DOCUMENT RESUME

ED 314 457	TM 014 321
TITLE	Guide to Test Development: A General Procedures Manual for the North Carolina Test Development Program.
INSTITUTION	North Carolina State Dept. of Public Instruction, Raleigh. Div. of Research.
PUB DATE	88
NOTE	234p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	*Achievement Tests; Curriculum Development; Educational Objectives; Elementary Secondary Education; *Item Analysis; Psychometrics; State Programs; *Test Construction; Testing Programs; Test Manuals
IDENTIFIERS	Curriculum Based Assessment; *North Carolina Test Development Program

ABSTRACT

This procedures manual for the North Carolina Test Development Program details the tasks involved in test development. The basic steps are described in an introductory section; they are (1) to specify the institutional objectives of the curriculum and ensure their instructional validity; and (2) to specify the characteristics, such as the purpose and time-constraints, of the test. The second chapter concerns writing the test items and conducting item content validity and content bias analyses to determine which items to keep. The next chapter is devoted to examining item psychometrics, which involves testing items and analyzing and revising them. The fourth chapter, on constructing tests, describes: (1) asembling the tests; (2) constructing the consensual validity and standards analyses; and (3) "testing" the test through pilot administration. Finally, a section on analyzing usage results discusses: (1) administering the test; (2) analyzing the posttest data; and (3) planning for new editions. Eight appendices comprising the greater part of the document include: a detailed list of test development procedures; a Curriculum Survey Form; a Sample Item Specification Form; the script for the audiovisual presentation "How to Write Multiple-Choice Achievement Items," with three work packets, and a sample Item Specification sheet; the text of the booklet "Instructions for Item Review of Achievement Test Items" with a sample item record for item review; a sample item record with all data included; sample test review instructions and questionnaires; and the technical manual for the North Carolina Test of Algebra 1 as a sample. (SLD)

Guide to Test Development

A general procedures manual for
the North Carolina Test Development Program

NC TESTS

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332
(919) 733-3809

Published 1988

Guide to
Test Development

A general procedures manual for the North Carolina Test Development Program

William C. Inman, Editor

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332
(919) 733-3809

NC TESTS

North Carolina Department of Public Instruction

A Craig Phillips, State Superintendent of Public Instruction

Research, Testing, and Accreditation Services

William J. Brown, Jr., Assistant State Superintendent

Division of Research

William C. Inman, Director
Clark Trivett, Assistant Director
Eleanor E. Sanford, Educational Consultant
Shirley Stoll, Research Assistant
Lisa Powell, Research Assistant
Perry Cumbie, Editor
Barbara Corey, Administrative Assistant

Table of contents

Introduction 1
    Specifying the Instructional Objectives 3
    Test Specifications 6

Creating Items 9
    Writing Items 9
    Conducting Item Content Validity and Content Bias Analyses 14

Examining Item Psychometrics 16
    Testing Items 16
    Analyzing Items and Making Revisions 23

Constructing Tests 26
    Assembling the Tests 26
    Constructing Consensual Validity and Standards Analyses 29
    Testing the Test: Pilot Test Administration 31

Analyzing Usage Results—and Beyond 32
    Administering the Test 32
    Analyzing Posttest Data 32
    Planning for the Next Editions 33

Appendices 35
    A Test Development Procedures
    B Curriculum Survey Form
    C Sample Item Specification Form
    D Script for How to Write Multiple-Choice Achievement Test Items, three Work Packets, and a sample Item Specification
    E Booklet, Instructions for Item Review of Achievement Test Items and Sample Item Record for Item Review
    F Sample Item Record with all Data Included
    G Test Review Instructions and Questionnaires
    H Technical Manual

Introduction

In developing a test, the most important task is that of defining the subject to be tested, of laying out in detail the anatomy of the topic. We start with concrete facts, but the final product is an abstraction, a rationally constructed analog of the world, not as it is, but as we choose to define it for our special purposes. The first part of this discussion will be directed toward defining the substructure that supports this ultimate abstraction.

First, all achievement tests deal with knowledge, broadly defined to include skills. Knowledge exists in two forms: *physical records* consisting of artifacts housed in libraries and museums, legal and other business documents preserved in files, information coded in electronic form and stored in computers and on tapes and disks, knowledge 'models' such as buildings and automobiles, archeological information buried beneath the ground, and other tangible artifacts, such as a stage performance, which is a physical record, albeit transitory in nature; and *mental records*, consisting of a living human being's invisible repertoire of knowledge. Tests deal with mental records.

