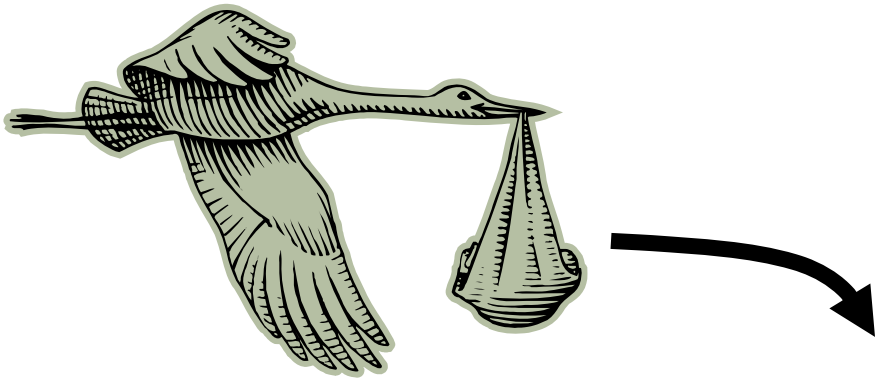
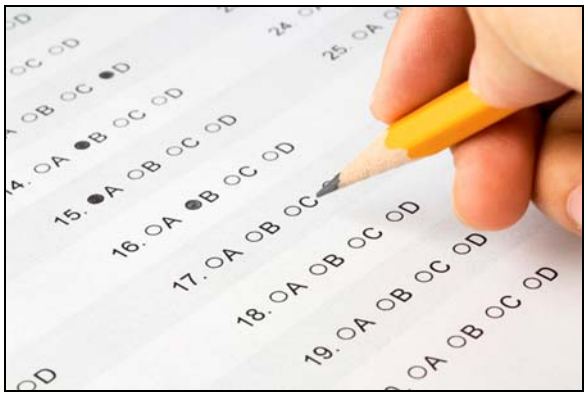
A BRIEF LOOK AT:

E&R Report No. 09.26

WHERE DO NORTH CAROLINA'S END-OF-GRADE (EOG) AND END-OF-COURSE (EOC) TESTS COME FROM?

Author: Brad McMillen





This document is a brief description of the test development process for the End-of-Grade and End-of-Course tests that are part of the North Carolina Testing Program. Most of the information presented here can be found on the North Carolina Department of Public Instruction (NCDPI) Web site, particularly in the Technical Manuals for those tests (see http://www.ncpublicschools.org/accountability/testing/technicalnotes). All information herein is reproduced or adapted from those sources with permission from NCDPI. The Wake County Public School System (WCPSS) would like to thank Nadine McBride of NCDPI for her feedback and comments on an earlier draft of this document.

WAKE COUNTY PUBLIC SCHOOL SYSTEM

www.wcpss.net/evaluation-research

The Test Development Cycle

Each End-of-Grade (EOG) and End-of-Course (EOC) test administered in North Carolina is developed via a detailed sequence of events that takes approximately four years to complete. The test development cycle begins when the state adopts a new curriculum for a particular subject area. From that point, the test specifications are developed; test items that measure the learning goals in the new curriculum are written and reviewed for quality; the new test items are field-tested and reviewed again; and then finally, the new test is put into place. The entire process is displayed in Table 1. Each of the remaining sections of this document then provides more detailed information about this process.

Table 1
The NC Test Development Process

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

[a]Activities done only at implementation of new curriculum
[b]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
Source: NC Mathematics Tests Edition 3 – Technical Report, p. 7 (Retrieved from http://www.ncpublicschools.org/docs/accountability/reports/mathtechmanualdrafted2.pdf on April 5, 2009).

Test Specifications or "Blueprints"

All EOG and EOC tests are derived from the state's official curriculum - the North Carolina Standard Course of Study (SCOS). The NC SCOS lists the specific competencies and skills that students are expected to master in each grade level and subject area. The entire SCOS for grades K through 12 can be found at http://www.ncpublicschools.org/curriculum/.

When a new curriculum for a particular subject or course is adopted by the state, the test development process begins. The new SCOS is used to develop the specifications (i.e., blueprint) for the new test. This blueprint specifies how many test items are to be developed for each goal and objective in the curriculum, the intended difficulty level of the items, and the percentage of test items that involve different levels of thinking skills (e.g., knowledge recall, analysis, application, etc.).

When tests are created, a test specification committee develops specific targets around how many test items should come from each of the goals and objectives (or strands) in the curriculum. This test specification committee consists of individuals from DPI Divisions of Accountability, Academic Services and Instructional Support (Curriculum), Exceptional Children, as well as content teachers, teachers of students with disabilities and English Language Learners, LEA curriculum specialists, and other experts and stakeholders. In consultation with the test specification committee, DPI decides approximately how much emphasis each strand of the curriculum should receive on the test, and the distribution of test items is often purposely uneven. An example of how test items are distributed across strands from the current Math curriculum is provided in Table 2. These percentages will vary by test and grade level, as will the priority of objective coverage.

