SBE Policy: Multiple-Choice Test Development

The official process for the development of state tests included in the North Carolina State Testing Program is as follows. The flowchart depicts the steps in the test development process for the state tests. A written description of each step in the test development process is included.

Questions regarding the Test Development Process should be directed to:

NC Department of Public Instruction
Accountability Services Division
Test Development Section
6314 Mail Service Center
Raleigh, NC 27699-6314

(919) 807-3774
# North Carolina Testing Program
## Test Development Process Flow Chart

| Step 1<sup>a</sup> | Step 2<sup>b</sup> | Step 3<sup>b</sup> | Step 4 | Step 5 | Step 6<sup>b</sup> | Step 7 | Step 8<sup>b</sup> | Step 9<sup>b</sup> | Step 10 | Step 11<sup>b</sup> | Step 12<sup>b</sup> | Step 13 | Step 14<sup>b</sup> | Step 15 | Step 16 | Step 17 | Step 18<sup>ab</sup> | Step 19 | Step 20<sup>ab</sup> | Step 21<sup>b</sup> | Step 22 |
|---------------------|-------------------|-------------------|--------|--------|-------------------|--------|-------------------|-------------------|--------|-------------------|-------------------|--------|-------------------|--------|--------|-------------------|-------------------|-------------------|-------------------|--------|
| Adopt Content Standards | Review Item Tryout Statistics | Conduct Bias Reviews | Review Item Tryout Forms | Assemble Field Test Forms | Review Field Test Forms | Review Item Tryout Forms | Assemble Equivalent and Parallel Forms | Review Items for Field Test | Review Assembled Test | Assemble Field Test Forms | Administer Field Test | Review Item Tryout Forms | Develop Test Specifications (Blueprint) | Develop New Items | Review Field Test Forms | Final Review of Test | Administer Test as Pilot | Score Test | Establish Standards | Administer Test as Fully Operational | 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:**  
1. For novel items or new curriculum, item tryouts should precede field-testing items.  
2. Professional development opportunities are integral and ongoing to the curriculum and test development process.
North Carolina Testing Program

TEST DEVELOPMENT PROCESS

Introduction

North Carolina tests are curriculum-based tests designed to measure the objectives found in the state-adopted content standards. The responsibility of updating the state-adopted content standards falls to the North Carolina Department of Public Instruction (NCDPI) K-12 Curriculum and Instructional Division. Curriculum specialists, teachers, administrators, university professors, and others assist in the process of updating curricula. Once curricula are adopted or tested objectives are approved by the North Carolina State Board of Education, in areas where statewide tests are required, the test development process begins.

The state-adopted content standards are periodically reviewed for possible revisions; however, test development is continuous. The NCDPI Accountability Services/Test Development Section test development staff members begin developing operational test forms for the North Carolina Testing Program when the State Board of Education (SBE) determines that such tests are needed. The need for new tests may result from mandates from the federal government or the North Carolina General Assembly. New tests can also be developed if the SBE determines the development of a new test will enhance the education of North Carolina students. The test development process consists of six phases and takes approximately four years. The phases begin with the development of test specifications and end with the reporting of operational test results.

PHASE 1: DEVELOP THE TESTING PLAN

Step 1: Develop the Test Specifications (Blueprint)

Prior to developing test specifications, it is important to outline the purpose of a test and what types of inferences (e.g. diagnostic, curriculum mastery) are to be made from test scores. Millman and Greene (1993, in Robert Linn, ed)\(^1\) offer a rationale for delineating the purpose of the test. “A clear statement of the purpose provides the overall framework for test specification, item development, tryout, and review. A clear statement of test purpose also contributes significantly to appropriate test use in practical contexts.” Using a test’s purpose as the guiding framework, NCDPI curriculum specialists, teachers, NCDPI test development staff, and other content, curriculum, and testing experts establish the test specifications for each of the grade levels and content areas assessed. In general, test specifications include the following:

\( (1) \) Percentage of questions from higher or lower thinking skills and classification of each test question in the two dimensions of difficulty\(^2\) and thinking skill level\(^3\).

---


\(^2\)Difficulty Level. Difficulty level describes how hard the test questions are. Easy questions are ones that about 70 percent of the students would answer correctly. Medium test questions are ones that about 50 percent to 60 percent of the students would
(2) Percentage of item types such as multiple choice, constructed response, technology-enhanced, or stimulus-based and other specialized constraints.

(3) Percentage of test questions that measure a specific goal, objective, domain, or category.

(4) For tests that contain reading selections, the percentage of types of selections (e.g., literary vs. informational texts, etc.).

