. These standards require that

*Effective with the class entering ninth grade for the first time in the 2006-2007 school year, students who are following the career preparation, college technical preparation, or college/university preparation courses of study shall meet the following exit standards:*

*(A) successfully complete a senior project that is developed, monitored, and scored within the LEA using state-adopted rubrics; and*

*(B) score at proficiency level III or above on the end-of-course assessment for English I, U.S. History, Biology, Civics and Economics, and Algebra I.*

*(16 NCAC 6D .0503 State graduation requirements, section E subsection 2).*

The Grade Level Proficiency Guidelines, approved by the State Board of Education (February, 1995), established Level III (of those achievement levels listed above) as the standard for each grade level. The EOC tests measure a student's mastery of course-level material. Scale scores for end-of-grade tests use a developmental (vertical) scale.

### **1.3 The North Carolina Science Tests**

This Technical Report for the North Carolina Science Tests discusses tests aligned with the North Carolina Science 2004 *Standard Course of Study (SCS)*. Following a five-year revision cycle, the North Carolina State Board of Education adopted the Science SCS in 2004 to replace the 1999 SCS. The End-of-Course Tests in Physical Science, Biology, Chemistry, and Physics were administered as field tests in school year 2006-2007 and were administered operationally for the first time in school year 2007-2008.

The purpose of this document is to provide an overview of and technical documentation for the North Carolina EOC Science Tests. Chapter One provides an overview of the North Carolina Science Tests. Chapter Two describes the test development process. Chapter Three outlines the test administration. Chapter Four describes the construction of the developmental scale, the scoring of the tests, and the standard setting process. Chapter Five provides an outline of reporting of test results. Chapters Six and Seven provide the technical properties of the tests

such as descriptive statistics from the first operational year, reliability indices, and evidence of validity. Chapter Eight is an overview of quality control procedures.

# **Chapter Two: Test Development Process**

## **2.1 Test Development Process for the North Carolina Testing Program**

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

**Figure 1:** Flow chart of the test development process used in development of North Carolina Tests

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

<sup>a</sup>Activities done only at implementation of new curriculum

<sup>b</sup>Activities involving NC teachers

Phase 1 (step 1) requires 4 months

Phase 2 (steps 2-7) requires 12 months

Phase 3 (steps 8-14) requires 20 months

Phase 4 (steps 15-20) requires 4 months for EOC and 9 months for EOG

Phase 5 (step 21) requires 4 months

Phase 6 (step 22) requires 1 month

TOTAL 44-49 months

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

## 2.2 The Curriculum Connection

North Carolina wants its students to graduate with the skills necessary to compete in the global marketplace, to be prepared for further education, and to participate effectively as citizens.

The previous revision to the science North Carolina *Standard Course of Study* (NCSCS) was 1999. Following the North Carolina five-year revision cycle, the 2004 revisions reflects “the development of National Science Education Standards better. The 2004 revision further reflects the recommendations of the Third International Mathematics and Science Study (TIMSS) and the 1996 National Assessment of Educational Progress (NAEP) science framework and assessment. The SCS has been written to expand the intent of previous documents and represents an evolutionary process of curriculum refinement” (pg 6). The *Standard Course of Study* is available at <http://www.ncpublicschools.org/curriculum/science/scos/>

The North Carolina Science *Standard Course of Study* clearly defines a curriculum focused on what students will need to know and be able to do to be successful and contributing citizens in our state and nation in the years ahead. As defined in the 2004 North Carolina Science *Standard Course of Study*, the goals of science education are for students to develop science literacy as defined by National Science Education Standards as "the knowledge and understanding of scientific concepts and processes required for scientific decision making, participation in civic and cultural affairs, and economic productivity." (p. 22)

Testing of North Carolina students’ skills relative to the competency goals and objectives in the *Standard Course of Study* (SCS) is one component of the North Carolina Testing Program. At the High School level, students are tested in Science at the end of courses in Biology, Chemistry, Physical Science, and Physics.

Each item on the End-of-Course Science Tests is aligned to an objective from the NC SCS for Science for the applicable course. While some objectives can be measured readily by multiple-choice questions and are assessed by the tests, other objectives address the skills and background knowledge that are needed to do well on the tests, but are not easily measured in a multiple-choice format.

## 2.3 Test Specifications

Delineating the purpose of a test must come before the test design. A clear statement of purpose provides the overall framework for test specifications, test blueprint, item development, tryout, and review. A clear statement of test purpose also contributes significantly to appropriate test use in practical contexts (Millman & Greene, 1993). The tests in the North Carolina Testing Program are designed in alignment with the NCSCS. The purpose of the North Carolina EOC Tests of Science is legislated by General Statute *115C-174.10* and focuses on the measurement of individual student science skills and knowledge as outlined in the NCSCS.

Test specifications for the North Carolina science tests are developed in accordance with the competency goals and objectives specified in the NCSCS. A summary of the test specifications

is provided in Appendix A. These test specifications also are generally designed to include the following:

- (1) percentage of questions from higher or lower thinking skills and classification of each test question into level of difficulty;
- (2) percentage of test questions that measure a specific goal, and rank order of emphasis for objectives within a goal;

Test blueprints, specific layouts or “road maps” to ensure the parallel construction of multiple test forms, were developed from the test specifications. These blueprints identify the exact numbers of items from each objective that are used in the creation of the test forms. At the objective level, the tests are comprised of items that are a random domain sample from the superordinate goal, and as such there may be more than one layout. However, at the goal level and in terms of the relative emphasis of the objective coverage, all test blueprints conform to the test specifications.

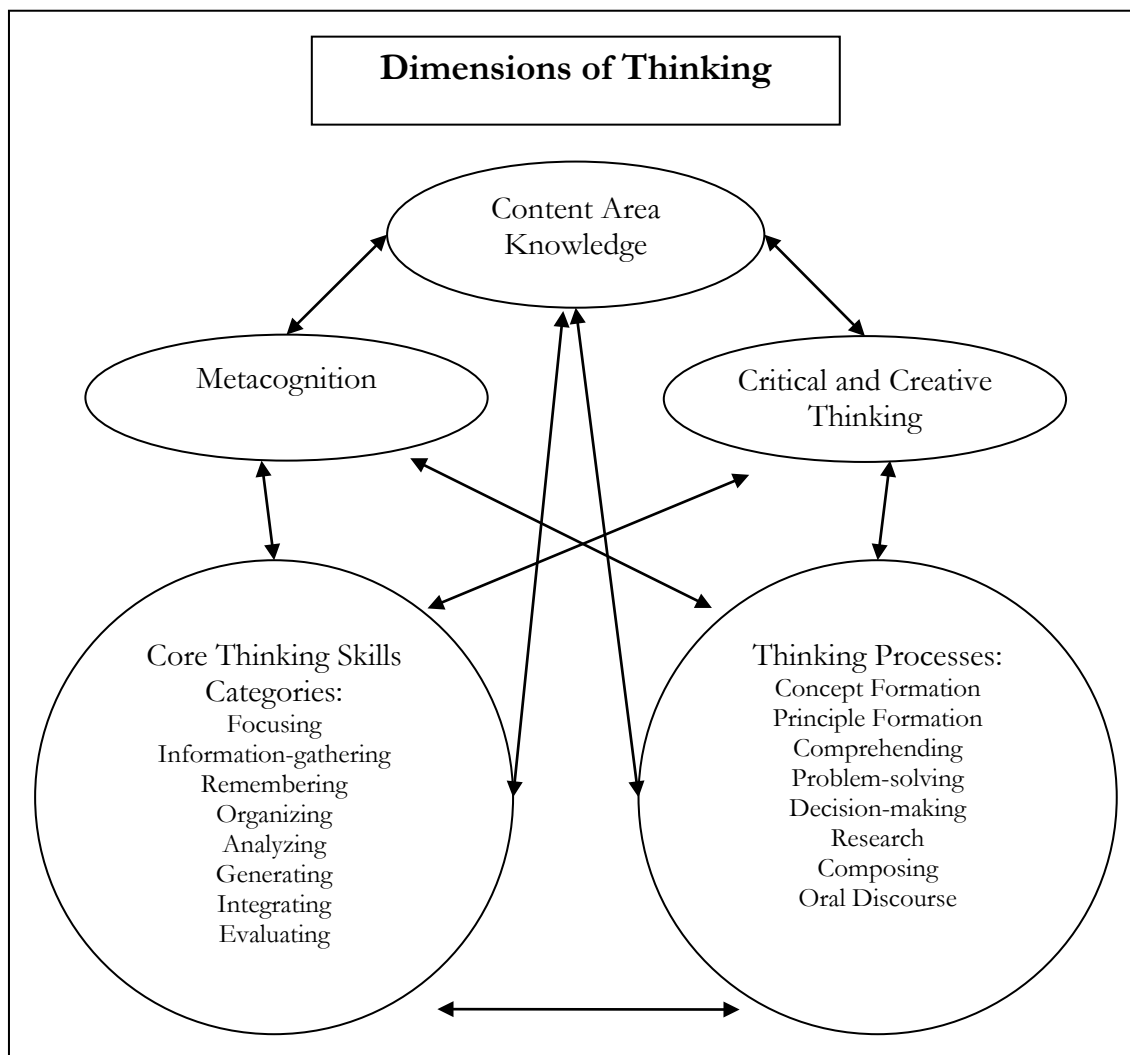
## **2.4 Item Development**

Each item is written to be aligned with a specific objective in the NCSCS. Items on the North Carolina EOC Tests of Science are developed using level of difficulty and thinking skill level. Item writers use these frameworks when developing items. The purpose of the categories is to ensure a balance of difficulty as well as a balance across the different cognitive levels among the items in the North Carolina Science tests.

For the purposes of guiding item writers to provide a variety of items, item writers were instructed to classify the items into three levels of difficulty: easy, medium, and hard. Easy items are those items that the item writer believes can be answered correctly by approximately 70% of the examinees. Medium items can be answered correctly by 50%–60% of the examinees. Difficult items can be answered correctly by approximately 30% of the examinees. The item writers were further instructed to write approximately 25% of their items at the hard level, 25% at the easy level, and the remaining 50% at the medium level of difficulty. These targets are used for item pool development to ensure an adequate range of difficulty.

A more recent consideration for item development is the classification of items by thinking skill level, the cognitive skills that an examinee must use to solve a problem or answer a test question. Thinking skill levels are based on an adaptation of *Dimensions of Thinking* by Marzano et al. (1988). Thinking skill levels, in addition to their usefulness in framing achievement tests, also provide a practical framework for curriculum development, instruction, assessment, and staff development. Thinking skills begin with the basic skill of remembering and move to more complex thinking skills, such as analysis, integration, and evaluation. Figure 2 below shows a visual representation of the framework.

**Figure 2:** Thinking skills framework used to develop the North Carolina Science Tests (adapted from Marzano et al., 1988)



## 2.5 Item Format

Items on the North Carolina science tests are four-foil multiple-choice items. Apart from what may be presented in an item, students do not have rulers, protractors, or formulas available for the science EOC Biology tests. Formulas are provided for the EOC Physical Science, Chemistry, and Physics tests.

## 2.6 Selection and Training of Item Writers

Once the test blueprints were finalized from the test specifications for the revised edition of the North Carolina Science tests, North Carolina educators were recruited and trained to write new items for the state tests. The diversity among the item writers and their knowledge of the current NCSCS was addressed during recruitment. The use of North Carolina educators to develop items strengthened the instructional validity of the items.

Potential item writers received training and materials designed in accordance with the Science curriculum, which included information on content and procedural guidelines as well as information on stem and foil development. The item-writing guidelines are included in Appendix B. The items developed during the training were evaluated by content specialists, who then provided feedback to the item writers on the quality of their items.

## 2.7 Reviewing Items for Field Testing

To ensure that an item was developed to NCSCS standards, each item went through a detailed review process prior to being placed on a field test. This review is represented by Step 9 on the Test Development Flow Chart (Figure 1). A new group of North Carolina educators was recruited to review items. Once items had been through an educator review, test development staff members, with input from curriculum specialists, reviewed each item. Items were also reviewed by educators and/or staff familiar with the needs of students with disabilities and limited English proficiency.

The criteria for evaluating each written item included the following:

- 1) Conceptual
  - objective match (curricular appropriateness)
  - thinking skill match
  - fair representation
  - lack of bias or sensitivity
  - clear statement
  - single problem
  - one best answer
  - common context in foils
  - credible foils
  - technical correctness
  
- 2) Language
  - appropriate for age
  - correct punctuation
  - spelling and grammar
  - lack of excess words
  - no stem or foil clues
  - no negative in foils (unless it fits the objective)

- 3) Format
  - logical order of foils
  - familiar presentation style, print size, and type
  - correct mechanics and appearance
  - equal/balanced length foils
  
- 4) Diagram/Graphics
  - necessary
  - clean
  - relevant
  - unbiased

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

## 2.8 Assembling Field Test Forms

Prior to creating an operational test, items for each course area were assembled into field test forms. Field test forms were organized according to the blueprints for the operational tests. All forms were administered as stand-alone field tests. All items were aligned with the 2004 North Carolina *Standard Course of Study (SCS)* content standards. Prior to field test administration, North Carolina educators reviewed the assembled field test forms for clarity, correctness, potential bias or sensitivity, cuing of items, and curricular appropriateness, following steps similar to operational test review.

The initial round of field tests for the Edition 3 Science consisted in stand-alone, rather than embedded field tests (see table 1). Because the 2004 *SCS* for Physical Science, Biology, Chemistry, and Physics was first implemented instructionally in academic year 2006-07, field testing these four EOC subjects occurred in that academic year. As some high schools provide instruction for those courses on a “block schedule” testing occurred in both fall 2006 and spring 2007.

**Table 1:** Number of items field tested for North Carolina EOC Tests of Science, 3<sup>rd</sup> Edition (Stand-alone field-testing 2006-07; Tests aligned with the 2004 *North Carolina Standard Course of Study* for Science)

<b>Grade or Course</b>	<b>Administration(s)</b>	<b>Number of Forms</b>	<b>Number of Items per Form</b>	<b>Total Number of Items</b>
Physical Science	Fall 2006 and Spring 2007	12	92	1104
Biology	Fall 2006 and Spring 2007	11*	92	1012
Chemistry	Fall 2006 and Spring 2007	12	92	1104
Physics	Fall 2006 and Spring 2007	6	92	552

\* An additional Biology test form, not reflected in this count, was a 2<sup>nd</sup> edition test form administered for moderating and cross-edition equating purposes.

## 2.9 Sampling Procedures and Field Test Sample Characteristics

Sampling for stand-alone field testing of the North Carolina Tests is typically accomplished using stratified random sampling of schools with the goal being a selection of students that is representative of the entire student population in North Carolina. Stratifying variables include

- gender
- ethnicity
- region of the state
- free/reduced lunch
- students with disabilities
- students with limited English proficiency
- previous year's test scores

Table 2 shows the demographic characteristics of the sample for the stand-alone field tests of the Edition 3 science tests.

Beginning with the first operational version of the science tests, field test items are embedded within each form to supplement the item pools. Embedded field test items are grouped into sections. Experimental sections are placed in operational forms, and the operational forms are spiraled within a classroom to obtain a randomly equivalent group of examinees on each form. This results in a demographic distribution nearly identical to that of the full population.

**Table 2:** Field test population (2006-07) for EOC Tests of Science, 3<sup>rd</sup> Edition

Course	N	Gender		Ethnicity						English Language Proficiency Status
		% Male	% Female	% Asian	% Black	% Hispanic	% American Indian	% Multiracial	% White	% LEP (Limited English Proficiency)
Physical Science	54,912	51.4	48.6	1.7	32.2	7.0	2.6	3.2	53.3	3.8
Biology	22,389	49.5	50.5	2.1	27.2	6.5	0.7	3.2	60.3	3.0
Chemistry	42,791	44.1	55.9	3.6	20.3	3.9	1.0	2.9	68.4	1.2
Physics	8,596	59.3	40.7	5.9	13.7	3.3	0.6	1.9	74.5	0.5

Notes: The Physics field tests were administered via computer-based testing. The Physical Science, Chemistry, and Physics field test administrations were census field tests. For Biology, field testing was accomplished by drawing demographically representative samples. The percentages for demographic categories are for all examinees with available demographic data.

## 2.10 Item Analysis

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

Field test data were analyzed by the North Carolina Department of Public Instruction (NCDPI) psychometric staff. Item statistics and descriptive information were then included on the item record for each item. The item records contain the statistical, descriptive, and historical information for an item, a copy of the item as it was field tested, comments by reviewers, and curricular and psychometric notations.

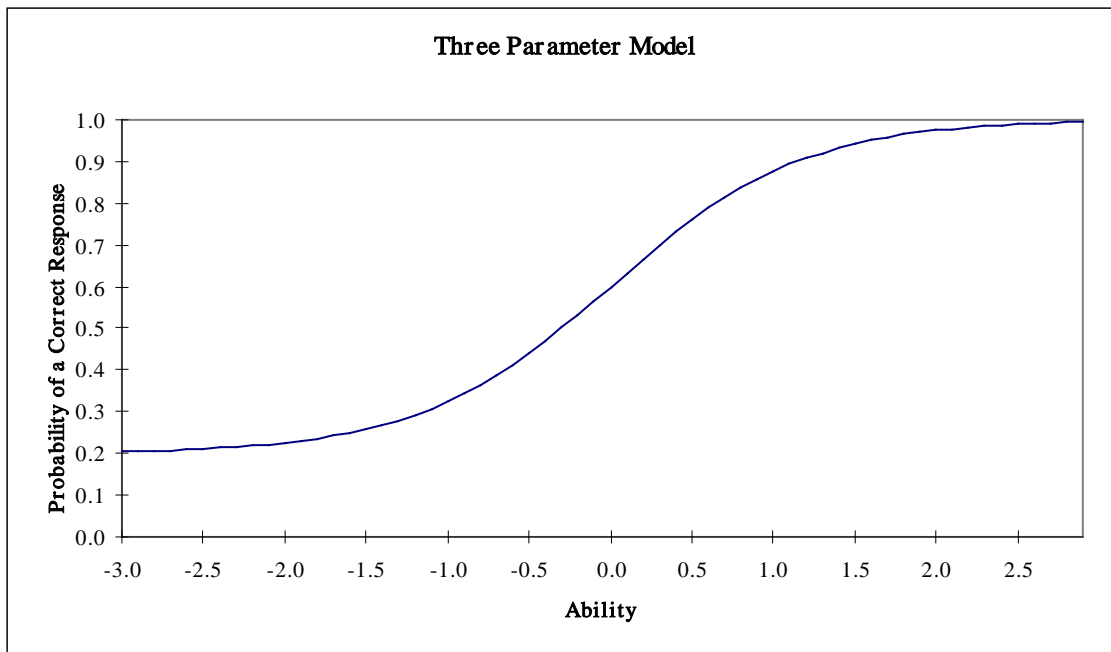
## 2.11 Classical Measurement Analysis

For each item, the p-value (proportion of examinees answering an item correctly), 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. While the p-value is an important statistic and one component used in determining the selection of an item, the North Carolina Testing Program also uses IRT to provide additional item parameters to determine the psychometric properties of the North Carolina Science tests.

## 2.12 Item Response Theory (IRT) Analysis

To provide additional information about item performance, the North Carolina Testing Program also uses IRT statistics to determine whether an item should be included on the test. IRT is, with increasing frequency, being used with large-scale achievement testing. "The reason for this may be the desire for item statistics to be independent of a particular group and for scores describing examinee proficiency to be independent of test difficulty, and for the need to assess reliability of tests without the tests being strictly parallel" (Hambleton, 1983, p. 148). IRT meets these needs and provides two additional advantages: the *invariance of item parameters* and the *invariance of ability parameters*. Regardless of the distribution of the sample, the parameter estimates will be linearly related to the parameters estimated with some other sample drawn from the same population. IRT allows the comparison of two students' ability estimates even though they may have taken different items. An important characteristic of IRT is item-level orientation. IRT makes a statement about the relationship between the probability of answering an item correctly and the student's ability or the student's level of achievement. The relationship between a student's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an Item Characteristic Curve (ICC). This function specifies that as the level of the trait increases, the probability of a correct response to an item increases. The following figure shows the ICC for a typical 4-option multiple-choice item.

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



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

- (1) unidimensionality—only one ability is assessed by the set of items (for example, a spelling test only assesses a student's ability to spell);
- (2) local independence—when abilities influencing test performance are held constant, an examinee's responses to any pair of items are statistically independent (conditional independence, i.e., the only reason an examinee scores similarly on several items is because of his or her ability); and
- (3) the ICC specified reflects the true relationship among the unobservable variable (ability) and the observable variable (item response).

