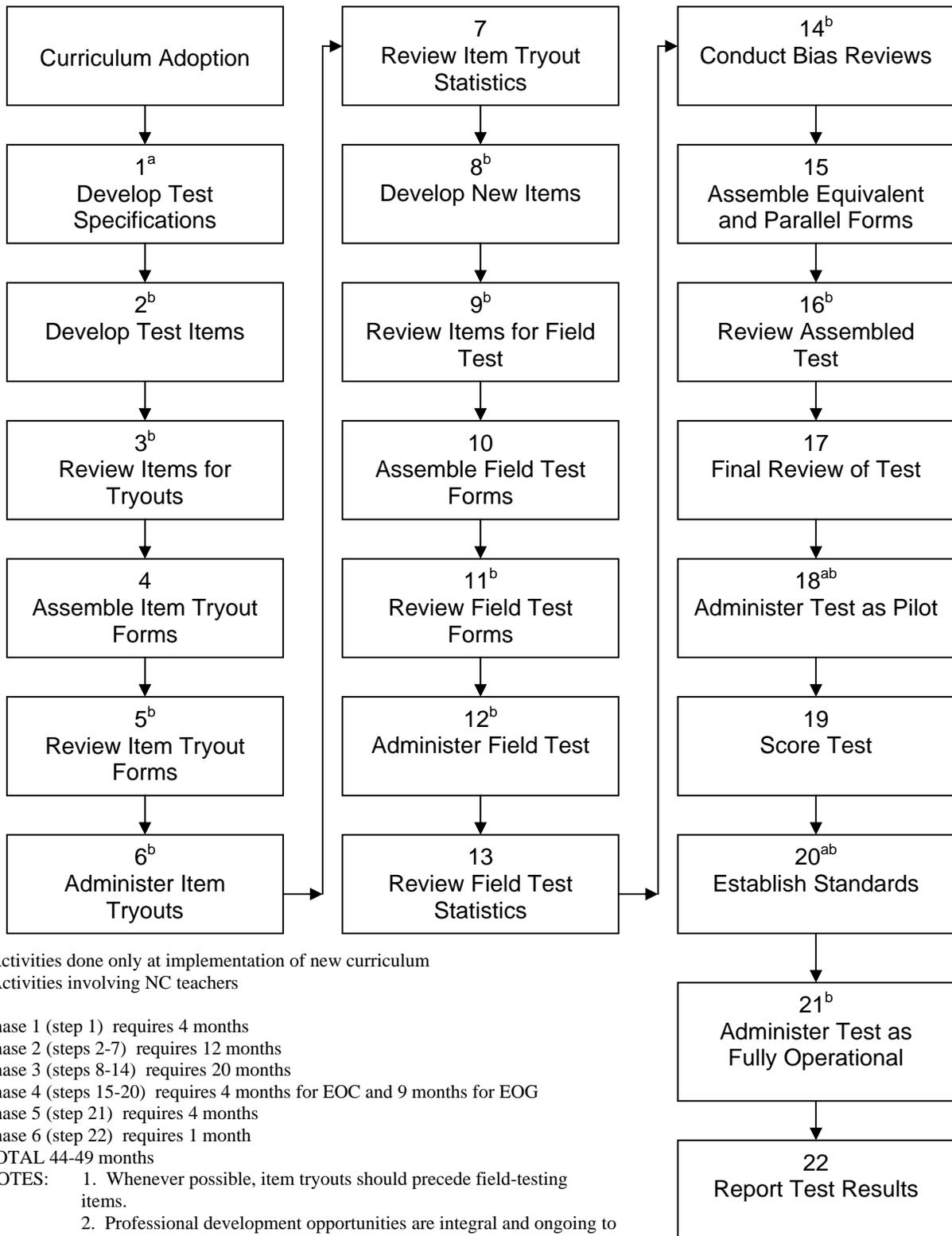


North Carolina Testing Program Multiple-Choice Test Development Process Flow Chart