(5) For tests of mathematics, the percentage of questions where a student is allowed to use a calculator.

PHASE 2: ITEM DEVELOPMENT (ITEM TRYOUTS AND REVIEW)

Step 2: Develop Test Items

While objectives for the new curriculum might not yet be implemented in the field, there are larger ideas that carry over from the previous curriculum cycle. These objectives are known as common curriculum objectives. Items can be developed from old test items that are categorized as common curriculum items or they can be developed as new items.

Old test items include those items from the previous curriculum cycle that were developed but not field tested. They can also be items that were field tested but not used in the statewide operational administration. If a curricular match is found for certain items, these items will be retained for further development with the new curriculum and tests. Items may be switched from grade to grade or from course to course to achieve a curriculum match. For example, a mathematics item may be moved from grade 5 to grade 4. If they are moved from grade to grade or course to course, they are considered to be new curriculum objective items. If they remain in the same grade or course, they are considered to be common curriculum items. Any item that has been used in a statewide operational test that matches the new curriculum may be released for training or for teachers to use in the classroom.

In many cases, the purpose of the item tryout is to examine item types that the students have not previously been exposed to. In those cases, the items must be newly developed and will follow the process outlined in Step 8. While additional training may be required for writing new item types, the teachers can begin item development of common curriculum items due to their existing familiarity with the content.

answer correctly. Hard test questions are ones that only about 20 percent or 30 percent of the students would answer correctly. Difficulty level may be estimated based on judgment prior to statistics having been collected on the items or statistically determined through field testing.

Thinking Skill Level. Thinking skill level describes the cognitive skills that a student must use to solve the problem or respond to the question. One test question may ask a student to classify several passages based on their genre; another question may ask the student to select the best procedure to use for solving a problem. Passages are selected on other criteria, including readability. They must be interesting to read, be complete (with a beginning, middle, and end), and be from sources students might actually read. Advisory Groups, curriculum specialists, the NCDPI Division of Instructional Services, and the NCDPI Division of Accountability Services/Testing Section select passages for state tests.

NCDPI Test Development Section reserves the right to waive the “item tryout” component if time and other resources do not support the practice or if requirements for field testing are limited.
Step 3: Review Items for Tryouts

The review process for items developed for the item tryout is the same as it would be for the review of newly written items developed for any statewide test. The review process is described in detail in the “Phase 3 Field Test Development” section. In some cases where there are new item types developed that are different from what had previously been seen by students, additional reviews may be incorporated.

Step 4: Assemble Item Tryout Forms

As time and other resources permit, item tryouts are conducted as the first step in producing new tests. Item tryouts are 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 may be tested. Conducting item tryouts has several advantages. One important advantage is that an opportunity exists, during this process, to provide items for field-testing that are known to be psychometrically sound. In addition, it provides an opportunity to refine a new or novel type of item, such as technology-enhanced items, for presentation to students. Having this data prior to field-testing and operational testing informs the item development and the test development process.

Step 5: Review Item Tryout Forms

Content specialists at the NCDPI Test Development Section and the Technical Outreach for Public Schools (TOPS) review the item tryout forms for clarity, correctness, potential bias, and curricular appropriateness. The NCDPI staff members, who specialize in the education of children with special needs, also review the forms.

Step 6: Administer Item Tryouts

When item tryouts are administered as a stand-alone test, a limited number of forms are produced, thus minimizing the number of children and schools impacted. Once these items are embedded in operational forms, the types of novel items that can be evaluated are severely constrained.

Item tryouts may include additional research, such as think-alouds or the evaluation of item modifications. Such research allows for the refinement of items for field testing.

Step 7: Review Item Tryout Statistics

Item statistics are examined to determine items that have a poor curricular match, poor response choices (foils), and confusing language. In addition, differential item functioning analyses can be run and a bias committee can review flagged items for revision. During a first-year item tryout, timing data can be collected to determine how long the new tests should be or to determine the amount of time needed for a given number of items. All of this information provides an opportunity to correct any flaws in the items that are to be included in the field tests.
PHASE 3: FIELD TEST DEVELOPMENT

Step 8: Develop New Items

North Carolina educators are recruited and trained as item writers for state tests. The diversity among the item writers and their knowledge of the current state-adopted content standards are addressed during recruitment. The use of classroom teachers from across the state as item writers and developers ensures that instructional validity is maintained through the input of professional educators with current classroom experience. In cases where item development is contracted to an external vendor, the vendor is encouraged to use North Carolina educators in addition to professional item writers to generate items for a given project.