The formula for the 3PL model is

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

where

- $P_i(\theta)$ —the probability that a randomly chosen examinee with ability ( $\theta$ ) 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 “difficulty parameter,” the point on the ability scale where the probability of a correct response is 50% **when  $c = 0$**  (the threshold of a typical item is 0.00)
- $c$ —the asymptote, or “guessing parameter”, the proportion of the examinees who got the item correct but did poorly on the overall test (the [theoretical] asymptote of a typical 4-choice item is 0.25)
- $D$ —a scaling factor, 1.7, to make the logistic function as close as possible to the normal ogive function (Hambleton, 1983, p.125).

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

### 2.13 Differential Item Functioning Analysis

It is important to know the extent to which an item on a test performs differently for different students. As a third component of the item analysis, differential item functioning (DIF) analyses examine the relationship between the score on an item and group membership, while controlling for ability, to determine if an item may be behaving differently for a particular gender or ethnic group. While the presence or absence of true bias is a qualitative decision, based on the content of the item and the curriculum context within which it appears, DIF can be used to quantitatively identify items that should be subjected to further scrutiny.

In developing the North Carolina Science tests, the North Carolina Testing Program staff used the Mantel-Haenszel procedure to examine DIF by examining  $j \times 2 \times 2$  contingency tables, where  $j$  is the number of different levels of ability actually achieved by the examinees (actual total scores received on the test). The focal group is the focus of interest, and the reference group serves as a basis for comparison for the focal group (Dorans & Holland, 1993; Camilli & Shepherd, 1994). For example, females might serve as the focal group and males might serve as the reference group to determine if an item may be biased toward or against females.

The Mantel-Haenszel (MH) chi-square statistic (only used for  $2 \times 2$  tables) tests the alternative hypothesis that a linear association exists between the row variable (score on the item) and the column variable (group membership). The  $\chi^2$  distribution has one degree of freedom (df) and its significance is determined by the correlation between the row variable and the column variable (SAS Institute, 1985). The MH Log Odds Ratio statistic in SAS was used to determine the direction of DIF. This measure was obtained by combining the odds ratios ( $a_j$ ) across levels with the formula for weighted averages (Camilli & Shepherd, 1994, p. 110).

For the Mantel-Haenszel statistic, the null hypothesis is that there is no relationship between score and group membership: the odds of getting the item correct are equal for the two groups. The null hypothesis was not rejected when the odds ratio equaled 1. For odds ratios greater than

1, the interpretation was that an individual at score level  $j$  of the Reference Group had 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 was that an individual at score level  $j$  of the Focal Group had a greater chance of answering the item correctly than an individual at score level  $j$  of the Reference Group. The Breslow-Day Test was used to test whether the odds ratios from the  $j$  levels of the score were all equal. When the null hypothesis was true, the statistic was distributed approximately as a chi-square with  $j-1$  degrees of freedom (SAS Institute, 1985).

The ethnic (Black/White) and gender (Male/Female) 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.

## **2.14 Expert Review**

All items, statistics, and comments were reviewed by curriculum specialists and testing consultants. Items found to be inappropriate for curricular or psychometric reasons were deleted. In addition, items flagged for exhibiting ethnic or gender DIF were then reviewed by a bias review committee. Differential item functioning is a purely statistical judgment without regard to the actual content of the item; the determination of actual bias is a qualitative judgment based on the content of the item.

The bias review committee members, selected because of their knowledge of the curriculum area and their diversity, evaluated test items with a DIF flag using the following questions:

1. Does the item contain language that is not commonly used statewide or has different connotations in different parts of the state or in different cultural or gender groups?
2. Does the item contain any local references that are not a part of the statewide curriculum?
3. Does the item portray anyone in a stereotypical manner? (These could include activities, occupations, or emotions.)
4. Does the item contain any demeaning or offensive materials?
5. Does the item have offensive, stereotyping, derogatory, or proselytizing religious references?
6. Does the item assume that all students come from the same socioeconomic background? (e.g., a suburban home with two-car garage)
7. Does the artwork adequately reflect the diversity of the student population?
8. Are there other bias or sensitivity concerns?

An answer of yes to any of these questions resulted in the unique item production number being recorded on an item bias sheet along with the nature of the bias or sensitivity. Items that were consistently identified as exhibiting bias or sensitivity were flagged for further review by the NCDPI curriculum specialists.

Items that were flagged by the bias review committee were then reviewed by the NCDPI curriculum specialists. If these experts found the items measured content that was expected to

be mastered by all students, the item was retained for test development. Items that were determined by both review committees to exhibit true bias were deleted from the item pool.

## 2.15 Criteria for Inclusion in Item Pool

All of the item parameter data generated from the above analyses were used to determine if an item displayed sound psychometric properties. Items could be potentially be flagged as exhibiting psychometric problems or bias due to ethnicity/race or gender according to the following criteria.

Items with these characteristics were deleted:

- weak discrimination—the slope ( $a$  parameter) was less than 0.50
- low correlation with total score—the item correlation (r-biserial) was less than 0.15
- guessing—the asymptote ( $c$  parameter) was greater than 0.45
- too difficult—the threshold ( $b$  parameter) was greater than 3.0 or the p-value was less than 0.10

Items with these characteristics were used sparingly (held in reserve):

- weak discrimination—the slope ( $a$  parameter) was between 0.70 and 0.50
- low correlation with total score—the item correlation (r-biserial) was between 0.25 and 0.15
- guessing—the asymptote ( $c$  parameter) was between 0.35 and 0.45
- too difficult—the threshold ( $b$  parameter) was between 2.5 and 3.0; or the p-value between 0.10 and 0.15
- too easy—the threshold ( $b$  parameter) was between  $\bar{2}$ .5 and  $\bar{3}$ .0; or the p-value was between 0.85 and 0.90.

Items with these characteristics underwent additional reviews:

- ethnic bias—the log odds ratio was greater than 1.50 (favored whites) or less than 0.67 (favored blacks)
- gender bias—the log odds ratio was greater than 1.50 (favored females) or less than 0.67 (favored males)

Items with threshold less than  $\bar{3}$ .0 or p-value greater than 0.90, provided all other statistical and content information supported keeping the item, were submitted for consideration in an alternative assessment targeted toward students with persistent academic disabilities.

The average item pool parameter estimates based on field test data are provided in the next section.

## 2.16 Item Pool Parameter Estimates

See Table 3 below.

**Table 3:** Average item pool parameter estimates for EOC Tests of Science by course (statistics from 2006-07 stand-alone field-testing ).

Course	Biserial Correlation	P-value	IRT Parameters			DIF (Odds-Ratio Logit)	
			Threshold <i>d</i> ( <i>b</i> )	Slope ( <i>a</i> )	Asymptote ( <i>c</i> )	Ethnicity	Gender
Physical Science	0.404	0.4377	0.915	1.036	0.210	1.026	1.000
Biology	0.44	0.5035	0.591	1.018	0.221	1.044	1.007
Chemistry	0.404	0.4595	0.865	0.98	0.214	1.004	1.002
Physics	0.434	0.4880	0.698	.974	0.193	1.056	1.028

**Notes:** The item pool averages shown in this table are for all items from the field-testing that upon post-field-test review of content and psychometric properties were retained as candidates for potential use on operational test forms.

## 2.17 Operational Test Construction

The final item pool was based on approval by content and curriculum experts for curricular match and testing experts and psychometricians for psychometrically sound item performance. Once the final items were identified for the item pool, operational tests were constructed according to the test blueprints. For a summary of the test specifications, see Appendix B.

## 2.18 Establishing the Target p-value for Operational Tests

P-value is a measure of the difficulty of an item. P-values can range from 0 to 1. The letter “p” symbolizes the proportion of examinees that answer an item correctly. So an item with a p-value of 0.75 was correctly answered by 75% of the students who answered the item during the field test, and one might expect that roughly 75 of 100 examinees will answer it correctly when the item is put on an operational test. An easy item has a p-value that is high, which means that a large proportion of the examinees got the item right during the field test. A difficult item has a low p-value, meaning that few examinees answered the item correctly during field testing. Note that items usually have higher p-values on the operational form than on stand-alone field tests, due to factors which may include higher motivation on the operational test, which has stakes for the student; increased or improved background preparation in earlier grades as the curriculum is implemented; and/or improved instruction in the content in the second and subsequent years of a new curriculum.

The NCDPI psychometric staff must choose a target p-value for each operational test prior to assembling the tests. Ideally, the average p-value of a test would be 0.625, which is the theoretical average of a student getting 100% correct on the test and a student scoring a chance performance (25% for a 4-foil multiple-choice test). That is  $(100 + 25)/2$ .

The actual target was chosen by first looking at the distribution of the p-values for a particular item pool. While the goal is to set the target as close to 0.625 as possible, it is often the case that the target

p-value is set between the ideal 0.625 and the average p-value of the item pool. The average p-value and the target p-value for operational forms are provided below for comparison.

## 2.19 Comparison of Item Pool p-Values with Operational p-Values

**Table 4:** Comparison of p-value of item pool with p-values of assembled forms and operational p-values

Subject	p-Value of Item Pool*	p-Value of Forms*	Operational p-Values*
Physical Science	0.4374	.4426	.5837
Biology	0.5088	.5052	.6005
Chemistry	0.4648	.4711	.6136
Physics	0.4777	.4910	.6305

\* Initial p-values are from 2006-07 stand-alone field testing. Operational p-values are from the 2007-08 administrations.

To develop equivalent forms, the test forms were balanced on P+, the sum of the p-values of the items. The sections also have matching or highly similar profiles in terms of numbers of items addressing higher and lower thinking skills and numbers of items categorized as easy, medium, or hard. Finally, to the extent possible, the sections were balanced on slope. Although all form-level values are reported as an average across forms, actual P+ differences between assembled forms within the same course were less than 0.01.

Because of the concerns about student motivation and opportunity to learn on the stand-alone field tests, p-values from the first operational administrations of the tests were also calculated and are included here.

## 2.20 Review of Assembled Operational Tests

Once forms were assembled to meet test specifications, target P+-values, and item parameter targets, a group of North Carolina educators and curriculum supervisors then reviewed the assembled forms. Each group of subject area teachers and curriculum supervisors worked independently of the test developers. The criteria for evaluating each group of forms included the following:

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

Reviewers were instructed to take the tests (circling the correct responses in the booklet as well as recording their responses on a separate sheet) and to provide comments and feedback next to each item. After reviewing all the forms, each reviewer independently completed a survey asking for his or her opinion as to how well the tests met the five criteria listed above. During the last part of the session, the group discussed the tests and made comments as a group. The test review ratings along with the comments were aggregated for review by test development staff and consultants. Items that were determined to be problematic at this point were replaced, and the forms rebalanced. Items may have been removed from a form because of cuing, overemphasis on a particular subtopic, or for maintaining statistical equivalency. If a form has more than 10% of its items replaced as a result of this process, the NCDPI psychometric policy is to send the form through review again, as it is no longer really the same form that was reviewed previously. No test forms exceeded this criterion. As a final review, test development staff members, with input from curriculum staff, content experts, and editors, conducted a final psychometric, content, and grammar check for each test form.

## 2.21 Establishing the Test Administration Time

Additional important considerations in the construction of the North Carolina Science tests were the number of items to be included and the time necessary to complete the test. Since the tests are power tests, requiring higher-level thinking for many items, students were provided with ample time to complete the test. The Test Administration Manual provided test administrators with suggested times, based on the times of 95% of the students finishing the stand-alone field test. See Table 5 below for suggested time on testing (exclusive of distributing materials, reading directions, and so forth).

Through the 2006-2007 school year, students who were working productively were allowed as much time as they needed to complete the test. Beginning with the 2007-2008 school year, the maximum time allowed for regular students on the End-of-Course tests in Physical Science, Biology, Chemistry, and Physics was four hours. This change was enacted after several accounts of test administrations that exceeded a normal school day.

Any student with documented special needs requiring accommodations, such as *Scheduled Extended Time*, of course may exceed these maximum times. Students requiring time beyond the suggested time in the manuals continue to receive 3-minute breaks after every hour of testing.

**Table 5:** Number of items per test and time allotted by subject

Subject	Number of Items*	Suggested Time in Minutes**
Physical Science	80	150
Biology	80	150
Chemistry	80	150
Physics	84	150

\* includes embedded field-test items

\*\* includes only the estimated testing time; does not include time allotted for instructions, breaks, etc. (which increases suggested time to 164 minutes)

# Chapter Three: Test Administration

## 3.1 Test Administration

The purpose of end-of-course tests is to sample a student's knowledge of subject-related concepts specified in the North Carolina *Standard Course of Study* and to provide a global estimate of the student's mastery of the material in a particular content area. The Science end-of-course tests were developed to provide accurate measurement of individual student knowledge and skills specified in the science component of the North Carolina *Standard Course of Study*. Effective with the 2007-2008 school year, the North Carolina End-of-Course Science Tests are multiple-choice tests that measure the goals and objectives of the science curriculum adopted in 2004 by the North Carolina State Board of Education for each course. The competency goals and objectives are organized into four strands: (1) Nature of Science, Science as Inquiry, (2) Science and Technology, (3) Science in Personal, and (4) Social Perspectives.

In schools that follow a traditional calendar, all end-of-course tests are administered within the final 10 days of the course to students enrolled for credit in courses where end-of-course tests are required. For schools which operate under a "block" or semester schedule the tests are administered in the last five days of the course.

## 3.2 Training for Test Administrators

The North Carolina Testing Program uses a train-the-trainer model to prepare test administrators to administer North Carolina tests. Regional accountability coordinators (RACs) receive training in test administration from the NCDPI Testing Policy and Operations staff at regularly scheduled monthly training sessions. Subsequently, the RACs provide training on conducting a proper test administration to local education agency (LEA) test coordinators. LEA test coordinators provide training to school test coordinators. The training includes information on the test administrators' responsibilities, proctors' responsibilities, preparing students for testing, eligibility for testing, policies for testing students with special needs (students with disabilities and students with limited English proficiency), accommodated test administrations, test security (storing, inventorying, and returning test materials), and the *Testing Code of Ethics*.

## 3.3 Preparation for Test Administration

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

Test Administrator’s Manual to ensure a standardized administration, and must read aloud all directions and information to students as indicated in the manual.

### **3.4 Test Security and Handling Materials**

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

*school systems shall (1) account to the department (NCDPI) for all tests received; (2) provide a locked storage area for all tests received; (3) prohibit the reproduction of all or any part of the tests; and (4) prohibit their employees from disclosing the content of or discussing with students or others specific items contained in the tests. Secure test materials may only be stored at each individual school for a short period prior to and after the test administration. Every effort must be made to minimize school personnel access to secure state tests prior to and after each test administration.*

At the individual school, the principal shall account for all test materials received. As established by APA 16 NCAC 6D .0306, the principal shall store test materials in a secure, locked area except when in use. The principal shall establish a procedure to have test materials distributed immediately prior to each test administration. Before each test administration, the building level coordinator shall collect, count, and return all test materials to the secure, locked storage area. Any discrepancies are to be reported to the school system test coordinator immediately and a report must be filed with the regional accountability coordinator.

### **3.5 Student Participation**

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

#### *Physical Science, Biology, Chemistry, and Physics End-of-Course Tests*

All students, including students with disabilities, enrolled in a yearlong (i.e., traditional calendar) course for credit must be administered the end-of-course test, which may be a corresponding alternate or alternative assessment if so indicated by the student’s IEP or LEP documentation, in the final 10 days of the course. In schools which operate on a “block” or semester schedule, all students, including students with disabilities, who are enrolled in a course for credit must be administered the EOC test in the final five days of the course. Students enrolled for credit in a course that has an end-of-course test must be administered the EOC test. Students who are repeating the course for credit must also be administered the EOC test. The student’s most recent test score will be used for the purpose of state accountability. In addition, starting with the 2001-2002 school year, LEAs shall use results from all multiple-

choice EOC tests as at least twenty-five percent of the student’s final grade for each respective course. LEAs shall adopt policies regarding the use of EOC test results in assigning final grades.

In 2006, the NC State Board of Education revised policy HSP-N-004 (16 NCAC 6D.0503): students entering the ninth grade for the first time in 2006-07 and beyond are now required to perform at Achievement Level III (with one standard error of measurement) or above on five required end-of-course (EOC) assessments, including Biology, in order to graduate. Multiple retest opportunities are available; however, the first test score is used for the purpose of AYP and federal accountability.

### 3.6 Alternate and Alternative Assessments

The North Carolina End-of-Course Testing Program currently offers the North Carolina Checklist of Academic Skills (NCCLAS) and the *NCEXTEND2* tests as options for meeting the assessment requirements at the state and federal levels. The chart below shows which end-of-course science assessments are available.

**Table 6:** Available assessments in the North Carolina EOC Science Testing Program

Subject	General		Modified Format	Modified Achievement Standards
	Without Accommodations	with Accommodations	NCCLAS	<i>NCEXTEND2</i>
Physical Science	X	X	X	
Biology	X	X	X	
Chemistry	X	X	X	
Physics	X	X	X	
OCS*				X

\*The Occupational Course of Study (OCS) is followed by high school students with disabilities for whom the general curriculum is not accessible. The OCS Extend2 Test of Life Skills Science assesses mastery of Life Skills Science Courses I and II within this course of study. As of 2008-09 the OCS test is not used for AYP.

The NCCLAS is an assessment process in which teachers utilize a checklist to evaluate student performance on curriculum benchmarks in the areas of reading, mathematics, and/or writing. Student performance data are provided to the NCDPI at the end of the school year (summative), although teachers gather evidence throughout the year. The NCCLAS measures competencies on the North Carolina *Standard Course of Study*. The Individualized Education Program (IEP) team determines if a student, due to the nature of his/her special needs, is eligible to participate in the NCCLAS. Typically, students who are being assessed on the NCCLAS should be those students who are unable to access the paper-and-pencil test, even with accommodations. Additionally, students who are limited English proficient (that is, students who have been assessed on the state-identified English language proficiency tests as below Intermediate High in reading and been enrolled in U.S. schools for less than two years) may also participate in NCCLAS for reading, mathematics, and/or science. These students have received instruction on the grade-level academic content standards outlined in the NCSCS and are held to the same grade-level academic achievement standards.

The *NCEXTEND2* tests are based on grade-level content standards for the grade in which the student is enrolled and are challenging for eligible students, but the items may be less difficult than the general assessment and the grade-level academic achievement standards are modified accordingly. These tests are also multiple-choice, but only have three foils (response options) rather than four foils as on the general assessments. Eligible students for the *NCEXTEND2* tests are identified by the IEP team and meet the criteria outlined below.

- The student’s progress in response to high-quality instruction is such that the student is not likely to achieve grade-level proficiency within the school year covered by the IEP.
- The student’s disability has precluded the student from achieving grade level proficiency, as demonstrated by objective evidence, (e.g., results from standardized state tests, IQ tests, achievement tests, aptitude tests, and psychological evaluations).
- Beginning in 2007-08, the student’s IEP must include goals that are based on grade-level content standards and provide for monitoring of the student’s progress in achieving those goals.

At the high school level, some of these students will follow the Occupational Course of Study (OCS). The OCS tests are structured in the same way as the end-of-grade *NCEXTEND2* tests.

The determination of a significant cognitive disability is one criterion for student participation in the *NCEXTEND1*. The *NCEXTEND1* uses standardized tasks to assess student performance on the NCSCS Extended Content Standards. These extended content standards capture the essence of the grade-level content standards but allow for students with disabilities to access the curriculum at a different level. Determination of student proficiency is based on alternate academic achievement standards. The IEP team determines if the disability of a student is a significant cognitive disability; other criteria include the following:

- The student requires extensive and explicit instruction to acquire, maintain, and generalize new reading, mathematics, science, and writing skills for independent living.
- The student exhibits severe and pervasive delays in multiple areas of development and in adaptive behavior (e.g. mobility, communication, daily living skills, and self-care).
- The student is receiving instruction in the grade-level *Standard Course of Study (SCS)* Extended Content Standards for the subject(s) in which the students are being assessed. For 2007-08, this last element was clarified to read “in **ALL** assessed content areas.” The revised eligibility requirements clearly state that the *NCEXTEND1* is not appropriate for students who receive instruction in any of the general course content standards of the NCSCS.

Beginning in 2007-08, the eligibility requirements were amended to more explicitly define a significant cognitive disability as exhibiting “severe and pervasive delays in **ALL** areas of conceptual, linguistic and academic development and also in adaptive behavior areas, such as communication, daily living skills and self-care.”

### **3.7 Testing Accommodations**

On a case-by-case basis where appropriate documentation exists, students with disabilities and students with limited English proficiency may receive testing accommodations. The need for accommodations must be documented in a current Individualized Education Program (IEP), Section 504 Plan, or LEP plan. The accommodations must be used routinely during the student's instructional program and similar classroom assessments. For information regarding appropriate testing procedures, test administrators who provide accommodations for students with disabilities must refer to the most recent publication of *Testing Students with Disabilities* and any published supplements or updates. The publication is available through the local school system or at <http://www.ncpublicschools.org/accountability/policies/tswd/>. Test administrators must be trained in the use of the specified accommodations by the school system test coordinator or designee prior to the test administration.