Table 2
Distribution of EOG Test Items Across Curriculum Strands, 4th Grade Mathematics

Curriculum Strand	% of Test Items	Priority of Objective Coverage
The learner will read, write, model and compute with non-negative rational numbers (Number Sense)	35-40%	1.03, 1.04, 1.01 (emphasis is on decimals), 1.02
The learner will understand and use perimeter and area (Measurement)	10-12%	N/A
The learner will recognize and use geometric properties and relationships (Geometry)	10-12%	3.02, 3.03, 3.01
The learner will understand and use graphs, probability, and data analysis (Data Analysis and Probability)	15-18%	4.02, 4.01, 4.04, 4.03
The learner will demonstrate an understanding of mathematical relationships (Algebra)	20-25%	5.01, 5.02, 5.03

Source: NC Mathematics Tests Edition 3 – Technical Report, Appendix B, p. 134 (Retrieved from http://www.ncpublicschools.org/docs/accountability/reports/mathtechmanualdrafted2.pdf on May 15, 2009).

During the item development process, items are written with this distribution in mind so that the final pool of items will match these specifications. The state also specifies that items be distributed in a certain fashion with respect to the type of thinking skill required to answer the item correctly. The framework used in this process is adapted from one popularized by Robert Marzano in the book *Dimensions of Thinking* (Marzano, Presseisen, Jones, Suhor, & Brandt, 1988), and is detailed in Table 3. The thinking skills in the taxonomy are listed in order of complexity; however, both the curriculum and the test acknowledge the importance of each skill. According to DPI, approximately 60-65% of the test items on an EOG or EOC test are written at or above the *analyzing* level of the taxonomy[1].

Table 3
Thinking Skills Underlying EOG and EOC Test Items

Thinking Skill	**Specific Examples**
Knowing	• Defining problems: clarifying needs, discrepancies, or puzzling situations • Setting goals: establishing direction and purpose • Observing: obtaining information through one or more senses • Formulating questions: seeking new information through inquiry • Encoding: storing information in long-term memory • Recalling: retrieving information from long-term memory **Useful Verbs: list, name, label, recall, identify, match, choose**
Organizing	• Arranging information so it can be used effectively • Comparing: noting similarities and differences between or among entities • Classifying: grouping and labeling entities on the basis of their attributes • Ordering: sequencing entities according to a given criterion • Representing: changing the form but not the substance of information **Useful Verbs: categorize, group, classify, compare, contrast**
Applying	• Demonstrating prior knowledge within a new situation. The task is to bring together the appropriate information, generalizations, or principles that are required to solve a problem. **Useful Verbs: apply, make, show, record, construct, demonstrate, illustrate**
Analyzing	• Clarifying existing information by examining parts and relationships • Identifying attributes and components: determining characteristics or parts of something • Identifying relationships and patterns: recognizing ways in which elements are related • Identifying main idea: identifying the central element; for example, the hierarchy of key ideas in a message or line of reasoning • Identifying errors: recognizing logical fallacies and other mistakes and, where possible, correcting them **Useful Verbs: outline, diagram, differentiate, analyze**

[1] This is based on test specifications for the current editions of the mathematics tests (Retrieved from http://www.ncpublicschools.org/docs/accountability/reports/mathtechmanualdrafted2.pdf on May 15, 2009).

Thinking Skill	Specific Examples
Generating	• Producing new information, meaning, or ideas • Inferring: going beyond available information to identify what reasonably may be true • Predicting: anticipating next events or the outcome of a situation • Elaborating: explaining by adding details, examples, or other relevant information **Useful Verbs: conclude, predict, explain, elaborate, infer**
Integrating	• Connecting and combining information • Summarizing: combining information efficiently into a cohesive statement • Restructuring: changing existing knowledge structures to incorporate new information **Useful Verbs: combine, summarize, design, imagine, generalize**
Evaluating	• Assessing the reasonableness and quality of ideas • Establishing criteria: setting standards for making judgments • Verifying: confirming the accuracy of claims **Useful Verbs: judge, evaluate, rate, verify, assess, define criteria**

Source: Adapted from online training materials produced by NCDPI and North Carolina State University (Retrieved from https://cuacs8.mck.ncsu.edu/moodle/course/ on April 22, 2009).

Test Item Development

EOG and EOC test items are created in a multiple-choice format, with four possible responses (called *foils*) for each item. One of the foils is the correct answer, and the other three incorrect foils are referred to as *distractors*. The actual test question itself is referred to as the *stem*.