Step 9: Review Items for Field Test

Another group of teachers is recruited for reviewing the written test items. Each item reviewer receives training in item writing and reviewing test items. Based on the comments from the reviewers, items are revised and/or rewritten, item-objective matches are re-examined and changed where necessary, and introductions and diagrams for passages are refined. Analyses occur to verify there is alignment of the items to the curriculum. Additional items are developed as necessary to ensure sufficiency of the item pool. Test development staff members, as well as curriculum specialists, review each item. Representation for students with special needs is included in the review. This process continues until a specified number of test items are written to each objective, edited, reviewed, edited, and finalized. Test development staff members, with input from the curriculum staff and other content, curriculum, and testing experts, approve each item to be field-tested.

Step 10: Assemble Field Test Forms

Items for each subject/course area are assembled into forms for field-testing. Although these are not the final versions of the tests, the forms are organized according to the specifications for the operational tests (test blueprints). New items or those that have been substantially changed since the item tryouts are analyzed after field testing. The item performance should be markedly better and the item rejection rates much lower for those items that were included in item tryouts as the items are mainly newly written and do not have item statistics. Parallel forms can be assembled which match test specifications and are parallel in terms of content coverage; however, difficulty of the forms cannot be addressed statistically.

Step 11: Review Field Test Forms

Content specialists at the NCDPI Test Development Section and the Technical Outreach for Public Schools (TOPS) review the field test forms to ensure that clarity, correctness, content coverage, and curricular appropriateness are addressed. Additionally, assembled tests forms are sent to an outside content expert who is not employed directly by the testing program. Such experts are typically professors or other staff of the university, college, or community college system.
**Step 12: Administer Field Tests**

For a stand-alone or explicit field test, a representative sample of students is selected to take the field test forms. Schools are selected from across the state's regions and LEAs to represent the state based on gender, ethnic/racial, geographic, and performance characteristics of the student population, including scores on previous versions of the tests and other appropriate characteristics for developing assessments.

The administration of the field test forms must follow the routine that will mimic the statewide operational administration of a test. The test administrator’s guide for the field test administration includes instructions about the types of data to be collected in addition to student responses to the test items during the test administration. Examples of the types of data collected during field testing are item information, student demographic information, students’ anticipated course grades as recorded by teachers, teachers’ judgments of students’ achievement level, field test administration time, and/or accommodations used for students with disabilities or identified as Limited English Proficient.

After the development of initial forms, field test items are embedded into the operational tests. At that point, all students take a small subset of field test items with their operational forms, and will no longer be aware of which items are experimental. Embedded field test items reduce the need for full forms of field test items and ensures students respond to field test items with the same motivation as they would an operational item.

**Step 13: Review Field Test Statistics**

The field test data for all items are analyzed by the NCDPI in conjunction with services contracted at Technical Outreach for Public Schools (TOPS). The classical measurement model and the three-parameter logistic item response theory (IRT) model (including p-value, biserial correlation, foil counts, slope, threshold, asymptote, and Mantel-Haenszel differential item functioning statistics) are used in the analyses. Teacher comments on field test items are also reviewed. Only the items approved by the NCDPI Division of Accountability Services/Test Development Section staff members, with input from staff members from the K-12 Curriculum and Instructional Services Division are sent to the next step.

**Step 14: Conduct Sensitivity/Fairness Reviews**

A separate committee conducts sensitivity/fairness reviews to address potential bias in test items. The NCDPI Division of Accountability Services/Test Development Section “casts a wide net” when statistically identifying potentially biased test items in order to identify more items for review instead of fewer items. Bias Review Committee members are selected for their diversity, their experience with special needs students, or their knowledge of a specific curriculum area. The NCDPI K-12 Curriculum and Instructional Services Division and additional content specialists review items identified by the field test data as functioning differentially for subgroups. Items are retained for test development only if there is agreement among the content specialists and testing specialists that the item appropriately measures knowledge/skills that every student should know based on the state-adopted content standards.
PHASE 4: PILOT/OPERATIONAL TEST DEVELOPMENT

Step 15: Assemble Equivalent and Parallel Forms

The final item pool is based on approval by the (1) NCDPI K-12 Curriculum and Instructional Services Division for curriculum purposes and (2) NCDPI Division of Accountability Services/Test Development Section for psychometrically sound item performance. To develop equivalent forms, the test forms are built to an IRT test characteristic curve. Each test form matches the test specifications. The test development staff members, in collaboration with the NCDPI K-12 Curriculum and Instructional Services Division, reviews the reliability and timing data to determine the appropriate number of test items. Curriculum content specialists also review the forms to determine if the test specifications have been implemented and to ensure that test forms by grade are parallel in terms of curricular coverage.