### **3.8 Students with Limited English Proficiency**

Per HSP-C-021(d), last revised in April 2007, students identified as limited English proficient shall be included in the statewide testing program as follows: standard test administration, standard test administration with accommodations, or the state-designated alternate assessment. Students identified as limited English proficient who have been assessed on the state English language proficiency tests as below Intermediate/High in reading and who have been enrolled in United States schools for less than two years may participate in the state-designated alternate assessment in the areas of reading and mathematics at grades 3 through 8 and 10, science at grades 5 and 8, and in high school courses in which an end-of-course assessment is administered. To be identified as limited English proficient students must be assessed using the state English language proficiency tests at initial enrollment. All students identified as limited English proficient must be assessed using the state English language proficiency test annually thereafter during the spring testing window. A student who enrolls after January 1 does not have to be retested during the same school year.

Schools must administer state reading, mathematics, end-of-course assessments, and writing tests for students identified as limited English proficient who score at or above Intermediate/High on the state English language proficiency reading test during their first year in U.S. schools. Results from these assessments shall be included in the ABCs and AYP. Additionally, schools must include students previously identified as limited English proficient, who have exited limited English proficient identification during the last two years, in the calculations for determining the status of the limited English proficient subgroup for AYP only if that subgroup already met the minimum number of 40 students required for a subgroup.

### **3.9 Medical Exclusions**

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

For requests that involve significant medical emergencies and/or conditions, the LEA superintendent or charter school director is required to submit a justification statement that

explains why the medical emergency and/or condition prevents participation in the respective test administration during the testing window and the subsequent makeup period. The request must include the name of the student, the name of the school, the LEA code, and the name of the test(s) for which the exception is being requested. Medical documents are not included in the request to the NCDPI. The request is to be based on information housed at the central office. The student's records must remain confidential. Requests must be submitted prior to the end of the makeup period for the respective test(s).

### **3.10 Reporting Student Scores**

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

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

### **3.11 Confidentiality of Student Test Scores**

State Board of Education policy states that “any written material containing the identifiable scores of individual students on tests taken pursuant to these rules shall not be disseminated or otherwise made available to the public by any member of the State Board of Education, any employee of the State Board of Education, the State Superintendent of Public Instruction, any employee of the North Carolina Department of Public Instruction, any member of a local board of education, any employee of a local board of education, or any other person, except as permitted under the provisions of the Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g.”

## Chapter Four: Scaling, Equating, and Standard-Setting for the North Carolina Tests of Science

The North Carolina Tests of Science scores are reported as scale scores, achievement levels, and percentiles. Scale scores are advantageous in reporting because:

- scale scores can be used to compare test results when there have been changes in the curriculum or changes in the method of testing;
- scale scores on pretests or released test forms can be related to scale scores used on secure test forms administered at the end of the course;
- scale scores can be used to compare the results of tests that measure the same content area but are composed of items presented in different formats; and
- scale scores can be used to minimize differences among various forms of the tests.

### 4.1 Conversion of Raw Test Scores

Each student's score is determined by counting the number of items he or she answered correctly and then converting the number of correct responses to a developmental scale score. Items are assigned a score of 0 if the student did not answer the item correctly, and a score of 1 if the student did answer the item correctly. Software developed at the L.L. Thurstone Psychometric Laboratory at the University of North Carolina at Chapel Hill converts raw scores (total number of items answered correctly) to scale scores using the three IRT parameters (threshold, slope, and asymptote) for each item. The software implements the algorithm described by Thissen and Orlando (2001, pp. 119-130). Because different items are placed on each form of a subject's test, unique score conversion tables are produced for each form of a test for each grade or subject area. Each scale score has a conditional standard error of measurement associated with it.

Because the EOC Tests of Science are not developmental in nature, the scales were independently calibrated in the norming year to have a mean of 150 and a standard deviation of 10 for each test.

### 4.2 Setting the Standards

For tests developed under the North Carolina Testing Program, academic achievement standard setting, the process of determining cut scores for the different achievement levels, has historically been accomplished through the use of contrasting groups, and this method continues to be one source of information that is considered when setting standards. Contrasting groups is an examinee-based method of standard setting, which involves categorizing students into the four achievement levels by expert judges who are knowledgeable of students' achievement in various domains outside of the testing situation and then comparing these judgments to the distributions of students' actual scores. For the North Carolina Science tests, North Carolina teachers were considered expert judges under the rationale that teachers were able to make informed judgments about students' academic achievement because they had observed the breadth and depth of the students' work during the school year.

For the academic achievement standard setting for the new North Carolina EOC tests of Science, students were placed into categories by approximately 1,500 teachers for Biology;

1,500 teachers for Physical Science; and 1,000 teachers for Chemistry. Teachers categorized students into one of the four achievement levels as described by generic policy-level achievement level descriptors.

For the North Carolina EOC Science tests, the proportions of students expected to score in each of the four achievement levels were collected during the first operational administration in Fall of 2008. These proportions were applied to the distribution of student scores from the first administration to arrive at one possible set of cut points.

Internal meetings were held to review impact data. The meeting participants unanimously voted to present the achievement levels as predicted by the teachers to the State Board of Education.

In summer of 2008, the state conducted a modified Bookmark/item mapping standard setting facilitated by an external vendor. The results of this study confirmed the percentages of students classified by the contrasting group method. Therefore, the State Board of Education marked the interim achievement levels as final on October 2, 2008.

### **4.3 Score Reporting for the North Carolina Tests**

Scores from the North Carolina Science tests are reported as scale scores, achievement levels, and percentile ranks. The scale scores are computed through the use of raw-to-scale score conversion tables. The scale score determines the achievement level in which a student falls.

Score reports are generated at the local level to depict performance for individual students, classrooms, schools, and local education agencies. The data can be disaggregated by subgroups of gender and race/ethnicity, as well as other demographic variables collected during the test administration or through more authoritative source data collection throughout the school year, such as migrant census, school nutrition data, and so forth. Demographic data are reported on variables such as free/reduced lunch status, limited English proficient status, migrant status, Title I status, disability status, and parents' levels of education. The results are reported in aggregate at the state level usually in the middle of July of each year; disaggregated results are available later in the summer. The NCDPI uses the data for school accountability, student accountability, and to satisfy other federal requirements under the No Child Left Behind Act of 2001.

### **4.4 Achievement Level Descriptors**

The four policy-level generic achievement descriptors in the North Carolina Testing Program were defined previously. Using these policy-level generic descriptors and the results of the standard-setting, panels of teachers and curriculum experts created a content-based set of achievement level descriptors. After the final standards were approved by the State Board of Education, the achievement level descriptors were reviewed and refined by NCDPI curriculum staff, content experts, and teachers. The goal was to create achievement level descriptors that adequately described what content-specific skills a student should be able to demonstrate to differentiate performance across the four categories without tying student performance to a single test form or administration. The final content-specific achievement level descriptors

adopted by the State Board of Education are included as part of the test specifications in Appendix A.

#### 4.5 Achievement Level Cut Scores

The achievement level cut scores for the North Carolina Science tests are shown in the table below.

The scaling for the new tests and the scale score ranges corresponding to the four achievement levels were adopted on an interim basis for school year 2007-08 based on contrasting group methodology applied to Fall 2007 test data, along with other considerations. Upon review of full-year 2007-08 test data, and following a formal test-based item mapping process for confirmatory standard setting, the interim scale and achievement level scale score ranges were reaffirmed and adopted for use in subsequent academic years as a final articulation of achievement standards for the 3<sup>rd</sup> edition tests of EOC Science. Final achievement level descriptors were also adopted at this time.

**Table 7:** EOC Tests of Science, achievement levels and corresponding scale scores

<b>Subject</b>	<b>Level I</b>	<b>Level II</b>	<b>Level III</b>	<b>Level IV</b>
Physical Science	≤ 139	140-148	149-159	≥160
Biology	≤137	138-146	147-158	≥159
Chemistry	≤136	137-145	146-157	≥158
Physics	≤138	139-142	143-156	≥157

#### 4.6 Achievement Level Trends

The percentage of students in each of the achievement levels is provided below by grade for selected school years. The years shown include two selected academic years of administration (2002-03 and 2005-06) for the 2<sup>nd</sup> edition tests and the first academic year (2007-08) of administration for the 3<sup>rd</sup> edition tests. It is important to note that the 2007-08 tests are tied to a new curriculum (2004 SCS) and new achievement levels.

**Table 8:** Achievement level trends for Physical Science

<b>Physical Science</b>	<b>2002-2003</b>	<b>2005-2006</b>	<b>2007-2008</b>
Level I	5.3	2.8	13.6
Level II	30.7	28.0	28.1
Level III	46.4	56.0	40.8
Level IV	17.6	13.2	17.5

**Table 9:** Achievement level trends for Biology

<b>Biology</b>	<b>2002-2003</b>	<b>2005-2006</b>	<b>2007-2008</b>
Level I	10.4	8.4	8.9
Level II	28.6	28.2	23.0
Level III	45.6	45.3	45.3
Level IV	15.3	18.1	22.9

**Table 10:** Achievement level trends for Chemistry

<b>Chemistry</b>	<b>2002-2003</b>	<b>2005-2006</b>	<b>2007-2008</b>
Level I	5.8	4.7	7.2
Level II	20.0	18.2	21.0
Level III	39.9	39.6	44.2
Level IV	34.3	37.5	27.7

**Table 11:** Achievement level trends for Physics

<b>Physics</b>	<b>2001-2002</b>	<b>2005-2006</b>	<b>2007-2008</b>
Level I	2.9	2.8	4.1
Level II	13.7	12.1	12.8
Level III	42.1	40.2	49.1
Level IV	41.4	44.9	34.0

#### **4.7 Percentile Ranking**

The percentile rank for each scale score is the percentage of scores less than or equal to that score. A percentile is a score or a point on the original measurement scale. The percentile rank provides relative information about a student's score on a test relative to other students in the norming year. The percentile ranks for the scores on the North Carolina Science tests are calculated based on the first operational administration of the tests. The use of percentile rank reporting allows a meaningful comparison to be made among Science scores at the total test score level.

## Chapter Five: Reports

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

The North Carolina Testing Program provides reports at the student level, school level, and state level. The North Carolina Testing Code of Ethics (see Appendix F) dictates that educators use test scores and reports appropriately. This means that educators recognize that a test score is only one piece of information and must be interpreted together with other scores and indicators. Test data help educators understand educational patterns and practices. Data analysis of test scores for decision-making purposes should be based upon disaggregation of data by student demographics and other student variables as well as an examination of grading practices in relation to test scores, growth trends, and goal summaries for state-mandated tests.

### 5.2 Reporting by Student

The state provides scoring equipment and software in each school system so that administrators can score all state-required multiple-choice tests. This scoring generally takes place within two weeks after testing so the individual score report can be given to the student and parent before the end of the school year. School districts who test earlier in the window submit their data to the NCDPI for quality control purposes; those districts are strongly encouraged to not print any reports until the quality control procedures have been completed and the data are certified.

Each student who takes the EOC Science tests is given an “Individual Student Report.” This single sheet provides information on that student’s performance on each EOC Science test. A flier titled, “Understanding Your Child’s EOC Score,” is provided with each Individual Student Report. This publication offers information for understanding student scores as well as suggestions on what parents and teachers can do to help students in the areas of science.

The student report also shows how that student’s performance compared to the average scores for the school, the school system, and the state. A four-level achievement scale is used for the tests. A set of global policy-level generic descriptors are used for all subjects:

Achievement Level I represents insufficient mastery of the subject.

Achievement Level II is inconsistent mastery of the subject.

Achievement Level III is consistent mastery and the minimum goal for students.

Achievement Level IV is superior mastery of the subject.

Additionally, content-specific achievement level descriptors are developed as an outgrowth of the standard-setting process. Additionally, the appropriate achievement level descriptor is included on the Individual Student Report.

Beginning in the 2007-08 school year, the Individual Student Report is being redesigned to provide more feedback to parents and teachers about a student’s strengths and weaknesses in the content area. Although the new report only includes information at the goal (superordinate) level, it is actually less prone to misinterpretation or misuse than the older reports which included information at the objective (subordinate) level. Additionally, the goal summary now

is scaled in order to provide somewhat more meaningful interpretations of an individual student's strengths or weaknesses.

Of course it is clearly understood by the NCDPI that reporting at a grain finer than the total test score level is less reliable. Reliability can be bolstered by statistical means such as regression or IRT methods. However, these methods tend to mask individual student profiles by either forcing the student goal-level profile to look more like the state profile, or by flattening the student profile and minimizing any real differences in performance by goal.

In order to attempt to provide meaningful sub-score reporting, a student score for each goal is calculated and scaled to have a mean of 10 and a standard deviation of 3; thus a student goal-level scale score of 10 means the student did about as well on that topic as the rest of the students in the state. A student goal-level scale score below 10 indicates the student did not do as well as other students on that topic, while a student goal-level score above 10 indicates the student did better on that topic than other students in the state. Conditional standard errors are indicated by shaded bands around the point estimate. Strong cautions are included with the report that these scores are less reliable than the total score, and that instructional or placement decisions should not be made on the basis of the sub-scores alone.

### **5.3 Reporting by Classroom**

Classroom rosters can be created from the state-supplied software. These rosters include, for each student, a summary of the information contained on the Individual Student Report. For Algebra I, the classroom roster also provides the numeric "grade" to be factored in to the student's course grade. The default conversion is provided by the state, but the actual conversion used is a local decision. Any district can make its conversion more strict than the state default, to be more in line with district grading policies.

### **5.4 Reporting by School**

Since 1997, the student performance on end-of-grade tests for each elementary and middle school has been released by the state through the ABCs of School Accountability. High school student performance began to be reported in 1998 in the ABCs of School Accountability. For each school, parents and others can see the actual performance for groups of students at the school in reading, mathematics, and writing; the percentage of students tested; whether the school met or exceeded goals that were set for it; and the status designated by the state.

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

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

### **5.5 Reporting by District**

Each district receives its own LEA summary of student performance on the tests that are in the ABCs accountability model as well as information on how the LEA performed in terms of AYP.

### **5.6 Reporting by the State**

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

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

## Chapter Six: Descriptive Statistics and Reliability

### 6.1 Descriptive Statistics for the First Operational Administration of the Tests

The EOC Tests of Science were administered for the first time in the 2005-06 school year (fall and spring block schedule, and traditional calendar). Descriptive statistics for the North Carolina Tests of Science' first operational year and operational administration population demographics are provided below.

### 6.2 Means and Standard Deviations for the First Year of Operational Administration

**Table 12:** Descriptive statistics by course for the 2007-08 administration of the North Carolina EOC Tests of Science

<b>Subject</b>	<b>N</b>	<b>Scale Score Mean</b>	<b>Scale Score Standard Deviation</b>	<b>Average p-Value of Tests</b>
Physical Science (2007-08)	54,425	150.44	9.36	.5837
Biology (2007-08)	97,394	151.00	9.51	.6005
Chemistry (2007-08)	45,139	151.28	9.72	.6136
Physics (2007-08)	9,686	151.32	9.35	.6305

### 6.3 Population Demographics for the First Operational Administration

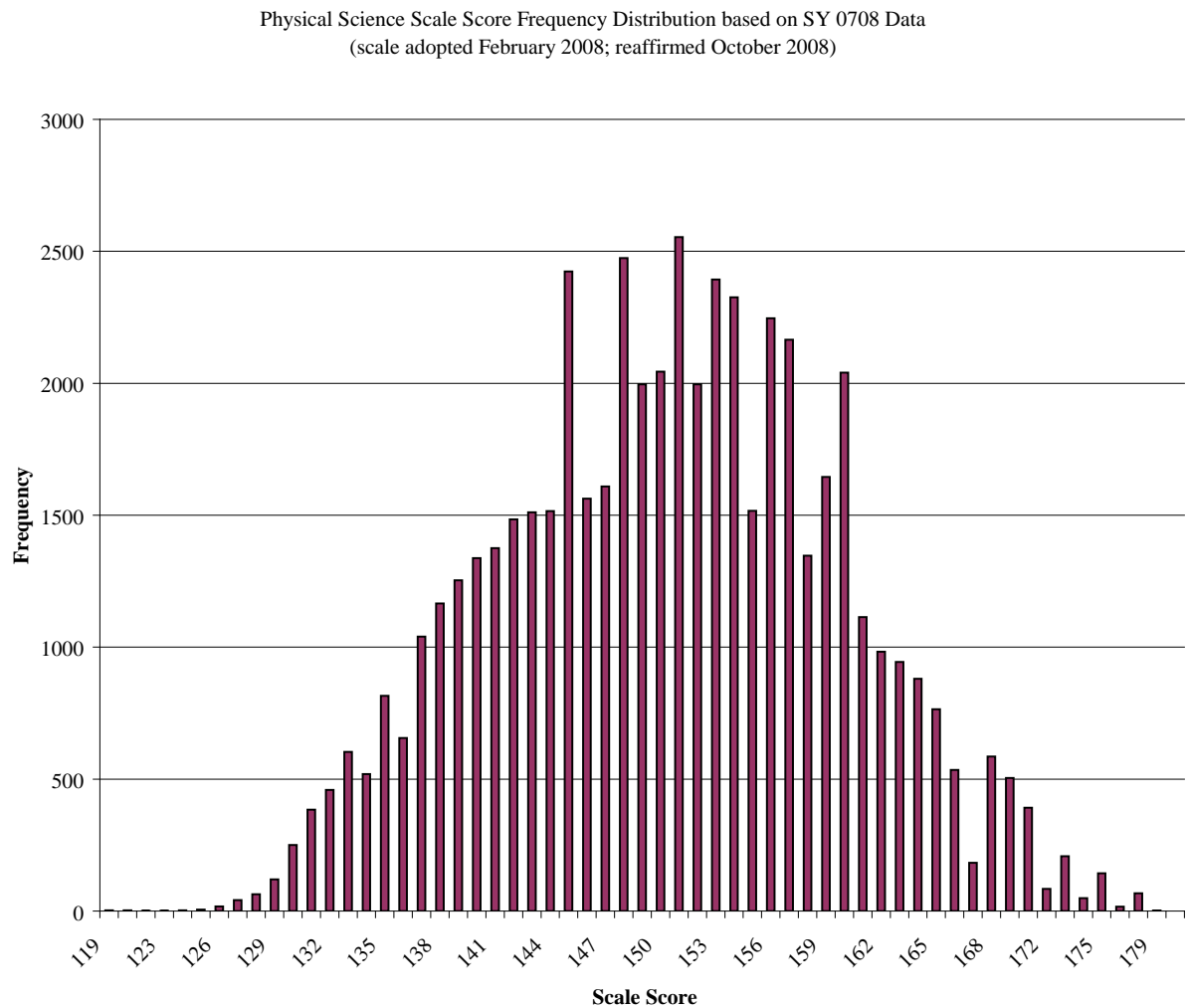
**Table 13:** Population demographics for the first administration of the North Carolina EOC Tests of Science

<b>Subject</b>	<b>% Male</b>	<b>% Female</b>	<b>% Asian</b>	<b>% Black</b>	<b>% Hispanic</b>	<b>% American Indian</b>	<b>% Multiracial</b>	<b>% White</b>	<b>% LEP (Limited English Proficiency)</b>	<b>% EDS (Economically Disadvantaged Student)</b>
Physical Science	51.4	48.6	1.5	32.5	7.0	1.4	2.1	55.4	3.4	40.3
Biology	49.4	50.6	2.4	29.1	6.6	1.2	2.3	58.4	3.3	34.6
Chemistry	44.2	55.8	3.7	20.0	4.6	0.8	2.3	68.7	1.5	20.1
Physics	59.9	40.1	6.5	13.6	3.6	0.5	2.0	73.9	1.4	13.1