Next, the state identifies and trains item writers to begin drafting possible test items. These item writers are generally North Carolina educators who have expertise in the content area that the test will be measuring. During training, these item writers are given specific guidelines as to how to write high-quality test items and how to ensure that those items are consistent with the test specifications. As an example, a list of general item development guidelines used for the new EOG mathematics tests implemented in 2006 is listed next.

Item Development Guidelines for Mathematics Tests

Content Guidelines

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

Procedural Guidelines

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

Stem Construction Guidelines

15. Items are to be written in the question format.
16. Ensure that the directions written in the stems are clear and that the wording lets the students know exactly what is being tested.
17. Avoid excessive verbiage when writing the 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 where phrases in the stem (clang associations) are repeated in the foils.
28. Avoid including ridiculous options.
29. Avoid grammatical clues to the correct answer.
30. 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 italic.
31. The correct response for items written should be evenly balanced among the response options. For a 4-option multiple-choice item, each correct response should be located at each option position about 25% of the time.
32. The items written should contain one and only one best (correct) answer.

Distractor Development

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

Source: NC Mathematics Tests Technical Report, March 2006, p. 102-103 (Retrieved from http://www.ncpublicschools.org/docs/accountability/testing/mathtechmanual.pdf on May 15, 2009).

Field Testing and Analysis

Once a pool of potential items is written, the items are *field-tested*. This process typically occurs by mixing these items in with the actual test items that are currently in use across the state. Therefore, whenever a student is taking an EOG or EOC test, that test typically consists of actual test items that will determine the student's score, along with a small number of field-test items which will not be used to determine their score. The student does not know which items count toward their score or which are being field-tested. However, because the student responds to all of those questions, the

state can then analyze the results of the field test questions to see if they are good enough to include on the new test that is being developed.

This assessment of the quality of field-test items occurs on several levels. Statistical analyses are performed to determine a number of things, including whether each field-test item:

- is too easy or too difficult;
- is more likely to be answered incorrectly by certain groups of students (boys vs. girls, Caucasian students vs. non-Caucasian students, etc.);
- has an unusual response pattern that does not adequately correlate to responses to other similar test items

The results of these statistical analyses are also reviewed by curriculum experts to determine whether any items contain any content or wording that might be biased in favor or against students from different backgrounds, and whether the items are not appropriately linked to the curriculum goals and objectives that they are supposed to be measuring. Items that are identified as problematic through this process are then removed from the item pool.

Creating the Pilot Tests

Once a sufficient pool of items is established, the state assembles multiple forms of the test which are referred to as pilot tests. There are typically multiple forms of the test at this point in the process, such that not every form of the test contains every single question. Each form of the test contains a portion of the items from the pool, and each form contains items from across the various goals and objectives in the curriculum. However, each form is built to the same test specifications, previously described. This sampling of items across multiple forms of the test helps to keep the administration time at a reasonable level given the ages of the students who will be taking the tests. Essentially, a balance is struck so that the test is short enough for students to complete without getting fatigued, but long enough to constitute a good, reliable measure of the student's mastery of the curriculum.

In order to ensure that various forms of the test are equal in difficulty, each test item is measured by something called a *p-value*, which is the probability that a given student will answer the question correctly. Since the tests measure the curriculum that all students are expected to learn, there is a common misconception that most of the items should be answered correctly by a student if they truly learned what they were supposed to learn. However, EOG and EOC tests are not built in that manner; they are built to purposely include a mixture of easy, medium, and hard items. On most forms of EOG and EOC tests, *a typical test item should be answered correctly by only 50-60% of the students who take it*. While only half of students answering a 4-option multiple-choice item correctly may seem low—especially when students have a 25% chance of getting the item correct just by guessing - creating tests with such a high difficulty level is important for statistical purposes in order to accurately measure how much students across all ability levels have learned. The average p-values (i.e., the proportion of items that an average student should answer correctly across all forms of the test) for the current EOG Mathematics tests are shown in Table 4.

Table 4
Number of items and average item difficulty, EOG Mathematics Tests (3rd Edition)

Grade Level	# test items per form	average p value*
Grade 3 Pretest	40	0.60
Grade 3	80	0.66
Grade 4	80	0.59
Grade 5	80	0.54
Grade 6	80	0.51
Grade 7	80	0.49
Grade 8	80	0.46

Note: * An average p-value of 0.59 means that the percentage of students who get a given item correct will be around 59%.
Source: NC Mathematics Tests Technical Report, March 2006, p. 30 (Retrieved on May 15, 2009 from: http://www.ncpublicschools.org/docs/accountability/testing/mathtechmanual.pdf)

Standard Setting (i.e., "Cut Scores")

Once tests are developed and administered for the first time, they are generally referred to as pilot tests. In order for the test to be deemed *operational* (i.e., a completed, official test), NCDPI has to engage in a standard setting process[2]. This process determines the score points that separate the four achievement levels that the state uses to report test results. The first two achievement levels theoretically contain score points that represent insufficient mastery of the tested material. Achievement Level Descriptors are provided for each test that describes what a typical child at each level is expected to know and do. An example of how the various score points on the scale are divided into four achievement levels is presented in Table 5.