Step 16: Review Assembled Tests

The assembled tests are carefully reviewed by content experts at the Technical Outreach for Public Schools (TOPS) and the NCDPI Test Development Section. Representation for students with special needs is included. The content team reviews the assembled tests for content validity and addresses the parallel nature of the test forms. Additionally, assembled tests forms are sent to an outside content expert for review.

At the operational stage, the types of edits allowed are quite limited to avoid invalidating the final item calibration. Should the item be determined to be unusable without the changes, it can be returned to the field test stage for revision and re-field testing. The field test items continue to be reviewed separately, since for those items, major revisions are still allowed.

Step 17: Final Review of Tests

Test development staff members, with input from curriculum staff, other content, curriculum, and testing experts and editors, conduct the final content and grammar check for each test form. If at this point a test item needs to be replaced, the test development staff must rebalance the entire form. If a large number of items are replaced after the series of reviews, the form is no longer considered to be the same form that originally went to review. Therefore the “new” form must go back through review.

Step 18: Administer Test as Pilot

A pilot test of the final forms allows any remaining glitches or “bugs” to be caught without negative ramifications for students or schools. This also allows for calibration of item parameters under instructed, motivated conditions. The pilot test mimics an administration of the operational test in every way except that the standards are not yet in place. Test scores are delayed until after the standard setting and final test administration data analyses.

---

5 Pilot tests are conducted only for new tests not for tests considered revised from a previous test.
Step 19: Score Tests

The NCDPI Division of Accountability Services/Testing Section must complete the following in order to provide local education agencies (LEAs) with the ability to scan multiple-choice answer sheets and report student performance at the local level:

1. Answer key text files must be keyed with the goal/objective information and then converted to the format used by the WINSCAN/SCANXX program.
2. A program converts the IRT files containing the item statistics to scale scores and standard errors of measurement. State percentiles must be added to create equating files.
3. The equating files are created so the appropriate conversions occur: (a) raw score to scale score with standard error of measurement and, (b) scale score to percentile.
4. Files that convert scale scores to achievement levels are added.
5. The test configuration file must be completed next. This file describes the layout of the header/answer sheets, Special Code instructions, answer keys, and the linkage test scores for WINSCAN/SCANXX.
6. Using the WINSCAN or the SCANXX program, header and answer sheets are scanned. This consists of selecting the appropriate test configuration file and scanning answer sheets. The program reads the answer key, equating the file and achievement level files. The individual items are compared to the answer keys and the raw score is calculated by summing the number correct. Each test item receives equal weight. Raw scores are then converted to other scores.

The student’s final score is based solely on performance on the operational sections of the test. Any embedded field test item is not included in the calculation of the student’s score.

Step 20: Establish Standards

Industry guidelines require that performance standards, or cut scores be set using data from a pilot test or first year of fully operational. A variety of established and accepted methods for setting standards are available. Test characteristics, such as inclusion of constructed response items, may dictate which methodology is chosen. In the past, North Carolina has used methods such as Contrasting Groups and Bookmark or Item Mapping to determine standards for state tests. Once the performance standards for a test are determined, typically they are not changed unless a new curriculum, revised test, or a new scale is implemented.

PHASE 5: OPERATIONAL TESTING

Step 21: Administer Tests as Fully Operational

The tests are administered statewide following all policies of the State Board of Education, including the North Carolina Testing Code of Ethics. Standardized test administration procedures must be followed to ensure the validity and reliability of test results. Students with disabilities and students identified as Limited English Proficient may use accommodations as identified by
their Individualized Education Programs, Section 504 Plans, and/or Limited English Proficiency (LEP) documentation when taking the tests.

**PHASE 6: REPORTING**

**Step 22: Reporting Test Results**

For tests containing only multiple-choice or other immediately scoreable items, reports are generated at the local level to depict performance for individual students, classrooms, schools, and LEAs. Results are available shortly after the tests are administered. For tests which contain items relying on human scoring, such as constructed response items, results may take longer. These data can be disaggregated by subgroups of gender and race/ethnicity as well as other demographic variables collected during the test administration. Demographic data are reported on variables such as free/reduced lunch status, LEP status, migrant status, Title I status, and disability status. The results are reported in aggregate at the state level usually at the end of June of each year. The NCDPI uses these data for school accountability and to satisfy other federal requirements (e.g., No Child Left Behind Act of 2001).
# TIMELINE FOR TEST DEVELOPMENT