^aActivities done only at implementation of new curriculum

^bActivities involving NC teachers

Phase 1 (step 1) requires 4 months

Phase 2 (steps 2-7) requires 12 months

Phase 3 (steps 8-14) requires 20 months

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

Phase 5 (step 21) requires 4 months

Phase 6 (step 22) requires 1 month

TOTAL 44-49 months

- NOTES:
1. Whenever possible, 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

MULTIPLE-CHOICE TEST DEVELOPMENT PROCESS

Introduction

North Carolina tests are curriculum-based tests designed to measure the objectives found in the North Carolina *Standard Course of Study (NCSCS)*. The responsibility of updating the *Standard Course of Study* falls to the North Carolina Department of Public Instruction (NCDPI) Division of Instructional Services. Curriculum specialists, teachers, administrators, university professors, and others assist in the process of updating curricula. Once curricula are adopted or tested objectives are approved (e.g. NC High School Comprehensive Test) by the North Carolina State Board of Education, in areas where statewide tests are required, the test development process begins.

The *Standard Course of Study* is reviewed for possible revisions every five years; however, test development is continuous. The NCDPI Accountability Services/Testing Section test development staff members begin developing **operational** test forms for the North Carolina Testing Program when the State Board of Education 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 Board determines that the development of a new test will enhance the education of North Carolina students (e.g. NC Tests of Computer Skills). 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 difficulty² and thinking skill level³

¹Millman, 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.

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 answer correctly. Hard test questions are ones that only about 20 percent or 30 percent of the students would answer correctly. The number of items at each difficulty level for a test is identified during Step 1. For example, for end-of-grade and end-of-

- (2) Percentage of item types such as graphs, charts, diagrams, political cartoons, analogies, and other specialized constraints
- (3) Percentage of test questions that measure a specific goal, objective, domain, or category
- (4) For tests that contain passages, the percentage of types of passages (e.g. literary vs. nonliterary passages, percentage of composition vs. literary analysis, etc.).

PHASE 2: ITEM DEVELOPMENT (ITEM TRYOUTS⁴ 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. Some examples of common curriculum objectives are historical trends in literature and theorems in geometry. 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 will be released for training or for teachers to use in the classroom. 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.

Step 3: Review Items for Tryouts

The review process for items developed from the common curriculum 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.

course mathematics tests, 25 percent of the items were specified to be written at the easy level, 50 percent of the items were specified to be written at the medium level, and 25 percent of the items were specified to be written at the difficult level. This 25/50/25 breakdown of easy, medium, and hard items is a standard industry practice.

³*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 (thinking skill: organizing); another question may ask the student to select the best procedure to use for solving a problem (thinking skill: evaluating). The thinking skills framework adopted by NCDPI in framing the *Standard Course of Study* is adapted from *Dimensions of Thinking* by Robert J. Marzano and others (1988). 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.

⁴NCDPI Testing Section reserves the right to waive the “item tryout” component if time and other resources do not support the practice, if no items are left from the old curriculum to put into item tryouts, or if requirements for field testing are limited.

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 are tested. Conducting item tryouts has several advantages. The most 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 identify the need for a particular type of item (e.g. analogies). Having this data prior to field-testing and operational testing informs the item development and the test development process.

Conducting item tryouts will become increasingly important as the state moves to embedded field tests. Item tryouts provide an opportunity to determine the feasibility of and best possible plan for embedding, which can vary by subject or grade. Experimental items or sections can be tried out to determine whether students perceive them to be radically different from other sections. In addition, item tryouts provide an opportunity to examine the impact of the experimental sections on students' performance.

Step 5: Administer Item Tryouts

When item tryouts are administered as a stand-alone item tryout, 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.

Step 6: Review Item Tryout Forms

Teachers are recruited to 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 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, bias analyses can be run and the 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 NCSCS 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 multiple-choice 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 to verify that there is a valid representation by objectives also occur. Additional items are developed as necessary to ensure sufficiency of the item pool. Test development staff members, with input from 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). If the items on the field test have been through the item tryout process, the field-test forms are parallel and can also be **quasi-equated** because the item-level statistics are already available for those items. 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. If the items have not been through tryouts (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

A new group of teachers is recruited to review the field test forms to ensure that clarity, correctness, potential bias, and curricular appropriateness are addressed. The NCDPI staff members from the Limited English Proficient (LEP) and Exceptional Children's Sections also review each field test form. The NCDPI test development staff, curriculum staff, and other content specialists (e.g. exceptional children, LEP) review teacher comments about the items, and necessary changes are made to items in the test. Teacher responses to the field test items are also used to verify the answer keys.