## 6.4 Scale Score Frequency Distributions

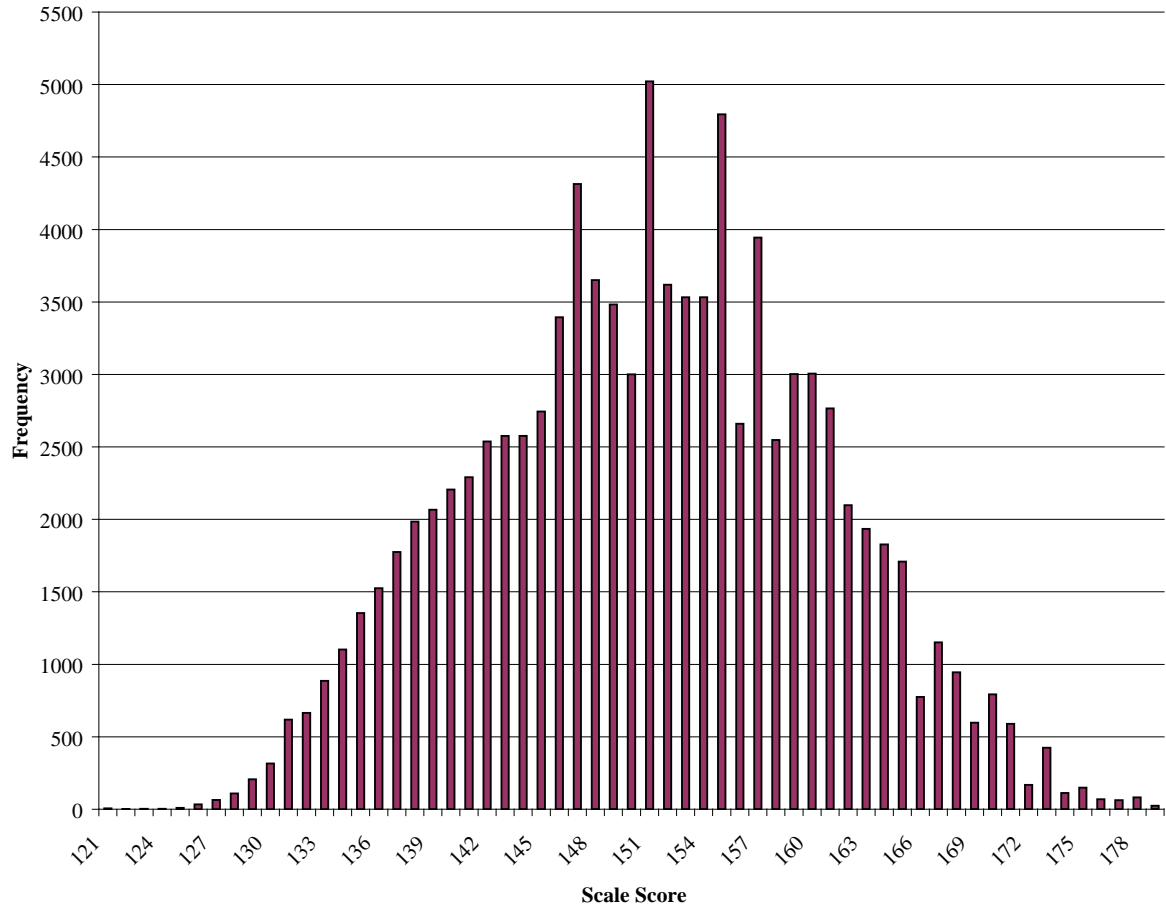
The following figures present the frequency distributions of the developmental scale scores from the first statewide administration of the North Carolina EOC Tests of Science. The frequency distributions are not smooth because of the conversion from raw scores to scale scores. Due to rounding in the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

**Figure 4:** Physical Science Scale Score Frequency Distribution



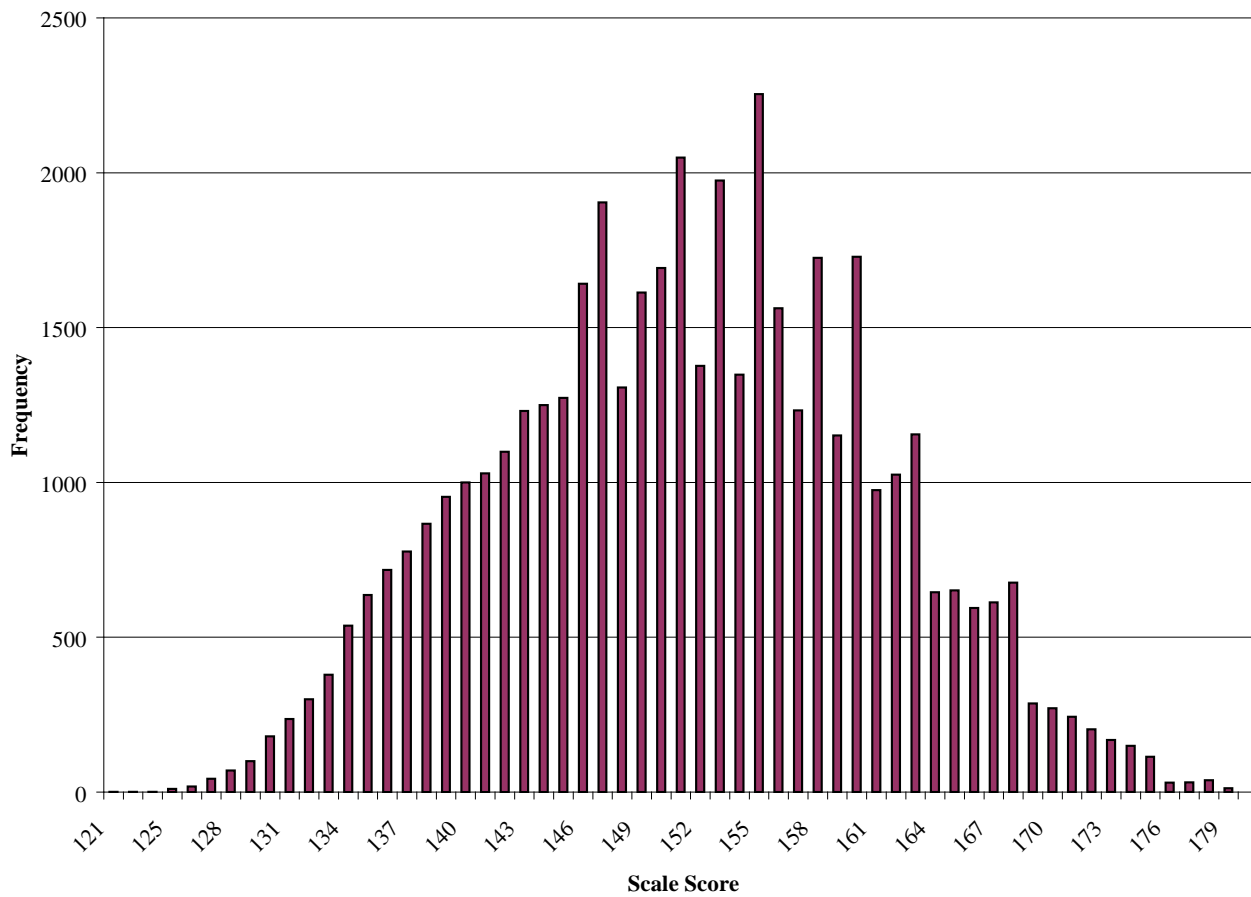
**Figure 5:** Biology Scale Score Frequency Distribution

Biology Scale Score Frequency Distribution based on SY 0708 Data  
(scale adopted February 2008; reaffirmed October 2008)



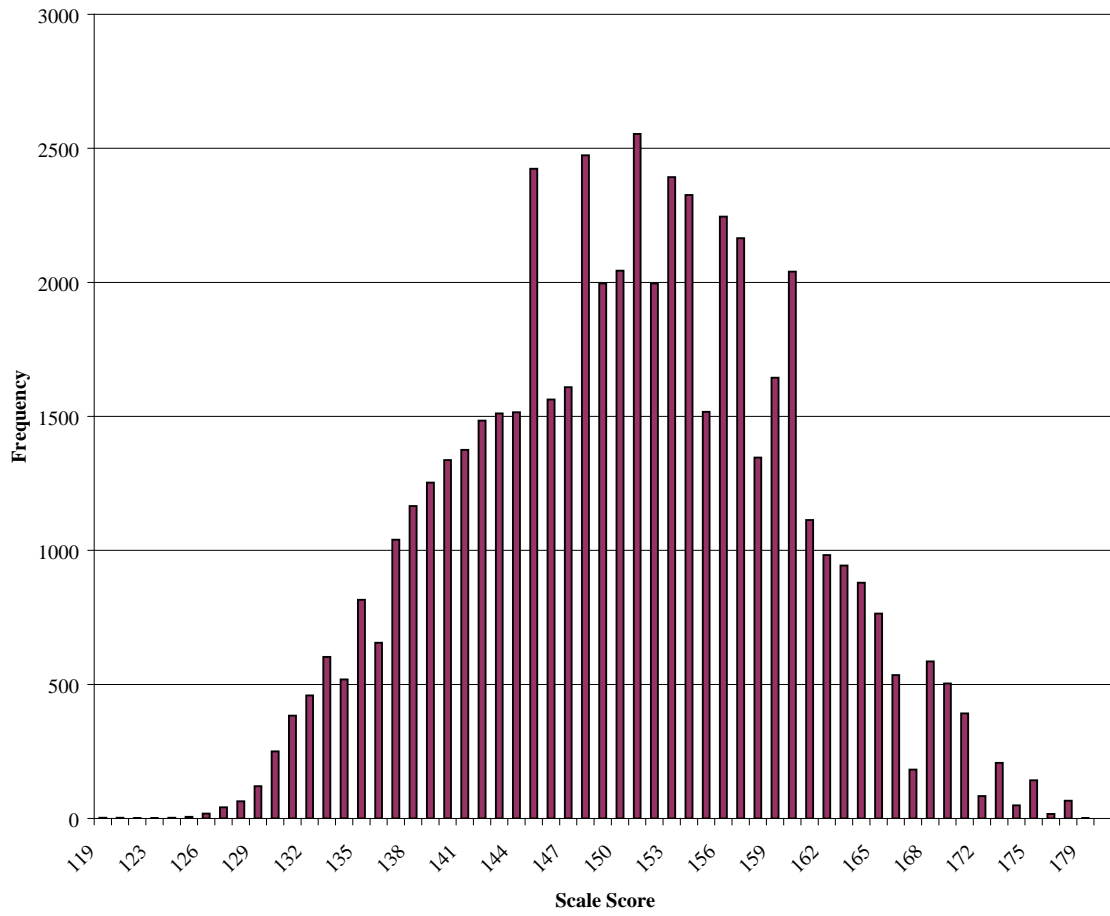
**Figure 6:** Chemistry Scale Score Frequency Distribution

Chemistry Scale Score Frequency Distribution based on SY 0708 Data  
(scale adopted February 2008; reaffirmed October 2008)



**Figure 7:** Physics Scale Score Frequency Distribution

Physics Scale Score Frequency Distribution based on SY 0708 Data  
(scale adopted February 2008; reaffirmed October 2008)



## 6.5 Reliability of the North Carolina Science Tests

Reliability refers to the consistency of a measure when the testing procedure is repeated on a population of individuals or groups. In testing, if use is to be made of some piece of information, then the information should be stable, consistent, and dependable. If any use is to be made of the information from a test, then the test results must be reliable. If decisions about individuals are to be made on the basis of test data, then it is desirable that the test results are reliable and replicable. For a high-stakes multiple-choice test, the reliability coefficient should be at least 0.85.

There are three broad categories of reliability coefficients recognized as appropriate indices for establishing reliability in tests: (a) coefficients derived from the administration of parallel forms in independent testing sessions (alternate-form coefficients); (b) coefficients obtained by administration of the same instrument on separate occasions (test-retest or stability coefficients); and (c) coefficients based on the relationships among scores derived from individual items or subsets of the items within a test, all data accruing from a single administration of the test. The last coefficient is known as an internal consistency coefficient (*Standards for Educational and Psychological Testing*, AERA, APA, NCME, 1985, p.27). An internal consistency coefficient, coefficient alpha, is the metric generally used to establish reliability for the North Carolina EOC Tests of Science.

## 6.6 Internal Consistency of the North Carolina Science Tests

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

**Table 14:** Reliability indices averaged across North Carolina EOC Tests of Science forms

Subject	N Operational Items on Test	Average Coefficient Alpha	Range of Coefficients Alpha
Physical Science	60	0.908	.905 - .910
Biology	60	0.910	.905 - .922
Chemistry	60	0.914	.908 - .920
Physics	63	0.912	.910 - .914

As noted above, the North Carolina EOC Tests of Science are highly reliable as a whole. In addition, it is important to note that this high degree of reliability extends across gender, ethnicity, and other characteristics, including disability status, limited English proficiency (LEP) status, Title 1 status, migrant status, and economically disadvantaged student (EDS) status. Looking at coefficients alpha for the different groups reveals that across all test forms, in all courses, 69% of the values were at or above 0.900, and all but one (1.4% of all reliability coefficients) were above 0.850.

**Table 15:** Reliability indices averaged across North Carolina EOC Test of Science forms (Gender)

Subject	Females	Males
Physical Science	0.903	0.913
Biology	0.904	0.916

Chemistry	0.910	0.919
Physics	0.905	0.913

**Table 16:** Reliability indices averaged across North Carolina EOC Tests of Science forms (Ethnicity)

Subject	Asian	Black	Hispanic	Native American	Multi-Racial	White
Physical Science	0.917	0.882	0.906	0.896	0.904	0.907
Biology	0.927	0.874	0.899	0.881	0.908	0.904
Chemistry	0.931	0.888	0.909	0.872	0.909	0.911
Physics	0.922	0.877	0.906	N/A*	0.915	0.902

\*The number of examinees in the Native American subgroup for Physics was too small to permit the calculation of a meaningful reliability index for that subgroup.

**Table 17:** Reliability indices averaged across North Carolina EOC Tests of Science forms (Other Characteristics)

Subject	No Disability	Disability	Not LEP	LEP	Not Title I	Title I	Not Migrant	Migrant	Not EDS	EDS
Physical Science	0.907	0.889	0.907	0.891	0.908	0.849	0.908	0.896	0.910	0.896
Biology	0.908	0.892	0.909	0.877	0.910	0.884	0.910	0.897	0.908	0.884
Chemistry	0.914	0.909	0.914	0.910	0.914	0.880	0.914	0.877	0.914	0.898
Physics	0.912	0.915	0.911	0.939	0.911	N/A*	0.912	N/A*	0.909	0.896

\* The numbers of examinees in the Title I and Migrant subgroups for Physics were too small to permit the calculation of meaningful reliability indices for those subgroups.

There was some variation among forms. Coefficients alpha that were below the 0.85 threshold were as follows:

- for the Title 1 subgroup: Physical Science Form A (.829) and Physical Science Form C (.811);
- for the Migrant subgroup: Chemistry Form B (.849)

Although the North Carolina Testing Program administers alternate forms of the test, it is not generally possible to calculate alternate-forms reliabilities on the tests within the context of a natural test setting. Students take the test one time, and only those students who do not achieve Level III are required to retake the test. Thus, the natural population of re-testers has a sharp restriction in range, which would lower the observed correlation. Additionally, North Carolina students are extremely test-wise. A study on test-retest reliability, where one of the administrations does not have stakes for the student, with this population would give questionable results.

## 6.7 Standard Error of Measurement

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

The standard error of measurement ranges for scores on the North Carolina EOC Tests of Science is provided in table 20 below. For students with scores within 2 standard deviations of the mean (95% of the students), standard errors are typically 2 to 3 points. For most of the EOC Tests of Science scale scores, the standard error of measurement in the middle range of scores, particularly at the cut point between Level II and Level III, is generally around 3 points. Scores at the lower and higher ends of the scale (above the 97.5 percentile and below the 2.5 percentile) have standard errors of measurement of approximately 5 to 6 points. This is typical as scores become more extreme due to less measurement precision associated with those extreme scores.

**Table 18:** Ranges of standard error of measurement for scale scores by subject

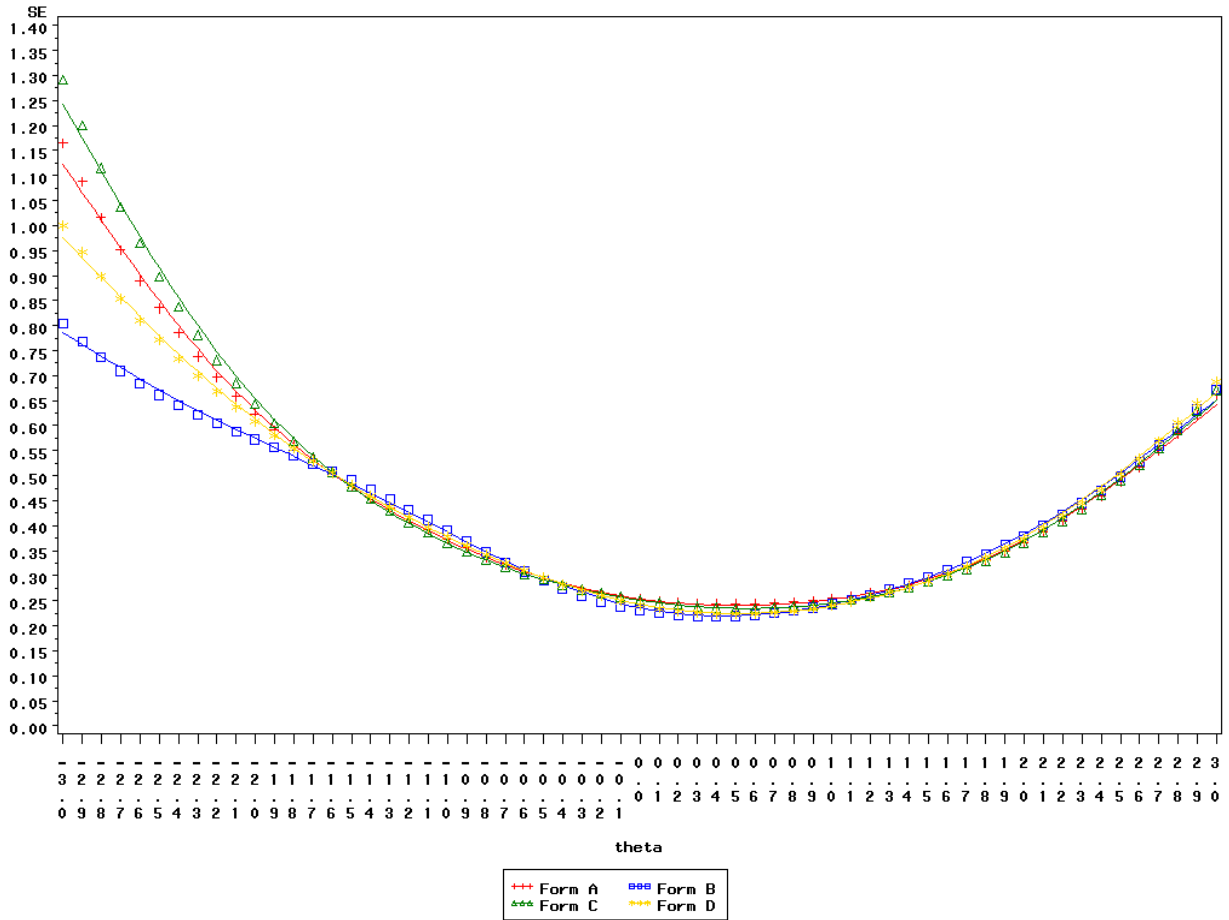
<b>Subject</b>	<b>SEM Range</b>	<b>SEM at I-II Cut Score</b>	<b>SEM at II-III Cut Score</b>	<b>SEM at III-IV Cut Score</b>
Physical Science	2–5	4	3	2–3 *
Biology	2–5	4–5 *	3	2–3 *
Chemistry	2–5	4–5 *	3	2–3 *
Physics	2–5	4	3-4*	2

\* SEM at Cut Score varies according to test form

Additionally, standard error curves are presented in the following figures. The horizontal axis represents the  $\theta$  estimate (the estimate of the test-taker’s true ability), on a scale from -3 to +3.