Table 5
Achievement Level Ranges for Selected EOC and EOG Tests, 2007-08 School Year

Test	Level I	Level II	Level III	Level IV
EOC - Algebra I	139 and below	140-147	148-157	158 or higher
EOC - English I	137 and below	138-145	146-156	157 or higher
EOG - 3rd Grade Math	311-328	329-338	339-351	352-370

There are a number of established methods for setting achievement level cut scores on educational tests, and the state has employed several different methods over the years. Most methods rely on some kind of expert judgment about either the abilities of the tested students or the difficulty of individual test items, or a combination of both. In all methods used by DPI, teachers and LEA

[2] In some cases, where an operational test must be given every year by law (e.g., Reading and Mathematics End-of-Grade Tests, etc.), the pilot test is simultaneously used as an "operational" test. In other cases, (e.g., Science End-of-Grade Tests in 2007-08), the pilot test results are not used officially, and the test does not become operational until the following year.

curriculum specialists provide input and guidance as a part of the standard setting process. These methods are part art and part science, and they generally require some kind of consensus-building process to determine where the cut scores should fall.

This standard setting process occurs after the first statewide administration of the pilot test. It can take a number of weeks or even months to complete after a pilot test is administered, depending on which method(s) are used. For this reason, the public release of scores from the first administration of a new test is usually delayed. Schools and students may have to wait weeks or even months to get the results of those tests.

Once the standard setting process is finalized, and the results of that process are applied to the test, it becomes a fully "operational" test, and the results are reported publicly. That test then remains in place as the official operational test of that curriculum until the curriculum changes (typically every 5 years or so, at the State Board of Education's discretion).

Test Administration

Once the operational test is developed, all of the aforementioned work will have helped ensure that the test is a reliable and valid measure of students' mastery of the curriculum. All of that work may go for naught, however, if the test is not administered under tightly controlled, standardized conditions. School districts receive training from NCDPI each year on the proper administration procedures for every test, and they in turn are required to train a designated Test Coordinator at each school in those procedures. This training must occur every year before the administration of every type of test, and Test Coordinators must attend this training each year even if they have served in that role in past years.

These procedures include guidelines for how test materials are to be handled and stored at the school, the organization of the room where the test is being administered, the specific words that are to be spoken when giving students instructions at the beginning of a testing session, and numerous other detailed steps that must be followed to ensure that the test is experienced in the same way by every student in the state. Failing to administer the test according to these procedures may result in the school district declaring a "misadministration", which essentially means that the scores of the students involved are deemed to be invalid. In that case, those students would be required to retake another form of the test on another occasion in order to obtain valid test scores.

Final Thoughts

This document is an attempt to briefly describe the long and arduous (yet essential) processes that occur leading up to the administration and scoring of the tests in the North Carolina statewide testing program. While a great deal of emphasis is placed on the test results that are the end product of this process, clearly there is a tremendous amount of work that must occur before that can happen. The general public and most educators tend to experience the testing program only at its very end stage, where the actual results of the test are publicly reported for individual students and schools. A basic understanding of the entire process, however, can help to debunk various myths about these tests, and to help educators and the general public understand and evaluate the logical foundation on which these tests are based.

Resources Related to the NC Statewide Testing Program

WCPSS Information on Testing and Accountability:
http://www.wcpss.net/evaluation-research/abcs_faqs.html

Testing Information for Parents from NCDPI:
http://www.ncpublicschools.org/accountability/parents/

NCDPI Test Development Process Outline
http://dpi.state.nc.us/docs/accountability/testing/policies/mctestdevelopment/RevisedTestDevelopmentProcessFinalFinal.pdf

Sample Test Items:

Math:
http://www.ncpublicschools.org/accountability/testing/eog/sampleitems/math
Reading:
http://www.ncpublicschools.org/accountability/testing/eog/sampleitems/reading
End-of-Course Tests:
http://www.ncpublicschools.org/accountability/testing/eoc/

Statewide Testing Results:
http://www.ncpublicschools.org/accountability/reporting/

WCPSS Testing Results:
http://www.wcpss.net/evaluation-research/reports/index.html
http://www.wcpss.net/test-scores/

References

Marzano, J. R., Presseinsen, B. Z., Jones, B. F., Suhor, C., & Brandt, R. S. (1988). *Dimensions of Thinking: A Framework for Curriculum & Instruction.* Alexandria, VA: Association for Supervision & Curriculum Development.