Step 12: Administer Field Tests

For a stand-alone or explicit field test, a stratified random sample of students is selected to take the field test forms. To ensure broad representation, schools are selected from across the state and are representative of the state based on the ethnic/racial characteristics of the student population, geographic location, and scores on previous versions of the tests among other characteristics. (Note that once field tests become embedded in operational tests, there will no longer be a need for stratified random sampling for field tests. The field test "sample" will

census the entire population of students with the exception of those students who take the alternate assessments. Periodic stand-alone item tryouts may be necessary for new item types.)

The administration of the field test forms must follow the routine that will mimic the statewide administration of a test. The test administrator's manual 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 Teacher Test Item Review Form, 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.

The above process will be modified for embedded field tests. For example, teachers will continue to provide the anticipated course grade and achievement judgments; however, they will no longer be able to complete the Teacher Item Review Form during the test administration since they will no longer be aware of which section is experimental.

Step 13: Review Field Test Statistics

The field test data for all items are analyzed by the NCDPI in conjunction with services contracted at the L. L. Thurstone Psychometric Laboratory, UNC-Chapel Hill and 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 bias statistics**) are used in the analyses. Only the items approved by the NCDPI Division of Accountability Services/Testing Section staff members, with input from staff members from the Division of Instructional Services are sent to the next step. For stand-alone field tests, teacher comments are also reviewed.

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/Testing 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 Division of Instructional Services and additional content specialists review items identified by the field test data as biased. 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 North Carolina *Standard Course of Study*.

PHASE 4: PILOT TEST DEVELOPMENT

Step 15: Assemble Equivalent and Parallel Forms

The final item pool is based on approval by the (1) NCDPI Division of Instructional Services for curriculum purposes and (2) NCDPI Division of Accountability Services/Testing Section for psychometrically sound item performance. To develop **equivalent** forms, the test forms are balanced on P+ (sum of p-values). If the tests have a subsection or exhibit dimensionality, the

subsections are equated. Finally, to the extent possible, the sections are balanced on slope. Each test matches the test specifications. The test development staff members, in collaboration with the NCDPI Division of Instructional Services, reviews the 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

A separate group of educators participates in the review of the assembled tests. Representation for students with special needs is included. The group reviews the assembled tests for content validity, responds to test items for an additional answer key check, and addresses the parallel nature of the test forms.

When embedding is fully implemented, teachers will review only the operational portions. 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 recalibration. The field test or item tryout sections will 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 to a teacher review.

Step 18: Administer Test as Pilot⁵

Because the field test forms are disassembled to form a global item pool from which the final tests are made, a **pilot test** of the final forms will allow any remaining glitches or “bugs” to be caught without negative ramifications for students or schools. The pilot test mimics an administration of the operational test in every way except that the standards are not yet in place. Thus the test can have no stakes for students. If there are stakes for schools they must be delayed until after the standard setting and final test administration data analyses.

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.

⁵ Pilot tests are conducted only for new tests not for tests considered revised from a previous test.

4/20/03

(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, (b) scale score to percentile, and (c) raw score to standard error of measurement.

(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, the student survey questions, 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 multiple-choice test item receives equal weight. Raw scores are then converted to other scores.

As mentioned earlier, when the move to an embedded model is complete for a subject or content area, the student's final score is based solely on performance on the operational sections of the test.

Step 20: Establish Standards

Industry guidelines require that standards be set using data from a pilot test or first year of fully operational. When data are not available from a pilot or first year fully operational test, interim standards are set using model based estimates from field tests. In addition, North Carolina has used the Contrasting Groups Method, a student-based method of standard setting, to determine standards for state tests. This method involves having students categorized into the various achievement levels by expert judges who are knowledgeable of the students' achievement. Teacher judgment of student achievement is compared to actual student performance on the operational tests. Analysis of this data is used in setting performance standards (e.g., achievement levels, cut scores) for the 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 when taking the tests.