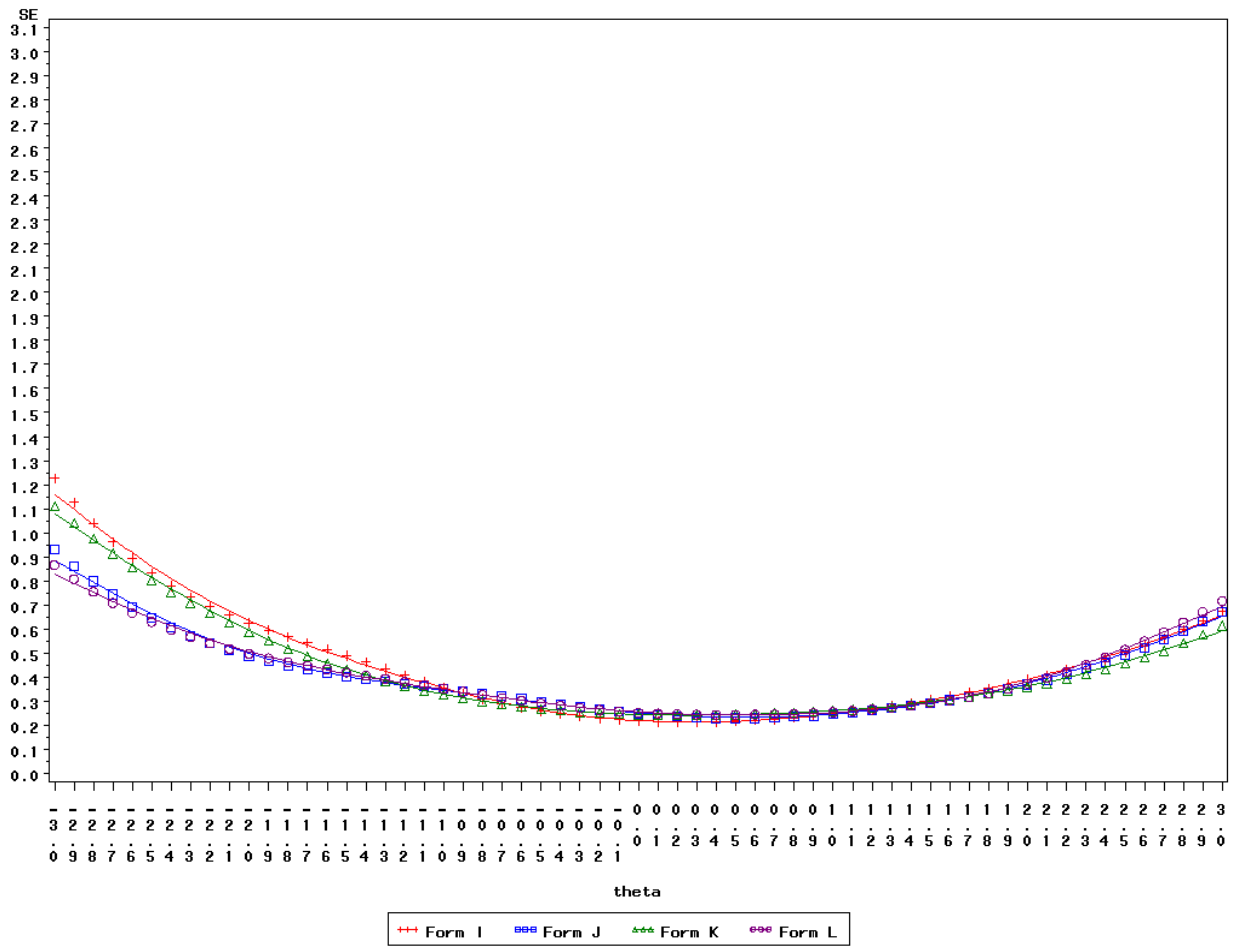
**Figure 8:** Standard Errors of Measurement on the Physical Science Test forms

Physical Science Forms ABCD: Standard Error Curves (2007–08 Op Parameters)



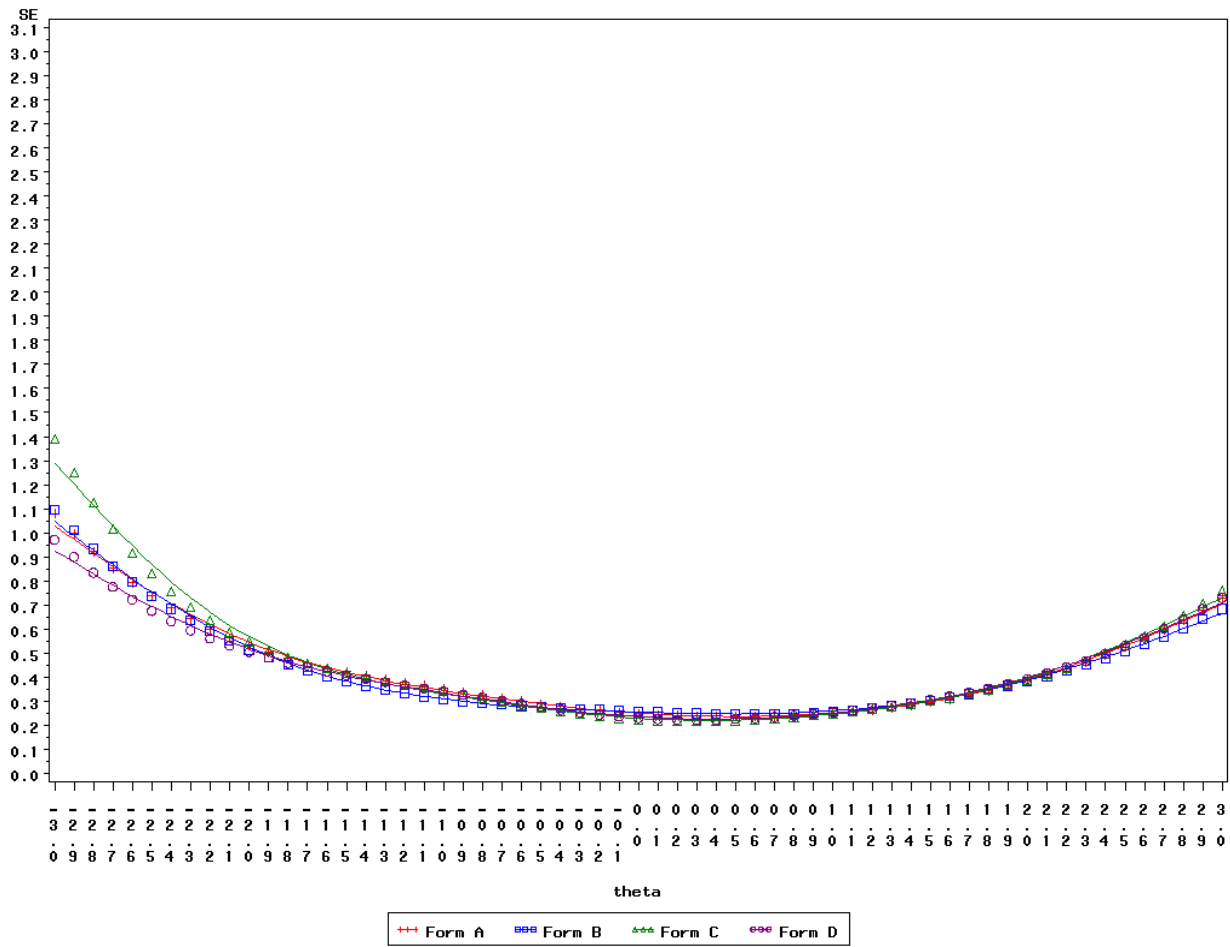
**Figure 9:** Standard Errors of Measurement on the Biology Test forms

EOC Biology 2007–2008 Forms IJKL: Standard Error Curves



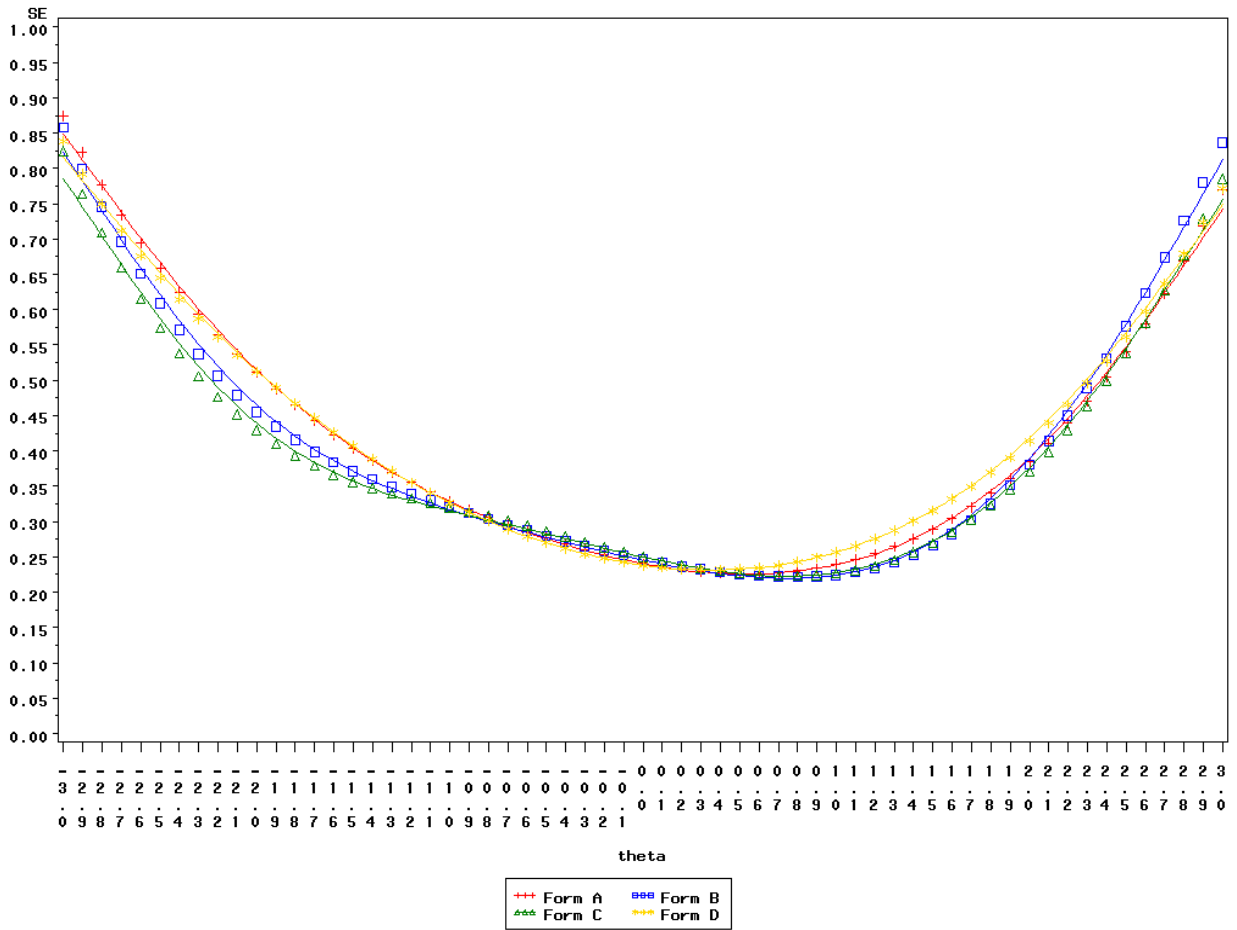
**Figure 10:** Standard Errors of Measurement on the Chemistry Test forms

EOC Chemistry 2007–2008 Forms ABCD: Standard Error Curves



**Figure 11:** Standard Errors of Measurement on the Physics Test forms

Physics Forms ABCD: Standard Error Curves (2007–08 Op Parameters)



## 6.8 Equivalency of Test Forms

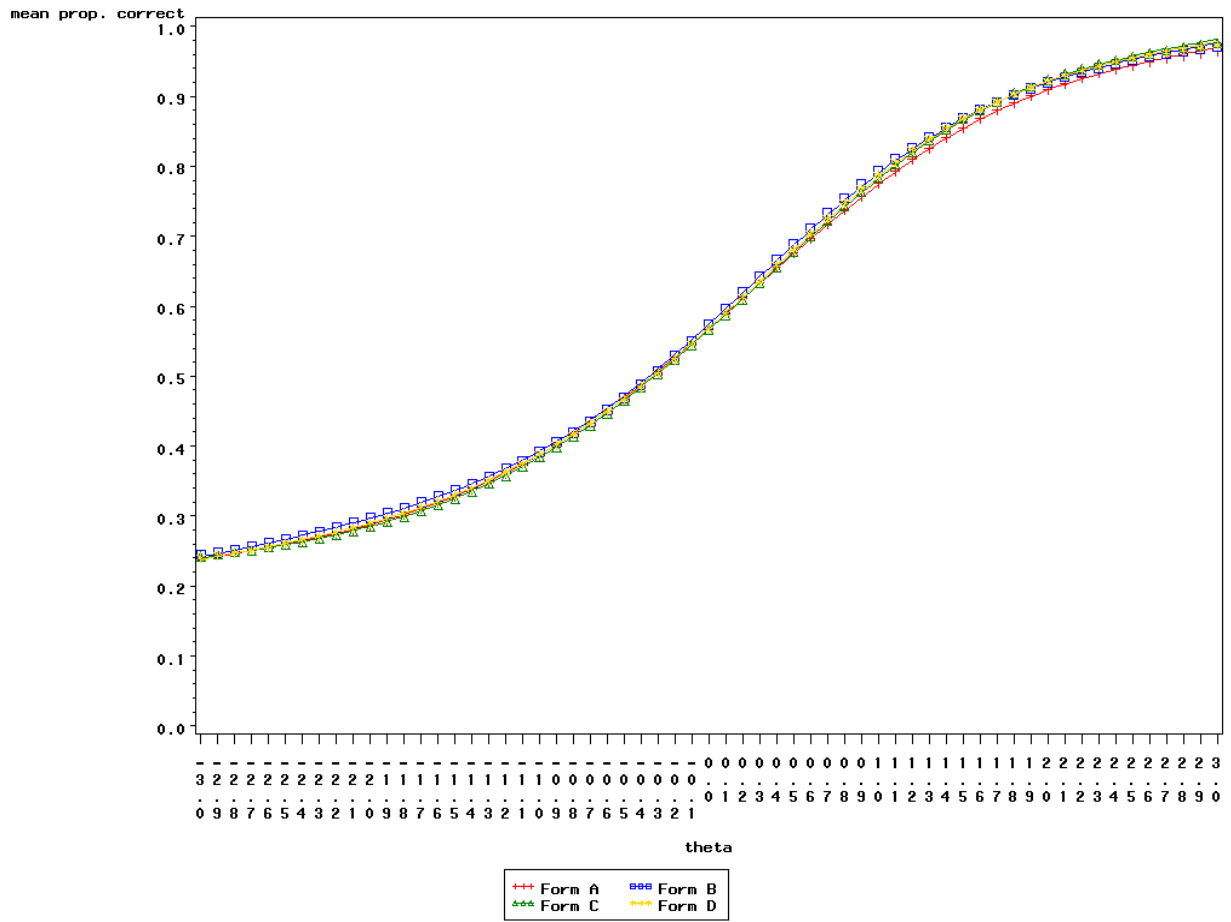
North Carolina administers multiple forms of each test during each testing cycle. This serves several purposes. First, it allows North Carolina to fully test the breadth and depth of each curriculum. The curricula are extremely rich, and administering a single form that fully addressed each competency would be prohibitively long. Additionally, the use of multiple forms reduces the incidence of one student copying from the test of another student.

The tests are parallel in terms of content coverage at the goal level. That is, each form has the same number of items from the number and operations strand (Goal 1) as every other form administered in that grade. The specific questions asked on each form are a random domain sample of the topics in that grade's goals, although care is taken to not overemphasize a particular topic on a single test form.

The following figures demonstrate the statistical equivalency of the multiple test forms. For each grade's set of test forms, the test characteristic curves are very nearly coincident for much of the range of  $\theta$ . Slight variations appear in the test curves at the extremes, as the tests were designed to have maximum sensitivity in the middle of the range of examinee ability.

**Figure 12:** Test Characteristic Curves for the Physical Science Test forms

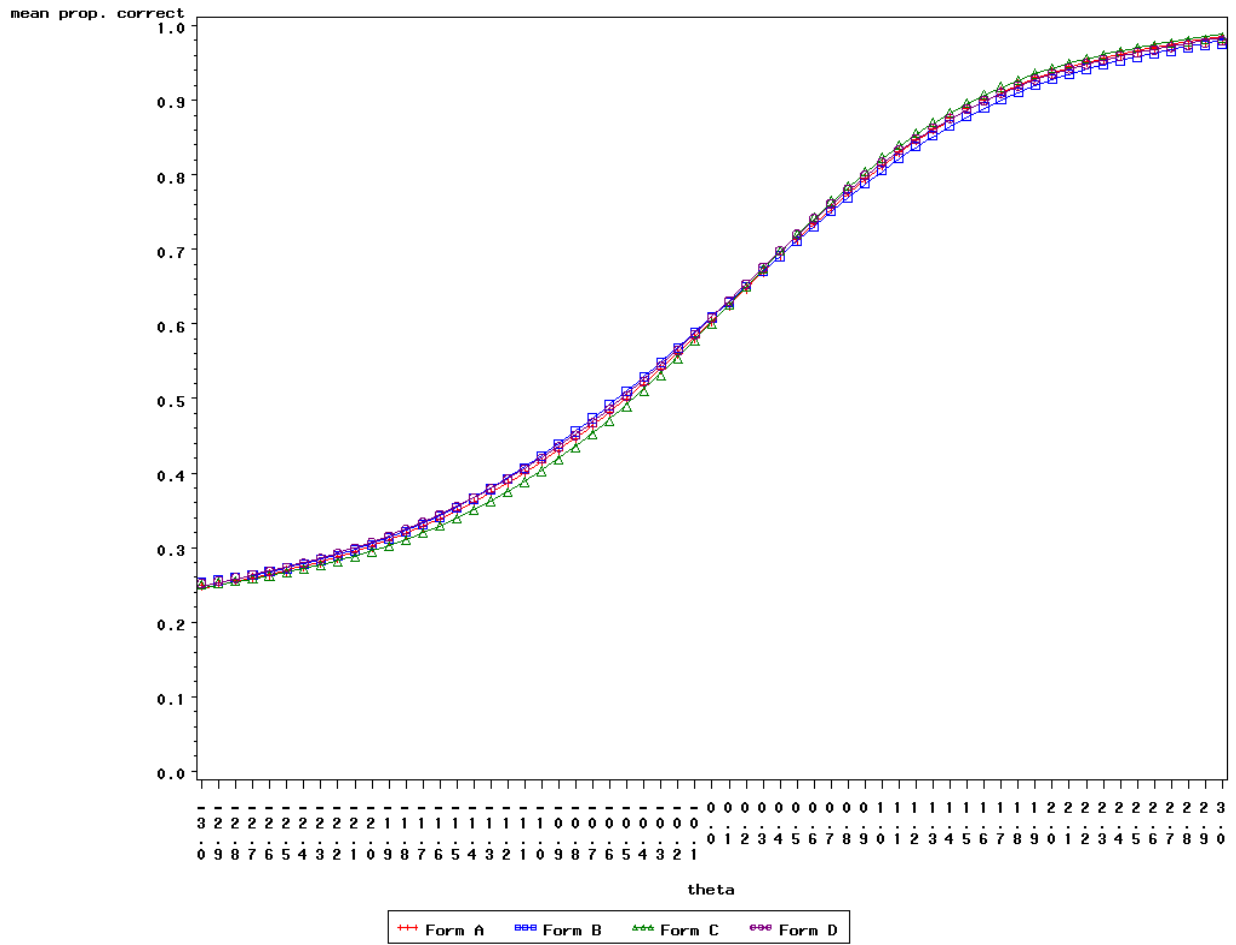
Physical Science Forms ABCD: Test Characteristic Curves (2007–08 Op Parameters)





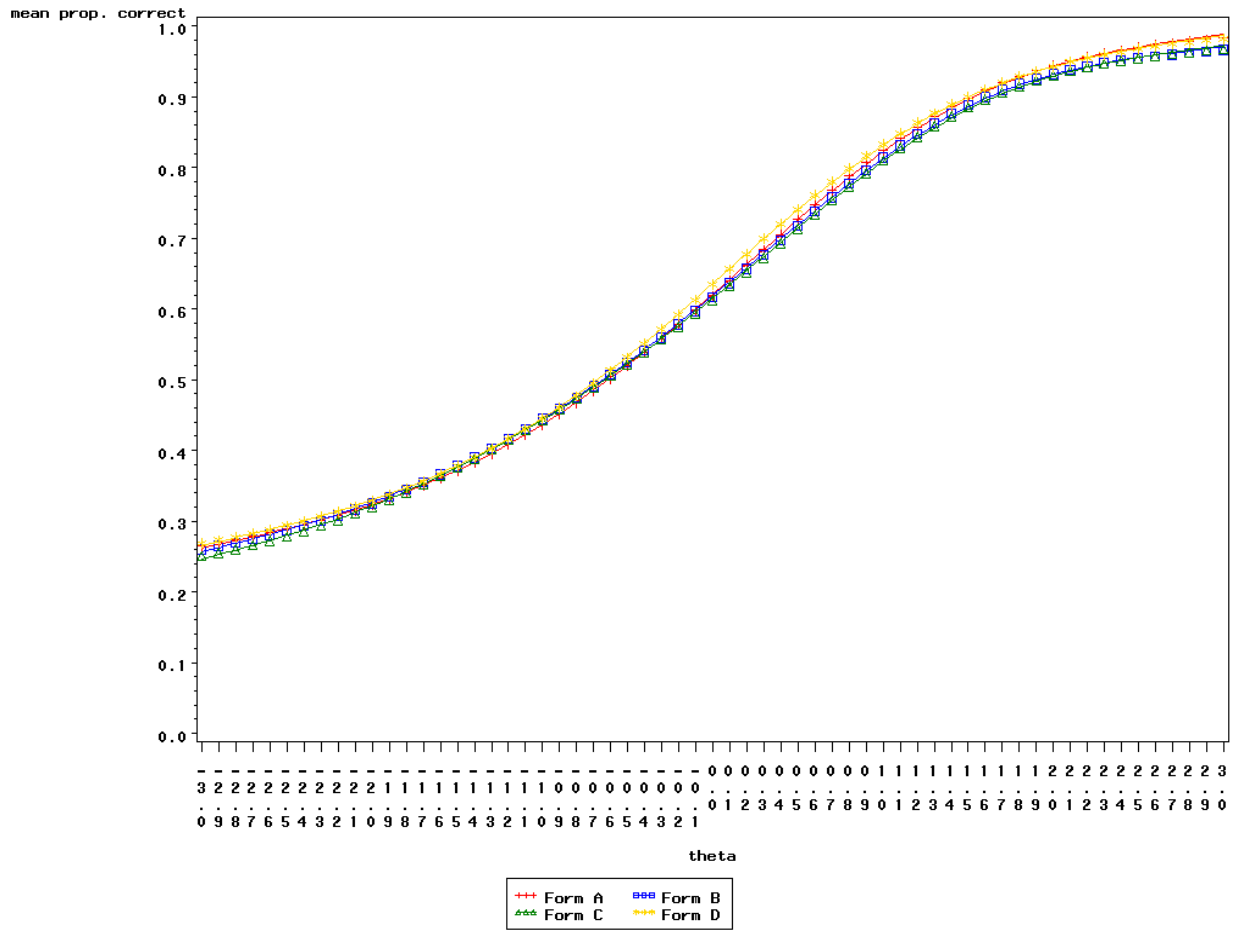
**Figure 14:** Test Characteristic Curves for the Chemistry Test forms

### EOC Chemistry 2007–2008 Forms ABCD: Test Characteristic Curves



**Figure 15:** Test Characteristic Curves for the Physics Test forms

Physics Forms ABCD: Test Characteristic Curves (2007–08 Op Parameters)



# Chapter Seven: Evidence of Validity

## 7.1 Evidence of Validity

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

## 7.2 Content Validity

Evidence of content validity begins with an explicit statement of the constructs or concepts being measured by the proposed test. Physical Science, Biology, Chemistry, and Physics comprise the EOC Tests of Science. These tests measure the different levels of science knowledge, skills, and abilities specific to the three areas with particular focus on assessing students' ability to process information and engage in higher order thinking. These elements of science measured by the North Carolina EOC Tests are also categorized into strands: *Nature of Science*, *Science as Inquiry*, *Science and Technology*, *Science in Personal and Social Perspectives*.