<table>
<thead>
<tr>
<th>Phase</th>
<th>Timeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: Develop Test Specifications (Blueprint)</td>
<td>4 months</td>
</tr>
<tr>
<td>Phase 2: Item Development for Item Tryout</td>
<td>12 months</td>
</tr>
<tr>
<td>Phase 3: Field Test Development and Administration</td>
<td>20 months</td>
</tr>
<tr>
<td>Phase 4: Pilot/Operational Test Development and Administration</td>
<td>4 months for EOC tests (9 months for EOG tests)</td>
</tr>
<tr>
<td>Phase 5: Fully Operational Test Development and Administration</td>
<td>4 months</td>
</tr>
<tr>
<td>Phase 6: Reporting Operational Test Results</td>
<td>Phase 6 completed as data become available.</td>
</tr>
<tr>
<td><strong>Total Time</strong></td>
<td><strong>44-49 months</strong></td>
</tr>
</tbody>
</table>

Note: Some phases require action by some other authority than the NCDPI Testing Section (e.g. contractors, field staff). These phases can extend or shorten the total timeline for test development.
**DEFINITION OF TERMS**

The terms below are defined by their application in this document and their common uses among North Carolina Test Development staff. 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.

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accommodations</td>
<td>Changes made in the format or administration of the test to provide options to test takers who are unable to access the test under standard test conditions. Accommodations do not alter the construct or content of the test.</td>
</tr>
<tr>
<td>Achievement Levels</td>
<td>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.</td>
</tr>
<tr>
<td>Asymptote</td>
<td>An item statistic that describes the proportion of examinees that endorsed a question correctly but did poorly on the overall test. Asymptote for a typical four choice item is 0.20 but can vary somewhat by test. (For math it is generally 0.15 and for social studies it is generally 0.22).</td>
</tr>
<tr>
<td>Biserial correlation</td>
<td>The relationship between an item score (right or wrong) and a total test score.</td>
</tr>
<tr>
<td>Common Curriculum</td>
<td>Objectives that are unchanged between the old and new curricula</td>
</tr>
<tr>
<td>Cut Scores</td>
<td>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.</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>The extent to which a test item measures more than one ability.</td>
</tr>
<tr>
<td>Embedded test model</td>
<td>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.</td>
</tr>
<tr>
<td>Equivalent Forms</td>
<td>Statistically insignificant differences between forms (i.e., the red form is not harder).</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
</tr>
<tr>
<td>Field Test</td>
<td>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.</td>
</tr>
<tr>
<td>Foil counts</td>
<td>Number of examinees that endorse each foil (e.g. number who answer “A,” number who answer “B,” etc.)</td>
</tr>
<tr>
<td>Item response theory</td>
<td>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.</td>
</tr>
<tr>
<td>Item Tryout</td>
<td>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 may be tested.</td>
</tr>
<tr>
<td>Mantel-Haenszel</td>
<td>A statistical procedure that examines the differential item functioning (DIF) or the relationship between a score on an item and the different groups answering the item (e.g. gender, race). This procedure is used to examine individual items for bias.</td>
</tr>
<tr>
<td>Operational Test</td>
<td>Test is administered statewide with uniform procedures and full reporting of scores, and stakes for examinees and schools.</td>
</tr>
<tr>
<td>p-value</td>
<td>Difficulty of an item defined by using the proportion of examinees who answered an item correctly.</td>
</tr>
<tr>
<td>Parallel Forms</td>
<td>Covers the same curricular material as other forms</td>
</tr>
<tr>
<td>Percentile</td>
<td>The score on a test below which a given percentage of scores fall.</td>
</tr>
<tr>
<td>Pilot Test</td>
<td>Test is administered as if it were “the real thing” but has limited associated reporting or stakes for examinees or schools.</td>
</tr>
<tr>
<td>Raw score</td>
<td>The unadjusted score on a test determined by counting the number of correct answers.</td>
</tr>
<tr>
<td>Scale score</td>
<td>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.</td>
</tr>
<tr>
<td><strong>Slope</strong></td>
<td>The ability of a test item to distinguish between examinees of high and low ability.</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Standard error of measurement</strong></td>
<td>The standard deviation of an individual’s observed scores usually estimated from group data.</td>
</tr>
<tr>
<td><strong>Test Blueprint</strong></td>
<td>The testing plan, which includes numbers of items from each objective to appear on test and arrangement of objectives.</td>
</tr>
<tr>
<td><strong>Threshold</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>WINSCAN Program</strong></td>
<td>Proprietary computer program that contains the test answer keys and files necessary to scan and score state multiple-choice tests. Student scores and local reports can be generated immediately using the program.</td>
</tr>
</tbody>
</table>