PHASE 6: REPORTING

Step 22: Reporting Test Results

For multiple-choice tests, reports are generated at the local level to depict performance for individual students, classrooms, schools, and LEAs. Results are distributed a week or two after the tests are administered. These data can be disaggregated by subgroups of gender and race/ethnicity as well as other demographic variables collected during the test administration. Demographic data are reported on variables such as free/reduced lunch status, LEP status, migrant status, Title I status, disability status, and parents' levels of education. The results are reported in aggregate at the state level usually at the end of June of each year. The NCDPI uses these data for school accountability and to satisfy other federal requirements (e.g. Annual Yearly Progress (AYP) requirement, No Child Left Behind Act of 2001).

4/20/03

TIMELINE FOR TEST DEVELOPMENT

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

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.

Accommodations	Changes made in the format or administration of the test to provide options to test takers who are unable to take the original test under standard test conditions.
Achievement Levels	Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum classified by broad ranges of performance.
Asymptote	An item statistic that describes the proportion of examinees that endorsed a question correctly but did poorly on the overall test. Asymptote for a typical four choice item is 0.20 but can vary somewhat by test. (For math it is generally 0.15 and for social studies it is generally 0.22).
Biserial correlation	The relationship between an item score (right or wrong) and a total test score.
Common Curriculum	Objectives that are unchanged between the old and new curricula
Cut Scores	A specific point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point.
Dimensionality	The extent to which a test item measures more than one ability.
Embedded test model	Using an operational test to field test new items or sections. The new items or sections are "embedded" into the new test and appear to examinees as being indistinguishable from the operational test.
Equivalent Forms	Statistically insignificant differences between forms (i.e., the red form is not harder).
Field Test	A collection of items to approximate how a test form will work. Statistics produced will be used in interpreting item

		behavior/performance and allow for the calibration of item parameters used in equating tests.
Foil counts		Number of examinees that endorse each foil (e.g. number who answer “A”, number who answer “B”, etc.)
Item response theory		A method of test item analysis that takes into account the ability of the examinee, and determines characteristics of the item relative to other items in the test. The NCDPI uses the 3-parameter model, which provides slope, threshold, and asymptote.
Item Tryout		A collection of a limited number of items of a new type, a new format or a new curriculum. Only a few forms are assembled to determine the performance of new items and not all objectives are tested.
Mantel-Haenszel		A statistical procedure that examines the differential item functioning (DIF) or the relationship between a score on an item and the different groups answering the item (e.g. gender, race). This procedure is used to examine individual items for bias.
Operational Test		Test is administered statewide with uniform procedures and full reporting of scores , and stakes for examinees and schools.
p-value		Difficulty of an item defined by using the proportion of examinees who answered an item correctly.
Parallel Forms		Covers the same curricular material as other forms
Percentile		The score on a test below which a given percentage of scores fall.
Pilot Test		Test is administered as if it were “the real thing” but has limited associated reporting or stakes for examinees or schools.
Quasi-equated		Item statistics are available for items that have been through item tryouts (although they could change after revisions); and field test forms are developed using this information to maintain similar difficulty levels to the extent possible.
Raw score		The unadjusted score on a test determined by counting the number of correct answers.

Scale score		A score to which raw scores are converted by numerical transformation. Scale scores allow for comparison of different forms of the test using the same scale.
Slope		The ability of a test item to distinguish between examinees of high and low ability.
Standard error of measurement		The standard deviation of an individual's observed scores usually estimated from group data.
Test Blueprint		The testing plan, which includes numbers of items from each objective to appear on test and arrangement of objectives.
Threshold		The point on the ability scale where the probability of a correct response is fifty percent. Threshold for an item of average difficulty is 0.00.
WINSCAN Program		Proprietary computer program that contains the test answer keys and files necessary to scan and score state multiple-choice tests. Student scores and local reports can be generated immediately using the program.