For test specification summaries, see Appendix A.

Almost all of the items are written by North Carolina teachers and other educators. Some of the science items were written under a contract with a major testing company to handle the logistics, but that contract specified that at least half of the items be written by teachers from North Carolina. Additionally, the items were all reviewed by North Carolina teachers. Where possible, the old items were recategorized into the appropriate grade or subject, goal, and objective. Additional items, representing the vast majority of the items written for the third edition, were written by North Carolina educators.

All item writers undergo training, during which time they are taught certain rules and guidelines for item (stem and foil) construction. The training also includes information on Universal Design and access issues for special populations such as students with disabilities and English language learners. The Universal Design training also includes a discussion of considerations for how various accommodations could impact the validity of an item and how to construct items to avoid invalidating an item. Finally, the item writer training includes information on what constitutes bias and how to minimize differential item functioning through careful item construction.

Additionally, all items written are reviewed by at least two content-area teachers from North Carolina. Because North Carolina educators not only deliver the *Standard Course of Study* every day in their classrooms, they are also the most familiar with the way in which students

---

---

learn and understand the material. Thus, North Carolina teachers are best able to recognize questions that not only match the *Standard Course of Study* for their particular course or grade, but also are relevant and comprehensible to the students at that level.

During the review process, the items are also reviewed by a specialist in Exceptional Children and a specialist in English as a Second Language. The specialist teachers review the items in a team with the content teachers in order to make the items as accessible as possible to all populations while preserving the integrity of the curriculum.

The state's teachers are also involved in other aspects of item development and test review (refer to Figure 1, the test development process).

### **7.3 Criterion-Related Validity**

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

Criterion-related validity of a test indicates the effectiveness of a test in predicting an individual's behavior in a specific situation. The criterion for evaluating the performance of a test can be measured at the same time (concurrent validity) or at some later time (predictive validity).

For the North Carolina EOC Tests of Science, teachers' judgments of student achievement, expected grade, and assigned achievement levels all serve as sources of evidence of concurrent validity. The Pearson correlation coefficient is used to provide a measure of association between the scale score and those variables listed above. The correlation coefficients for the North Carolina EOC Tests of Science range from 0.41 to 0.79 indicating a moderate to strong correlation between EOC scale scores and variables based on teacher judgment of the students' attainment of the content.\* The tables below provide the Pearson correlation coefficients for variables used to establish criterion-related validity for the North Carolina EOC Tests of Science.

*\*Note: By comparison, the uncorrected correlation coefficient between SAT score and freshman year grades in college is variously reported as 0.35 to 0.55 (Camera & Echternacht, 2000).*

**Table 19:** Pearson correlation coefficient table for variables used to establish criterion-related validity for the North Carolina EOC Tests of Science, 3<sup>rd</sup> Edition

<b>Subject</b>	<b>Physical Science</b>	<b>Biology</b>	<b>Chemistry</b>	<b>Physics</b>
Teacher Judgment of Achievement Level by Assigned Achievement Level	0.55	0.59	0.58	0.43
Teacher Judgment of Achievement by Expected Grade	0.75	0.77	0.79	0.73
Teacher Judgment of Achievement by Scale Score	0.58	0.63	0.62	0.47
Assigned Achievement Level by Expected Grade	0.49	0.54	0.53	0.41
Expected Grade by Scale Score	0.53	0.58	0.57	0.45

The variables used in the tables above are as follows.

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

The NCDPI found moderate to strong correlations between scale scores in science and variables such as teachers’ judgment of student achievement, expected grade, and assigned achievement levels (all measures of concurrent validity). Equally important is to demonstrate that the test scores are *not* correlated with external variables that should not be related to student proficiency or should only be moderately correlated. The department also found generally low correlations among the test scores and variables external to the test such as gender, limited English proficiency, and disability for all grades and subjects.

Gender was coded as male (0) or female (1); a negative coefficient for gender means that males did better than females.

The variables Disability, Migrant, EDS, and Title I status were all coded so that not-applicable was coded as 0. Thus, negative coefficients mean that students who were not, for example, of migrant status, did better than students who were of migrant status.

**Table 20:** Tetrachoric correlation coefficient table for additional, presumably uncorrelated variables used to establish criterion-related validity for the North Carolina EOC Tests of Science

<b>Score by Subject</b>	<b>× Gender</b>	<b>× Disability</b>	<b>× Migrant</b>	<b>× EDS</b>	<b>× Title I</b>
Physical Science	-0.03	-0.30	0.03	-0.25	-0.31
Biology	-0.04	-0.33	-0.13	-0.40	-0.28
Chemistry	-0.07	-0.13	-0.13	-0.26	-0.33
Physics	-0.23	-0.05	-0.06	-0.35	-0.37

Nearly half (45%) of the correlations between scores and external variables were less extreme than  $\pm 0.2$ , and only one of the correlations between scores and external variables was more extreme than  $\pm 0.4$ . None of these relationships exceeded the degree of relationship recorded for the selected measures of concurrent validity. These generalizations held across the full range of forms administered by DPI for all the courses.

#### **7.4 Concurrent and Predictive Validity**

Concurrent validity and predictive validity can be demonstrated if a test’s scores or interpretations show a strong relationship to the scores or interpretations from another, already validated, instrument that measures the same construct or a closely related construct. Conclusions about concurrent validity information, as the name suggests, can be drawn when the two measures occur at or nearly at the same time. Predictive validity, on the other hand, would imply the earlier measurement provides information on the performance of the test-taker on the second measure at a later point in time.

Because the North Carolina tests are the only tests that measure the North Carolina *Standard Courses of Study*, it is difficult to obtain obvious concurrent validity data. Instead, concurrent validity in this situation must be inferred from other tests of general scientific reasoning and problem-solving.

#### **7.5 Alignment**

A final element to ensure test validity is that the test is closely aligned to the content it is intended to measure. The NCDPI contracted with an outside provider to conduct an alignment study of the EOC Test of Biology to the 2004 Science *Standard Course of Study*. The report is not yet complete as of this writing.

Four elements of alignment will be quantified by the panelists:

**Categorical Concurrence:** *The categorical-concurrence criterion provides a very general indication of alignment if both documents incorporate the same content. The criterion of categorical concurrence between goals and assessment is met if the same or consistent categories of content appear in both documents. This criterion was judged by determining whether the assessment included items measuring content from each goal.*

*The analysis assumed that the assessment had to have at least six items measuring content from a goal in order for an acceptable level of categorical concurrence to exist between the goal and the assessment.*

**Depth of Knowledge:** *Depth-of-knowledge consistency between goals and assessment indicates alignment if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the goals.* For consistency to exist between the assessment and the goal, as judged in this analysis, at least 50% of the items corresponding to a goal had to be at or above the level of knowledge of the goal: 50%, a conservative cutoff point, is based on the assumption that a minimal passing score for any one goal of 50% or higher would require the student to successfully answer at least some items at or above the depth-of-knowledge level of the corresponding goal.

**Range of Knowledge:** The range-of-knowledge criterion is used to judge whether a comparable span of knowledge expected of students by a goal is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities. *The criterion for correspondence between span of knowledge for a goal and an assessment considers the number of standards within the goal with one related assessment item/activity. Fifty percent of the standards for a goal had to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable.*

**Balance of Representation:** *The balance-of-representation criterion is used to indicate the degree to which one standard is given more emphasis on the assessment than another.* An index is used to judge the distribution of assessment items. The index is computed by considering the difference in the proportion of standards and the proportion of hits assigned to the standard.

## **Chapter Eight: Quality Control Procedures**

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

### **8.1 Quality Control Prior to Test Administration**

Once test forms have been assembled, they are reviewed by a panel of subject experts. Once the review panel has approved a test form, test forms are then configured to go through the printing process. A .pdf file is sent directly to the printer once the final approval of the camera-ready copy has been obtained. Once all test answer sheets and booklets are printed, the operations specialist from NCDPI and the warehouse manager conduct a spot check of test booklets to ensure that all test pages are included and test items are in order.

### **8.2 Quality Control in Data Preparation and Test Administration**

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

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

### **8.3 Quality Control in Data Input**

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

### **8.4 Quality Control of Test Scores and Data Merging**

Once all tests are scanned, they are then sent through a secure system to the Regional Accountability Coordinators who check to ensure that all schools in all LEAs have completed and returned student test scores. The Regional Accountability Coordinators also conduct a spot check of data and then send the data through a secure server to the North Carolina Department of Public Instruction Division of Accountability Services. Data are then imported into a file and cleaned. When a portion of the data is in, NCDPI runs a CHECK KEYS program to flag areas where answer keys may need a second check.

As data come into the NCDPI, Student Information and Accountability Systems (a division of Technology and Information Services) staff import and clean data to ensure that individual student files are complete. Additionally, certain student demographic information is merged into the test data files from authoritative sources. For example, student Free and Reduced Lunch status is imported from the School Nutrition data collection activity. Other demographic variables that are imported from other data collections throughout the year are gender, ethnicity, and LEP status.

### **8.5 Quality Control in Reporting**

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

“The department hereby certifies the accuracy of the data from the North Carolina end-of-course tests for Fall 2007 provided that all NCDPI-directed test administration guidelines, rules, procedures, and policies have been followed at the district and schools in conducting proper test administrations and in the generation of the data. The LEAs may generate the required reports for the end-of-course tests as this completes the certification process for the EOC tests for the Fall 2007 semester.”

## Glossary of Key Terms

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

<b>Accommodations</b>	Changes made in the format or administration of the test to provide options to test takers who are unable to take the original test under standard test conditions.
<b>Achievement levels</b>	Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum classified by broad ranges of performance.
<b>Asymptote</b>	An item statistic that describes the proportion of examinees that endorsed a question correctly but did poorly on the overall test. Asymptote for a theoretical four-choice item is 0.25 but can vary somewhat by test. (For math it is generally 0.15 and for Science it is generally 0.22).
<b>Biserial correlation</b>	The relationship between an item score (right or wrong) and a total test score.
<b>Cut scores</b>	A specific point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.
<b>Dimensionality</b>	The extent to which a test item measures more than one ability.
<b>Embedded test model</b>	Using an operational test to field test new items or sections. The new items or sections are "embedded" into the new test and appear to examinees as being indistinguishable from the operational test.
<b>Equivalent forms</b>	Statistically insignificant differences between forms (i.e., the red form is not harder).
<b>Field test</b>	A collection of items to approximate how a test form will work. Statistics produced will be used in interpreting item behavior/performance and allow for the calibration of item parameters used in equating tests.

<b>Foil counts</b>	Number of examinees that endorse each foil (e.g. number who answer “A,” number who answer “B,” etc.)
<b>Item response theory</b>	A method of test item analysis that takes into account the ability of the examinee, and determines characteristics of the item relative to other items in the test. The NCDPI uses the 3-parameter model, which provides slope, threshold, and asymptote.
<b>Item tryout</b>	A collection of a limited number of items of a new type, a new format, or a new curriculum. Only a few forms are assembled to determine the performance of new items and not all objectives are tested.
<b>Mantel-Haenszel</b>	A statistical procedure that examines the differential item functioning (DIF) or the relationship between a score on an item and the different groups answering the item (e.g. gender, race). This procedure is used to identify individual items for further bias review.
<b>Operational test</b>	Test is administered statewide with uniform procedures and full reporting of scores, and stakes for examinees and schools.
<b>p-value</b>	Difficulty of an item defined by using the proportion of examinees who answered an item correctly.
<b>Parallel forms</b>	Covers the same curricular material as other forms
<b>Percentile</b>	The score on a test below which a given percentage of scores fall.
<b>Pilot test</b>	Test is administered as if it were “the real thing” but has limited associated reporting or stakes for examinees or schools.
<b>Raw score</b>	The unadjusted score on a test determined by counting the number of correct answers.
<b>Scale score</b>	A score to which raw scores are converted by numerical transformation. Scale scores allow for comparison of different forms of the test using the same scale.
<b>Slope</b>	The ability of a test item to distinguish between examinees of high and low ability.

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

## References

- Camera, W. J. & Echternacht, G. (2000). The SAT I and High School Grades: Utility in Predicting Success in College. *Research Notes RN-10, July 2000* (p.6). The College Board Office of Research and Development.
- Gregory, Robert J. (2000). *Psychological Testing: History, Principles, and Applications*. Needham Heights: Allyn & Bacon.
- Hambleton, Ronald K. (1983). *Applications of Item Response Theory*. British Columbia: Educational Research Institute of British Columbia.
- Hinkle, D.E., Wiersma, W., & Jurs, S. G. (1998). *Applied Statistics for the Behavioral Sciences* (pp. 69-70)
- Kolen, M.J., & Brennan, R.L. (1995). *Test equating methods and practices*. New York: Springer.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- Muraki, E., Mislevy, R.J., & Bock, R.D. (1991). PC-BiMain: Analysis of item parameter drift, differential item functioning, and variant item performance [Computer software]. Mooresville, IN: Scientific Software, Inc.
- Marzano, R.J., Brandt, R.S., Hughes, C.S., Jones, B.F., Presseisen, B.Z., Stuart, C., & Suhor, C. (1988). *Dimensions of Thinking*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Millman, J., and Greene, J. (1993). *The Specification and Development of Tests of Achievement and Ability*. In Robert Linn (ed.), *Educational Measurement* (pp. 335-366). Phoenix: American Council on Education and Oryx Press
- Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [December 13, 2007], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html>
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds), *Test Scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.
- Williams, V.S.L., Pommerich, M., & Thissen, D. (1998). A comparison of developmental scales based on Thurstone methods and item response theory. *Journal of Educational Measurement*, 35, 93-107.

## **Additional Resources**

- 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.
- Muraki, E., Mislevy, R.J., & Bock, R.D. PC-Bimain Manual. (1991). Chicago, IL: Scientific Software, Inc.
- North Carolina Department of Public Instruction. (1993). North Carolina End-of-Grade Testing Program: Background Information. Raleigh, NC: Author.
- North Carolina Department of Public Instruction. (1996). North Carolina Testing Code of Ethics. Raleigh, NC: Author.
- North Carolina State Board of Education. (1993). Public School Laws of North Carolina 1994. Raleigh, NC: The Michie Company.
- Nunnally, J. (1978). *Psychometric Theory*. New York: McGraw-Hill Book Company.
- Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill Book Company.
- SAS Institute, Inc. (1985). The FREQ Procedure. In SAS User's Guide: Statistics, Version 5 Edition. Cary, NC: Author.

## Appendix A. Test Specification Summaries

### *Biology*

Element	Comments
Purpose of the Test	<p>The North Carolina End-of-Course Tests are required by General Statute 115C-174.10 as a component of the North Carolina Annual Testing Program. As stated, the purposes of North Carolina state-mandated tests are “(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results.”</p> <p>For school, school system, and state accountability, prediction formulas (first used in 2000–2001) are used to determine growth expectations for end-of-course tests. The prediction formula is used to determine a student’s performance (average scores) on the North Carolina EOG or EOC tests, which serve as predictors of the same students’ performance in the EOC course where they are currently enrolled.</p> <p>Students entering the ninth grade for the first time in 2006–07 and beyond will be required to meet new exit standards. The exit standards will only apply to students following the Career Preparation, College/Technical Preparation, or College/University Preparation courses of study. These students will be required to perform at Achievement Level III (with one standard error of measurement) or above on the five EOC assessments (Algebra I, Biology, English I, Civics &amp; Economics, and U.S. History) and to successfully complete a senior project. Students following the Occupational Course of Study are required to meet rigorous exit standards as outlined in State Board of Education policy HSP-N-004 (16 NCAC 6D. 0503).</p>
Uses of the Test	<p>Student scores are used in determining student progress and proficiency under state-mandated Student Accountability Standards. According to State Board of Education policy, the standard for grade-level proficiency shall be a test score at Achievement Level III or above. Test results are also used to determine school, district, and state adequate yearly progress for the federal requirements per <i>No Child Left Behind</i>.</p>
Curriculum Cycle	<p>Test is based on the 2004 North Carolina Science <i>Standard Course of Study</i>.</p>

Content of the Test	
Subject/Course & Grade	Biology
Alignment	<p>The North Carolina End-of-Course Test of Biology assesses the 2004 North Carolina <i>Standard Course of Study</i>. Learners will study biological systems. The strands and unifying concepts provide a context for teaching content and process skill goals. Instruction should focus on the following unifying concepts:</p> <ul style="list-style-type: none"> <li>• Systems, Order and Organization</li> <li>• Evidence, Models, and Explanation</li> <li>• Constancy, Change, and Measurement</li> <li>• Evolution and Equilibrium</li> <li>• Form and Function</li> </ul> <p>The strands are: Nature of Science, Science as Inquiry, Science and Technology, and Science in Personal and Social Perspectives. They provide the context for teaching of the content goals and objectives.</p>
Dimensionality	The construct of the test is unidimensional, requiring the reporting of a total score for the test.
Weighting	By Goal
Obj. not/indirectly measured	To assess the student’s understanding of scientific inquiry and technological design, at least one question from each objective in Goal 1 will be tested in the context of the content of Goals 2 – 5. A variety of questions from each of the Goal 1 objectives will be included on every form.
Miscellaneous remarks	The test specifications committee and NCDPI – Curriculum staff made special requests about the following objectives: 2.02, 2.05, 3.01, 3.03, 3.05, 4.01, 4.02, 5.02
Design	
Population	All students enrolled in Biology in the state of North Carolina that want to receive credit for the course.
Format	
Item type	Multiple-choice: stem with four foils
Special item considerations	Items must be original and unique as well as free of bias (cultural, economic, gender, ethnic, or geographic). Distractors must be plausible and the language must be clear and precise.
Delivery Mode	

Mode	Paper & pencil	
Accommodations	As written in a student's IEP: Assistive Technology Devices, Braille, Braille Writer/Slate and Stylus, Cranmer Abacus, Dictation to Scribe, English/Native Language Dictionary or Electronic Translator, Home/Hospital Testing, Interpreter/Transliterators Signs/Cues Test, Keyboarding Devices, Large Print, Magnification Devices, Multiple Testing Sessions, One Test Item Per Page, Scheduled Extended Time, Student Reads Test Aloud to Self, Student Marks Answers in Test Book, Test Administrator Reads Test Aloud, Testing in Separate Room	
Number of Items (Total)	80	
Operational	60	
Embedded	20	
By section	20	
Time Limits	(NOTE: The current operational test is timed. The 2004 aligned tests will be untimed or at least will have limits that allow the vast majority of students to complete the assessment) TBD	
Universal Design	Items and selections are reviewed for accessibility by all students, specifically students with disabilities and students with limited English proficiency.	
<b>Item &amp; Test Characteristics</b>		
Item	Thinking Levels: Knowledge, Organizing, Applying, Analyzing, Generating, Integrating, Evaluating	
	<b>Lower-order thinking skills</b>	<b>Higher-order thinking skills</b>
	Knowledge	Analyzing
	Organizing	Generating
	Applying	Integrating
		Evaluating
	not more than <b>50%</b> per form	at least <b>50%</b> per form
	NCSCS Goal/Objective	
Difficulty Level(a priori): Easy, Medium, and Hard		
	<b>Level of Difficulty</b>	<b>Percent of items per form</b>

		Easy	25%
		Medium	50%
		Hard	25%
Test	Goal	Description of Category	Average Percentage
	1	The learner will develop abilities necessary to do and understand scientific inquiry.	28%-32%
	2	The learner will develop an understanding of the physical, chemical and cellular basis of life.	25%-30%
	3	The learner will develop an understanding of the continuity of life and the changes of organisms over time.	30%-35%
	4	The learner will develop an understanding of the unity and diversity of life.	25%-30%
	5	The learner will develop an understanding of the ecological relationships among organisms.	15%-20%
Cognitive Taxonomy	Dimensions of Thinking (Marzano et al.)		
Stimulus Materials	In order to assess students' ability to conduct investigations and make observations, each test form will contain 6–8 charts and tables; 8–10 graphs; and 8–12 diagrams, drawings, or other types of art		
Other Specifications			
Cost (Total: development, printing, shipping, etc.)			
Psychometric Properties			
P-value	.15<keep<.85 .85>reserve>.90 .10<reserve<.15		
Biserial Correlation	Keep >.25 Reserve >.15		
Slope	Keep >.7 .7 >Reserve >.5		
Asymptote	Keep<.35 .35<reserve<.45		
Threshold	-2.5<keep<2.5 2.5≤reserve≤3.0		

Dif Flags	.667<MH<1.5 not flagged
Minimum Reliability	.85 (multiple-choice)
Test Administration	
Guidelines & Procedures	Adhere to directions/script in <i>Test Administrator's Manual</i> .
Materials	Scrap paper for all students.
Testing Window	Last five weeks of traditional schedule. Last four weeks of semester or block schedule.
Scoring	
Methods	Scanned and scored locally (NCDPI-provided software)
Raw Scores	TBD
Scale Scores	TBD
Standard Setting	
Achievement Level Descriptors	<p style="text-align: center;">4 achievement levels</p> <p><b>Achievement Level I</b> Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level I recognize basic biological concepts and require extensive remediation to successfully complete the course. They can inconsistently identify the steps involved in the scientific process. Students at this level may recognize biological terms but demonstrate minimal understanding of the application of the terms.</p> <p><b>Achievement Level II</b> Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level II demonstrate limited understanding of biology concepts. These students can organize and interpret</p>

	<p>data with direct guidance. Students can identify basic cell structure/function, understand basic DNA structure, and solve simple Punnett squares. They can identify basic patterns of animal behavior and relationships among organisms in ecological systems.</p> <p><b>Achievement Level III</b> Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well prepared for a more advanced level in the content area.</p> <p>Students performing at Achievement Level III demonstrate mastery of biology concepts. These students understand and can conduct scientific inquiry, interpret and analyze data, and make predictions pertaining to various biological processes. They can understand form and function of biological systems. Students at this level interpret factors contributing to patterns of inheritance, animal behavior, and human health. They can assess how biotic and abiotic factors influence homeostasis within changing systems.</p> <p><b>Achievement Level IV</b> Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area.</p> <p>Students performing at Achievement Level IV demonstrate an advanced application of the scientific skills. These students demonstrate mastery of scientific inquiry skills including experimental design, identification of alternate interpretations of data, and application to real-world experiences. They can analyze the relationship between form and function of biological systems. They can also analyze and evaluate factors contributing to complex patterns of inheritance, animal behavior, and human health. Students at this level use their understanding of the universal genetic code to make predictions about unity and diversity of life.</p>
Method	Contrasting groups, item mapping
<b>Reporting</b>	
Levels of Reporting	Student, school, LEA, state
ABCs	Student report card, school report card, state report card

NCLB	Adequate Yearly Progress (AYP)
Appropriate Use	Measure of biology knowledge.
<i>History of Development</i>	
Committee Members	<p>Melanie Smith, Lead Consultant for Science Assessments, NCDPI Testing</p> <p>Laura Kramer, Senior Psychometrician, NCDPI Testing</p> <p>Bill Tucci, Section Chief for Science and Mathematics, NCDPI Instructional Services</p> <p>Eleanor Hasse, Science Consultant, High School, NCDPI Instructional Services</p> <p>Mildred Bazemore, Chief, Test Development Section, Accountability Services</p> <p>Lou Fabrizio, Director, Accountability Services</p> <p>Wandra Polk, Director, Instructional Services</p> <p>David Mills, Chief, Speech-Language, Areas of Exceptionality</p> <p>Tom Winton, Consultant, Assistive Technology, Areas of Exceptionality</p> <p>Frances Hoch, Chief, English as Second Language, Information &amp; Computer Skills</p> <p>Alesha McCauley, Consultant, English as Second Language, K–12</p>

*Chemistry*

Element	Comments
Purpose of the Test	<p>The North Carolina End-of-Course Tests are required by General Statute 115C-174.10 as a component of the North Carolina Annual Testing Program. As stated, the purposes of North Carolina state-mandated tests are “(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results.”</p> <p>For school, school system, and state accountability, prediction formulas (first used in 2000–2001) are used to determine growth expectations for end-of-course tests. The prediction formula is used to determine a student’s performance (average scores) on the North Carolina EOG or EOC tests, which serve as predictors</p>

	of the same students' performance in the EOC course where they are currently enrolled.
Uses of the Test	Student scores are used in determining student progress and proficiency under state-mandated Student Accountability Standards. According to State Board of Education policy, the standard for grade-level proficiency shall be a test score at Achievement Level III or above.
Curriculum Cycle	Test is based on the 2004 North Carolina Science <i>Standard Course of Study</i> .
Content of the Test	
Subject/Course & Grade	Chemistry
Alignment	<p>The Chemistry course encourages students to continue their investigations of the structure of matter along with chemical reactions and the conservation of matter and energy in those reactions. Inquiry is applied to the study of the composition, structure, properties and transformation of substances. The course focuses on basic chemical concepts and incorporates investigations to build understanding of these concepts. The unifying concepts and program strands provide a context for teaching content and process skill goals. All goals should focus on the unifying concepts:</p> <ul style="list-style-type: none"> <li>• Systems, Order and Organization</li> <li>• Evidence, Models, and Explanation</li> <li>• Constancy, Change, and Measurement</li> <li>• Evolution and Equilibrium</li> <li>• Form and Function</li> </ul> <p>The strands are: Nature of Science, Science as Inquiry, Science and Technology, Science in Personal and Social Perspectives. They provide the context for teaching of the content goals and objectives.</p>
Dimensionality	The construct of the test is unidimensional, requiring the reporting of a total score for the test.
Weighting	By goal
Obj. not/indirectly measured	To assess the student's understanding of scientific inquiry and technological design, at least one question from each objective in Goal 1 will be tested in the context of the content of Goals 2 – 5.

Miscellaneous remarks	The test specifications committee and NCDPI – Curriculum staff made special requests about the following objectives: 2.01, 2.03, 2.04, 2.05, 2.07, 3.01, 3.02, 3.03, 4.02, 4.03, 5.04.
Design	
Population	All students enrolled in Chemistry in the state of North Carolina that want to receive credit for the course.
Format	
Item type	-Multiple-choice: stem with four foils
Special item considerations	Items must be original and unique as well as free of bias (cultural, economic, gender, ethnic, or geographic). The reading level of the items must match the grade level of the test. Distractors must be plausible and the language must be clear and precise.
Delivery Mode	
Mode	Paper & pencil
Accommodations	As written in a student’s IEP: Assistive Technology Devices, Braille, Braille Writer/Slate and Stylus, Cranmer Abacus, Dictation to Scribe, English/Native Language Dictionary or Electronic Translator, Home/Hospital Testing, Interpreter/Transliterator Signs/Cues Test, Keyboarding Devices, Large Print, Magnification Devices, Multiple Testing Sessions, One Test Item Per Page, Scheduled Extended Time, Student Reads Test Aloud to Self, Student Marks Answers in Test Book, Test Administrator Reads Test Aloud, Testing in Separate Room
Number of Items (Total)	80
Operational	60
Embedded	20
By section	20
Time Limits	None. Estimated time is 162 minutes.

Universal Design	Items and selections are reviewed for accessibility by all students, specifically students with disabilities and students with limited English proficiency.										
Item & Test Characteristics											
Item	Thinking Levels: Knowledge, Organizing, Applying, Analyzing, Generating, Integrating, Evaluating										
	<b>Lower-order thinking skills</b>		<b>Higher-order thinking skills</b>								
	Knowledge		Analyzing								
	Organizing		Generating								
	Applying		Integrating								
			Evaluating								
	not more than <b>50%</b> per form		at least <b>50%</b> per form								
	NCSCS Goal/Objective										
Difficulty Level(a priori): Easy, Medium, and Hard											
		<table border="1"> <thead> <tr> <th>Level of Difficulty</th> <th>Percent of items per form</th> </tr> </thead> <tbody> <tr> <td>Easy</td> <td>25%</td> </tr> <tr> <td>Medium</td> <td>50%</td> </tr> <tr> <td>Hard</td> <td>25%</td> </tr> </tbody> </table>		Level of Difficulty	Percent of items per form	Easy	25%	Medium	50%	Hard	25%
Level of Difficulty	Percent of items per form										
Easy	25%										
Medium	50%										
Hard	25%										
Test	Goal	Description of Category	Average Percentage								
	1	The learner will develop abilities necessary to do and understand scientific inquiry.	28%-32%								
	2	The learner will develop an understanding of the structure and properties of matter.	30%-35%								
	3	The learner will build an understanding of regularities in chemistry.	25%-30%								
	4	The learner will build an understanding of energy changes in chemistry.	10%-15%								
	5	The learner will develop an understanding of chemical reactions.	25%-30%								
Cognitive Taxonomy	Dimensions of Thinking (Marzano et al.)										
Stimulus Materials	In order to assess students' ability to conduct investigations and make observations, each test form will contain 6–10 charts and tables; 7–10 graphs; and 12–16 diagrams, drawings, or other types of art.										
Other Specifications	NA										

Cost (Total: development, printing, shipping, etc.)	
Psychometric Properties	
P-value	.15<keep<.85 .85>reserve>.90 .10<reserve<.15
Biserial Correlation	Keep >.25 Reserve >.15
Slope	Keep >.7 .7 >Reserve >.5
Asymptote	Keep<.35 .35<reserve<.45
Threshold	-2.5<keep<2.5 2.5≤reserve≤3.0
Dif Flags	.667<MH<1.5 not flagged
Minimum Reliability	.85 (multiple-choice)
Test Administration	
Guidelines & Procedures	Adhere to directions/script in <i>Test Administrator's Manual</i> .
Materials	Scrap paper, scientific calculator and reference table.
Testing Window	Last five weeks of traditional schedule. Last four weeks of semester or block schedule.
Scoring	
Methods	Scanned and scored locally (NCDPI-provided software)
Raw Scores	TBD
Scale Scores	TBD

Standard Setting	
Achievement Level Descriptors	<p><b>Achievement Level Descriptors – Chemistry</b></p> <p><b>Achievement Level I</b>  Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level I have limited understanding with no mastery of chemical concepts. They have limited factual recall associated with atomic structure, nomenclature, and physical and chemical properties.</p> <p><b>Achievement Level II</b>  Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level II have a limited understanding of the structure and properties of matter, regularities, energy changes, and chemical reactions. Students demonstrate a limited ability to interpret and analyze data; and use reference materials. Students can balance simple equations, solve simple conversions, understand simple atomic structures, understand some periodic trends and substitute variables in given mathematical formulae.</p> <p><b>Achievement Level III</b>  Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well prepared for a more advanced level in the content area.</p> <p>Students performing at Achievement Level III demonstrate conceptual understanding and basic application of chemical concepts. They can analyze trends and relationships through interpretation of graphs and data and can write and balance equations including product prediction. They use their mathematical skills to solve equations and perform conversions. Students demonstrate an understanding of properties and structure of matter, energy changes, and kinetics.</p> <p><b>Achievement Level IV</b></p>

	<p>Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area.</p> <p>Students performing at Achievement Level IV demonstrate analysis, interpretation, and synthesis; understand the interrelationships of chemical concepts; and demonstrate a functional use of properties of structure and matter. They demonstrate superior comprehension of emission and absorption of EM energy as electrons change levels. Students calculate enthalpy changes for phase and temperature changes and interpret solubility rules to predict products for all types of reactions including net ionic reactions. They calculate stoichiometry multiple conversion problems and use IMFs to explain physical properties.</p>
Method	Contrasting groups, item mapping
Reporting	
Levels of Reporting	Student, school, LEA, state
ABCs	Student report card, school report card, state report card
NCLB	AYP
Appropriate Use	Measure of chemistry knowledge.
History of Development	
Committee Members	<p>Melanie Smith, Lead Consultant for Science Assessments, NCDPI Testing</p> <p>Laura Kramer, Senior Psychometrician, NCDPI Test Development</p> <p>Bill Tucci, Section Chief for Science and Mathematics, NCDPI Instructional Services</p> <p>Eleanor Hasse, Science Consultant, High School, NCDPI Instructional Services</p> <p>Clara Stallings, Science Consultant, 6 – 8, NCDPI Instructional Services</p> <p>Mike Jones, Science Content Specialist, NCSU/TOPS</p> <p>Myra Halpin, teacher, North Carolina School of Science and Mathematics</p>

	<p>Michael Jones, teacher, Pisgah High School Mildred Bazemore, Chief, Test Development Section, Accountability Services Lou Fabrizio, Director, Accountability Services Wandra Polk, Director, Instructional Services David Mills, Chief, Speech-Language, Areas of Exceptionality Tom Winton, Consultant, Assistive Technology, Areas of Exceptionality Frances Hoch, Chief, English as Second Language, Information &amp; Computer Skills Alesha McCauley, Consultant, English as Second Language, K- 12</p>
--	---

*Physical Science*

Element	Comments
Purpose of the Test	<p>The North Carolina End-of-Course Tests are required by General Statute 115C-174.10 as a component of the North Carolina Annual Testing Program. As stated, the purposes of North Carolina state-mandated tests are “(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results.”</p> <p>For school, school system, and state accountability, prediction formulas (first used in 2000–2001) are used to determine growth expectations for end-of-course tests. The prediction formula is used to determine a student’s performance (average scores) on the North Carolina EOG or EOC tests, which serve as predictors of the same students’ performance in the EOC course where they are currently enrolled.</p>
Uses of the Test	<p>Student scores are used in determining student progress and proficiency under state-mandated Student Accountability Standards. According to State Board of Education policy, the standard for grade-level proficiency shall be a test score at Achievement Level III or above.</p>
Curriculum Cycle	<p>Test is based on the 2004 North Carolina Science <i>Standard Course of Study</i>.</p>
Content of the Test	
Subject/Course & Grade	<p>Physical Science</p>
Alignment	<p>The Physical Science curriculum is designed to continue the investigation of the physical sciences begun in earlier grades. The Physical Science course will build a rich knowledge base to provide a foundation for the continued study of science. The investigations should be approached in both a qualitative and quantitative manner in keeping with the developing mathematical skills of the students. The unifying concepts and program strands provide a context for teaching content and process skill goals. All goals should focus on the unifying concepts:</p>

	<ul style="list-style-type: none"> <li>• Systems, Order and Organization</li> <li>• Evidence, Models, and Explanation</li> <li>• Constancy, Change, and Measurement</li> <li>• Evolution and Equilibrium</li> <li>• Form and Function</li> </ul> <p>The strands are: Nature of Science, Science as Inquiry, Science and Technology, Science in Personal and Social Perspectives. They provide the context for teaching of the content Goals and Objectives.</p>
Dimensionality	The construct of the test is unidimensional, requiring the reporting of a total score for the test.
Weighting	By goal
Obj. not/indirectly measured	To assess the student's understanding of scientific inquiry and technological design, at least one question from each objective in Goal 1 will be tested in the context of the content of Goals 2 – 6. A variety of questions from each of the Goal 1 objectives will be included on every form.
Miscellaneous remarks	The test specifications committee and NCDPI – Curriculum staff made special requests about the following objectives: 3.01, 3.02, 3.04, 4.02, 4.02, 5.02, 6.03.
Design	
Population	All students enrolled in Physical Science in the state of North Carolina that want to receive credit for the course.
Format	
Item type	Multiple-choice: stem with four foils
Special item considerations	Items must be original and unique as well as free of bias (cultural, economic, gender, ethnic, or geographic). The reading level of the items must match the grade level of the test. Distractors must be plausible and the language must be clear and precise.
Delivery Mode	
Mode	Paper & pencil
Accommodations	As written in a student's IEP: Assistive Technology Devices, Braille, Braille Writer/Slate and Stylus, Cranmer Abacus, Dictation to Scribe, English/Native Language Dictionary or Electronic Translator, Home/Hospital Testing, Interpreter/Transliterators/Signs/Cues Test, Keyboarding

	Devices, Large Print, Magnification Devices, Multiple Testing Sessions, One Test Item Per Page, Scheduled Extended Time, Student Reads Test Aloud to Self, Student Marks Answers in Test Book, Test Administrator Reads Test Aloud, Testing in Separate Room									
Number of Items (Total)	80									
Operational	60									
Embedded	20									
By section	20									
Time Limits	None. Estimated time is 162 minutes.									
Universal Design	Items and selections are reviewed for accessibility by all students, specifically students with disabilities and students with limited English proficiency.									
<b>Item &amp; Test Characteristics</b>										
Item	Thinking Levels: Knowledge, Organizing, Applying, Analyzing, Generating, Integrating, Evaluating									
	<b>Lower-order thinking skills</b>		<b>Higher-order thinking skills</b>							
	Knowledge		Analyzing							
	Organizing		Generating							
	Applying		Integrating							
			Evaluating							
	not more than <b>50%</b> per form		at least <b>50%</b> per form							
	NCSCS Goal/Objective									
	Difficulty Level(a priori): Easy, Medium, and Hard									
	<table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Level of Difficulty</th> <th>Percent of items per form</th> </tr> </thead> <tbody> <tr> <td>Easy</td> <td>25%</td> </tr> <tr> <td>Medium</td> <td>50%</td> </tr> <tr> <td>Hard</td> <td>25%</td> </tr> </tbody> </table>		Level of Difficulty	Percent of items per form	Easy	25%	Medium	50%	Hard	25%
Level of Difficulty	Percent of items per form									
Easy	25%									
Medium	50%									
Hard	25%									
Test	Goal	Description of Category	Average Percentage							

	1	The learner will develop abilities necessary to do and understand scientific inquiry.	38%-42%
	2	The learner will develop an understanding of forces and motion.	10%-15%
	3	The learner will analyze energy and its conservation.	15%-20%
	4	The learner will construct an understanding of electricity and magnetism.	15%-20%
	5	The learner will build an understanding of the structure and properties of matter.	15%-20%
	6	The learner will build an understanding of regularities in chemistry	30%-35%
Cognitive Taxonomy	Dimensions of Thinking (Marzano et al.)		
Stimulus Materials	In order to assess students' ability to conduct investigations and make observations, each test form will contain 10–12 charts and tables; 12–15 graphs; and 10–12 diagrams, drawings, or other types of art.		
Other Specifications	NA		
Cost (Total: development, printing, shipping, etc.)			
Psychometric Properties			
P-value	.15<keep<.85 .85>reserve>.90 .10<reserve<.15		
Biserial Correlation	Keep >.25 Reserve >.15		
Slope	Keep >.7 .7 >Reserve >.5		
Asymptote	Keep<.35 .35<reserve<.45		
Threshold	-2.5<keep<2.5 2.5≤reserve≤3.0		
Dif Flags	.667<MH<1.5 not flagged		
Minimum Reliability	.85 (multiple-choice)		

Test Administration	
Guidelines & Procedures	Adhere to directions/script in <i>Test Administrator's Manual</i> .
Materials	Scrap paper, four-function calculator and reference table.
Testing Window	Last five weeks of traditional schedule. Last four weeks of semester or block schedule.
Scoring	
Methods	Scanned and scored locally (NCDPI-provided software)
Raw Scores	TBD
Scale Scores	TBD
Standard Setting	
Achievement Level Ranges & Descriptors	<p>4 achievement levels</p> <p><b>Achievement Level I</b> Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area. Students performing at Achievement Level I do not have sufficient mastery of physical science concepts. They have minimal understanding of mechanics, energy, electricity and magnetism, wave motion and the nature of sound and light, structure and properties of matter, and regularities in chemistry.</p> <p><b>Achievement Level II</b> Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area. Students performing at Achievement Level II demonstrate inconsistent mastery of physical science concepts. They have limited understanding of mechanics, energy, electricity and magnetism, wave motion and the nature of sound and light, structure and properties of matter, and regularities in chemistry.</p> <p><b>Achievement Level III</b> Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well</p>

	<p>prepared for a more advanced level in the content area. Students performing at Achievement Level III demonstrate mastery of physical science concepts and are prepared for more advanced science courses. They have an adequate understanding of mechanics, energy, electricity and magnetism, wave motion and the nature of sound and light, structure and properties of matter, and regularities in chemistry.</p> <p><b>Achievement Level IV</b>  Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more Advanced level in the content area. Students performing at Achievement Level IV demonstrate superior understanding of physical science concepts and are very well prepared for more advanced science courses. They have an advanced level of understanding of mechanics, energy, electricity and magnetism, wave motion and the nature of sound and light, structure and properties of matter, and regularities in chemistry.</p>
Method	Contrasting groups, item mapping
Reporting	
Levels of Reporting	Student, school, LEA, state
ABCs	Student report card, school report card, state report card
NCLB	Adequate Yearly Progress (AYP)
Appropriate Use	Measure of physical science knowledge.
History of Development	
Committee Members	Melanie Smith, Lead Consultant for Science Assessments, NCDPI Testing Laura Kramer, Senior Psychometrician, NCDPI Testing Bill Tucci, Section Chief for Science and Mathematics, NCDPI Instructional Services Eleanor Hasse, Science Consultant, High School, NCDPI Instructional Services Clara Stallings, Science Consultant, 6 – 8, NCDPI Instructional

	<p>Services</p> <p>Mike Jones, Science Content Specialist, NCSU/TOPS</p> <p>Carolyn Elliott, teacher, South Iredell High School</p> <p>David English, teacher, Northside High School</p> <p>Barbara Wallace, teacher, North Gaston High School</p> <p>Mildred Bazemore, Chief, Test Development Section, Accountability Services</p> <p>Lou Fabrizio, Director, Accountability Services</p> <p>Wandra Polk, Director, Instructional Services</p> <p>David Mills, Chief, Speech-Language, Areas of Exceptionality</p> <p>Tom Winton, Consultant, Assistive Technology, Areas of Exceptionality</p> <p>Frances Hoch, Chief, English as Second Language, Information &amp; Computer Skills</p> <p>Alesha McCauley, Consultant, English as Second Language, K-12</p>
--	--

## Online Physics

Element	Comments
Purpose of the Test	<p>The North Carolina End-of-Course Tests are required by General Statute 115C-174.10 as a component of the North Carolina Annual Testing Program. As stated, the purposes of North Carolina state-mandated tests are “(i) to assure that all high school graduates possess those minimum skills and that knowledge thought necessary to function as a member of society; (ii) to provide a means of identifying strengths and weaknesses in the education process in order to improve instructional delivery; and (iii) to establish additional means for making the education system at the State, local, and school levels accountable to the public for results.”</p> <p>For school, school system, and state accountability, prediction formulas (first used in 2000–2001) are used to determine growth expectations for end-of-course tests. The prediction formula is used to determine a student’s performance (average scores) on the North Carolina EOG or EOC tests, which serve as predictors of the same students’ performance in the EOC course where they are currently enrolled.</p>
Uses of the Test	<p>Student scores are used in determining student progress and proficiency under state-mandated Student Accountability Standards. According to State Board of Education policy, the standard for grade-level proficiency shall be a test score at Achievement Level III or above.</p>
Curriculum Cycle	<p>Test is based on the 2004 North Carolina Science <i>Standard Course of Study</i>.</p>
Content of the Test	
Subject/Course & Grade	<p>Physics</p>
Alignment	<p>Physics, the most fundamental of the natural sciences, is quantitative in nature and uses the language of mathematics to describe natural phenomena. Inquiry is applied to the study of matter and energy and their interaction. Learners will study natural and technological systems. The program strands and unifying concepts provide a context for teaching content and process skill goals. All goals should focus on the unifying concepts:</p> <ul style="list-style-type: none"> <li>• Systems, Order and Organization</li> <li>• Evidence, Models, and Explanation</li> <li>• Constancy, Change, and Measurement</li> <li>• Evolution and Equilibrium</li> <li>• Form and Function</li> </ul>

	The strands are: Nature of Science, Science as Inquiry, Science and Technology, Science in Personal and Social Perspectives. They provide the context for teaching of the content Goals and Objectives.
Dimensionality	The construct of the test is unidimensional, requiring the reporting of a total score for the test.
Weighting	By goal
Obj. not/indirectly measured	To assess the student's understanding of scientific inquiry and technological design, at least one question from each objective in Goal 1 will be tested in the context of the content of Goals 2 – 8. A variety of questions from each of the Goal 1 objectives will be included on every form.
Miscellaneous remarks	
Design	
Population	All students enrolled in Physics in the state of North Carolina that want to receive credit for the course.
Format	
Item type	Multiple-choice: stem with four foils, simulations
Special item considerations	Items must be original and unique as well as free of bias (cultural, economic, gender, ethnic, or geographic). The reading level of the items must match the grade level of the test. Distractors must be plausible and the language must be clear and precise.
Delivery Mode	
Mode	Online
Accommodations	As written in a student's IEP: Assistive Technology Devices, Braille Writer/Slate and Stylus, Cranmer Abacus, Dictation to Scribe, English/Native Language Dictionary or Electronic Translator, Home/Hospital Testing, Interpreter/Transliterators Signs/Cues Test, Keyboarding Devices, Magnification Devices, Multiple Testing Sessions, Scheduled Extended Time, Student Reads Test Aloud to Self, Test Administrator Reads Test Aloud, Testing in Separate Room
Number of Items (Total)	84

Operational	63								
Embedded	21								
By section	21								
Time Limits	None. Estimated time is 162 minutes.								
Universal Design	Items and selections are reviewed for accessibility by all students, specifically students with disabilities and students with limited English proficiency.								
<b>Item &amp; Test Characteristics</b>									
Item	Thinking Levels: Knowledge, Organizing, Applying, Analyzing, Generating, Integrating, Evaluating								
	<b>Lower-order thinking skills</b>	<b>Higher-order thinking skill</b>							
	Knowledge	Analyzing							
	Organizing	Generating							
	Applying	Integrating							
		Evaluating							
	not more than <b>40%</b> per form	at least <b>60%</b> per form							
	NCSCS Goal/Objective  Difficulty Level(a priori): Easy, Medium, and Hard  <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th style="text-align: center;">Level of Difficulty</th> <th style="text-align: center;">Percent of items per form</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">Easy</td> <td style="text-align: center;">25%</td> </tr> <tr> <td style="text-align: center;">Medium</td> <td style="text-align: center;">50%</td> </tr> <tr> <td style="text-align: center;">Hard</td> <td style="text-align: center;">25%</td> </tr> </tbody> </table>		Level of Difficulty	Percent of items per form	Easy	25%	Medium	50%	Hard
Level of Difficulty	Percent of items per form								
Easy	25%								
Medium	50%								
Hard	25%								
Test	Goal	Description of Category							
	Average Percentage								
	1	The learner will develop abilities necessary to do and understand scientific inquiry.	33%-37%						
	2	The learner will build an understanding of linear motion.	10%-15%						
	3	The learner will build an understanding of two-dimensional motion including circular motion.	10%-15%						
4	The learner will develop an understanding of forces and Newton's laws of motion.	15%-20%							

	5	The learner will build an understanding of impulse and momentum.	10%-15%
	6	The learner will develop an understanding of energy as the ability to cause change.	15%-20%
	7	The learner will develop an understanding of wave motion and the wave nature of sound and light.	10%-15%
	8	The learner will build an understanding of static electricity and direct current electrical circuits.	10%-15%
Cognitive Taxonomy	Dimensions of Thinking (Marzano et al.)		
Psychometric Properties			
P-value	.15<keep<.85 .85>reserve>.90 .10<reserve<.15		
Biserial Correlation	Keep >.25 Reserve >.15		
Slope	Keep >.7 .7 >Reserve >.5		
Asymptote	Keep<.35 .35<reserve<.45		
Threshold	-2.5<keep<2.5 2.5≤reserve≤3.0		
Dif Flags	.667<MH<1.5 not flagged		
Minimum Reliability	.85 (multiple-choice)		
Test Administration			
Guidelines & Procedures	Adhere to directions/script in <i>Test Administrator's Manual</i> .		
Materials	Scrap paper, calculator, and Physics reference table for all students.		
Testing Window	Last five weeks of traditional schedule. Last four weeks of semester or block schedule.		
Scoring			

Methods	Scanned and scored locally (NCDPI-provided software)
Raw Scores	TBD
Scale Scores	TBD
Standard Setting	
Achievement Level Descriptors	<p style="text-align: center;">4 achievement levels</p> <p><b>Achievement Level I</b> Students performing at this level do not have sufficient mastery of knowledge and skills of the course to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level I demonstrate minimal understanding of the concepts related to physics. They can use reference tables to find and solve basic equations with limited manipulation. For example, they can calculate momentum given mass and velocity. Students can recognize the slope of a graph, can recognize linear, circular, and projectile motion, and can state Newton’s laws of motion. They can recognize different types of energy, basic wave characteristics, and a series versus a parallel circuit.</p> <p><b>Achievement Level II</b> Students performing at this level demonstrate inconsistent mastery of knowledge and skills of the course and are minimally prepared to be successful at a more advanced level in the content area.</p> <p>Students performing at Achievement Level II demonstrate a limited understanding of physics concepts used to predict outcomes. Students can solve basic problems involving manipulation of variables. They can compare data and have a limited ability to interpret graphical and tabular data. Students can follow a procedure in a lab but require assistance in design and analysis of laboratory experiments.</p> <p><b>Achievement Level III</b> Students performing at this level consistently demonstrate mastery of the course subject matter and skills and are well prepared for a more advanced level in the content area.</p>

	<p>Students performing at Achievement level III demonstrate mastery of the grade-level concepts related to physics. Students at this level begin to use abstract thinking to predict outcomes. Students can use graphical and mathematical tools to analyze linear and circular motion. Students can understand concepts of projectile motion and can quantitatively solve horizontally launched projectile problems. Students can analyze centripetal forces and net force on an object with horizontal and vertical force vectors. Students can consistently resolve vectors. Students can understand and apply conservation laws to solve problems. Students can graphically and mathematically analyze wave phenomena. Students can analyze a series or parallel circuit. Students can follow a procedure in a lab and only require limited assistance in design and analysis of laboratory experiments.</p> <p><b>Achievement Level IV</b> Students performing at this level consistently perform in a superior manner clearly beyond that required to be proficient in the course subject matter and skills and are very well prepared for a more advanced level in the content area.</p> <p>Students performing at Achievement Level IV demonstrate superior mastery of the grade-level concepts related to physics. They can think abstractly to solve new and unique problems, both qualitatively and quantitatively. Students can manipulate and solve complex, multi-step problems. Students can analyze, interpret and generate new information graphically, descriptively, and experimentally. Students can independently design, conduct, and analyze laboratory investigations.</p>
Method	Contrasting groups, item mapping
Reporting	
Levels of Reporting	Student, school, LEA, state
ABCs	Student report card, school report card, state report card
NCLB	Adequate Yearly Progress (AYP)

Appropriate Use	Measure of physics.
History of Development	
Committee Members	<p>Melanie Smith, Lead Consultant for Science Assessments, NCDPI Testing</p> <p>Laura Kramer, Senior Psychometrician, NCDPI Testing</p> <p>Bill Tucci, Section Chief for Science and Mathematics, NCDPI Instructional Services</p> <p>Eleanor Hasse, Science Consultant, High School, NCDPI Instructional Services</p> <p>Clara Stallings, Science Consultant, 6 – 8, NCDPI Instructional Services</p> <p>Mike Jones, Science Content Specialist, NCSU/TOPS</p> <p>David Corsetti, physics teacher, Needham Broughton High School</p> <p>Charlene Marsh, physics teacher, Trinity High</p> <p>Peter Sequeira, physics teacher, Thomasville High</p>

## Appendix B – Item-Development Guidelines

### Procedural Guidelines

1. Use the best-answer format.
2. Avoid writing complex multiple-choice items.
3. Format the items vertically, not horizontally.
4. Avoid errors of grammar, abbreviation, punctuation, and spelling.
5. Minimize student reading time.
6. Avoid tricky or misleading items.
7. Avoid the use of contractions.
8. Avoid the use of first or second person.

### Content Guidelines

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

### Stem Construction Guidelines

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

### General Foil Development

21. Each item must contain four foils (A, B, C, D).
  22. Order the answer choices in a logical order. Numbers should be listed in ascending or descending order.
  23. Each item written should contain foils that are independent and not overlapping.
  24. All foils in an item should be homogeneous in content and length.
  25. Do not use the following as foils: all of the above, none of the above, I don't know.
  26. Word the foils positively, avoiding any negative phrasing.  
The use of negatives, such as NOT and EXCEPT, is to be avoided.
  27. Avoid providing clues to the correct response. Avoid writing items so that phrases in the stem (clang associations) are repeated in the foils.  
Also avoid including ridiculous options.
  28. Avoid grammatical clues to the correct answer.
  29. Avoid specific determiners because they are so extreme that they are seldom the correct response. To the extent possible, specific determiners such as ALWAYS, NEVER,
- 
-

TOTALLY, and ABSOLUTELY should not be used when writing items. Qualifiers such as *best*, *most likely*, *approximately*, etc. should be bold and/or italic.

30. The correct response for items written should be evenly balanced among the response options. For a four-option multiple-choice item, correct responses should be located at each option position about 25 percent of the time.
31. The items written should contain one and only one best (correct) answer.

#### Distractor Development

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

## Appendix C. *Testing Code of Ethics*

### Testing Code of Ethics

---

#### Introduction

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

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

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

#### *Security*

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

#### *Preparation*

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

#### *Administration*

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

#### *Scoring, Analysis and Reporting*

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

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

---

---

### **.0306 TESTING CODE OF ETHICS**

- (a) This Rule shall apply to all public school employees who are involved in the state testing program.
  - (b) The superintendent or superintendent's designee shall develop local policies and procedures to ensure maximum test security in coordination with the policies and procedures developed by the test publisher. The principal shall ensure test security within the school building.
    - (1) The principal shall store test materials in a secure, locked area. The principal shall allow test materials to be distributed immediately prior to the test administration. Before each test administration, the building level test coordinator shall accurately count and distribute test materials. Immediately after each test administration, the building level test coordinator shall collect, count, and return all test materials to the secure, locked storage area.
    - (2) "Access" to test materials by school personnel means handling the materials but does not include reviewing tests or analyzing test items. The superintendent or superintendent's designee shall designate the personnel who are authorized to have access to test materials.
    - (3) Persons who have access to secure test materials shall not use those materials for personal gain.
    - (4) No person may copy, reproduce, or paraphrase in any manner or for any reason the test materials without the express written consent of the test publisher.
    - (5) The superintendent or superintendent's designee shall instruct personnel who are responsible for the testing program in testing administration procedures. This instruction shall include test administrations that require procedural modifications and shall emphasize the need to follow the directions outlined by the test publisher.
    - (6) Any person who learns of any breach of security, loss of materials, failure to account for materials, or any other deviation from required security procedures shall immediately report that information to the principal, building level test coordinator, school system test coordinator, and state level test coordinator.
  - (c) Preparation for testing.
    - (1) The superintendent shall ensure that school system test coordinators:
      - (A) secure necessary materials;
      - (B) plan and implement training for building level test coordinators, test administrators, and proctors;
      - (C) ensure that each building level test coordinator and test administrator is trained in the implementation of procedural modifications used during test administrations; and
      - (D) in conjunction with program administrators, ensure that the need for test modifications is documented and that modifications are limited to the specific need.
    - (2) The principal shall ensure that the building level test coordinators:
      - (A) maintain test security and accountability of test materials;
      - (B) identify and train personnel, proctors, and backup personnel for test administrations; and
      - (C) encourage a positive atmosphere for testing.
- 
-

- (3) Test administrators shall be school personnel who have professional training in education and the state testing program.
  - (4) Teachers shall provide instruction that meets or exceeds the standard course of study to meet the needs of the specific students in the class. Teachers may help students improve test-taking skills by:
    - (A) helping students become familiar with test formats using curricular content;
    - (B) teaching students test-taking strategies and providing practice sessions;
    - (C) helping students learn ways of preparing to take tests; and
    - (D) using resource materials such as test questions from test item banks, testlets and linking documents in instruction and test preparation.
- (d) Test administration.
- (1) The superintendent or superintendent's designee shall:
    - (A) assure that each school establishes procedures to ensure that all test administrators comply with test publisher guidelines;
    - (B) inform the local board of education of any breach of this code of ethics; and
    - (C) inform building level administrators of their responsibilities.
  - (2) The principal shall:
    - (A) assure that school personnel know the content of state and local testing policies;
    - (B) implement the school system's testing policies and procedures and establish any needed school policies and procedures to assure that all eligible students are tested fairly;
    - (C) assign trained proctors to test administrations; and
    - (D) report all testing irregularities to the school system test coordinator.
  - (3) Test administrators shall:
    - (A) administer tests according to the directions in the administration manual and any subsequent updates developed by the test publisher;
    - (B) administer tests to all eligible students;
    - (C) report all testing irregularities to the school system test coordinator; and
    - (D) provide a positive test-taking climate.
  - (4) Proctors shall serve as additional monitors to help the test administrator assure that testing occurs fairly.
- (e) Scoring. The school system test coordinator shall:
- (1) ensure that each test is scored according to the procedures and guidelines defined for the test by the test publisher;
  - (2) maintain quality control during the entire scoring process, which consists of handling and editing documents, scanning answer documents, and producing electronic files and reports. Quality control shall address at a minimum accuracy and scoring consistency.
  - (3) maintain security of tests and data files at all times, including:
    - (A) protecting the confidentiality of students at all times when publicizing test results; and
    - (B) maintaining test security of answer keys and item-specific scoring rubrics.
- (f) Analysis and reporting. Educators shall use test scores appropriately. This means that the educator recognizes that a test score is only one piece of information and must be interpreted together with other scores and indicators. Test data help
- 
-

educators understand educational patterns and practices. The superintendent shall ensure that school personnel analyze and report test data ethically and within the limitations described in this paragraph.

- (1) Educators shall release test scores to students, parents, legal guardians, teachers, and the media with interpretive materials as needed.
  - (2) Staff development relating to testing must enable personnel to respond knowledgeably to questions related to testing, including the tests, scores, scoring procedures, and other interpretive materials.
  - (3) Items and associated materials on a secure test shall not be in the public domain. Only items that are within the public domain may be used for item analysis.
  - (4) Educators shall maintain the confidentiality of individual students. Publicizing test scores that contain the names of individual students is unethical.
  - (5) Data analysis of test scores for decision-making purposes shall be based upon:
    - (A) disaggregation of data based upon student demographics and other collected variables;
    - (B) examination of grading practices in relation to test scores; and
    - (C) examination of growth trends and goal summary reports for state-mandated tests.
- (g) Unethical testing practices include, but are not limited to, the following practices:
- (1) encouraging students to be absent the day of testing;
  - (2) encouraging students not to do their best because of the purposes of the test;
  - (3) using secure test items or modified secure test items for instruction;
  - (4) changing student responses at any time;
  - (5) interpreting, explaining, or paraphrasing the test directions or the test items;
  - (6) reclassifying students solely for the purpose of avoiding state testing;
  - (7) not testing all eligible students;
  - (8) failing to provide needed modifications during testing, if available;
  - (9) modifying scoring programs including answer keys, equating files, and lookup tables;
  - (10) modifying student records solely for the purpose of raising test scores;
  - (11) using a single test score to make individual decisions; and
  - (12) misleading the public concerning the results and interpretations of test data.
- (h) In the event of a violation of this Rule, the SBE may, in accordance with the contested case provisions of Chapter 150B of the General Statutes, impose any one or more of the following sanctions:
- (1) withhold ABCs incentive awards from individuals or from all eligible staff in a school;
  - (2) file a civil action against the person or persons responsible for the violation for copyright infringement or for any other available cause of action;
  - (3) seek criminal prosecution of the person or persons responsible for the violation; and (4) in accordance with the provisions of 16 NCAC 6C .0312, suspend or revoke the professional license of the person or persons responsible for the violation.

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

*Amended Eff. August 1, 2000.*