Mental records can be regarded as physical records placed in working storage in an individual where they can be used for the benefit of the individual and of society in general. Tests attempt to describe in some manner an individual's current inventory of mental records. We take inventory by asking an individual to do something observable—select a correct answer among several possible answers to a problem, demonstrate the skill of riding a bicycle, sing a song, or exhibit other behavior that will give us a clue as to the individual's current state of knowledge in the area that interests us. In short, we ask for a physical record that corresponds in some sense to the mental record.

Mental records are limited in several ways: in extent, by what people can learn; by the amount of time and money that can be devoted to learning; by the utility of what is learned; and by the length of the human life cycle. Because of these limitations, care must be taken not to be extravagant in the selection of the physical or mental knowledge to be converted to new mental storage. Also, a distinction must be made between what is of general use to all members of a generation, and what should be targeted to only a small group of specialists.

When we plan to transfer a portion of society's existing physical and mental records to an as yet untutored individual, we create a curriculum (a course; a route or path). A curriculum can refer to a single course of study, or a collection of courses.

Role of Scholars

The role of scholars is central to the process of curriculum development. It is the place of scholars to inventory knowledge and describe its extent and significance. Scholars are also responsible for devising courses of study. To give substance to their judgments of worthy courses, scholars write textbooks and reference books, create films and tapes, and publish other information.

The functions satisfied by scholarship never have a mark of finality. New knowledge originated by scientists, inventors, and creative writers makes old syntheses obsolete. Conditions in the world change because of war, pestilence, famine, and other catastrophes, and bring new values to knowledge. A scholar tries to anticipate needs and directions for the use of knowledge, but sometimes reaches too far or not far enough, or simply misunderstands the material. Therefore, courses of study are constantly under revision.

Role of Curriculum Specialists

In total, the scope of curricular scholarship far exceeds the practical scope of a curriculum for elementary and secondary education. Much picking and choosing must occur before a curriculum is finally endorsed for use in public schools. Curriculum specialists play an important role in this process of selection.

Curriculum specialists, however, can only propose a curriculum. Here we need to distinguish between a *proposed* curriculum and an *established* curriculum. A proposed curriculum may or may not be in place in schools. For example, a national committee of curriculum specialists may decide that U.S. History should cover 114 curricular objectives, and a test specialist may develop a test to assess the knowledge of students with respect to those objectives. When the results of the test administrations are known, it may be found that students have little knowledge of the history covered by the 114 objectives. This does not speak to whether the material was not taught, was taught but not learned, or even whether it should have been taught in public schools.

It is this last concern—that of what *should* be taught—that leads us to the *established* curriculum. The established curriculum is a curriculum that has been legitimated by a diffuse process of decision-making. The general subjects to be offered are defined at the state level where the basic responsibility for public education rests; additional subjects may be added at the district level; some of those subjects may be taught in some schools and not others; and some subjects may be accepted or rejected by the students themselves or their parents, or accepted for different years. The courses of study defined at the state level include subject goals and objectives. A state-appointed committee decides what textbooks the state will make available; the local school districts may add books and further define the goals and objectives; and teachers may follow curricular objectives rigidly or use them in only the most general way.

The Common Core Curriculum

Thus it is difficult to define curriculum exactly for even a single classroom, let alone a state. Yet knowledge clusters relating to grade levels and subjects have sufficient unity to permit the establishment of a common core curriculum. The first task of test development is to identify the portion of this curriculum that is to be covered by the test.

To many people first participating in the process of test development, this point may seem a good time to revise the curriculum or even to carry out a curricular revolution. Closer inspection of the circumstances shows this not to be the case.

First, it is not ethical (or even legal) to test a student on material the student has not been taught, and then make the results a part of the student's record where the results may be interpreted as poor scholarship. Usually, a year's instruction with the new curriculum would be needed to avoid this element of unfairness, assuming that no untaught courses are prerequisite to the revised subject. Second, the lead time needed to install a new curriculum would likely require another year for teacher training. Also, if the changes were substantial, additional texts would have to be obtained and new ones would perhaps need to be written and published. Furthermore, professional and public debate over the changes would need to be resolved. Therefore, any substantive change in curriculum could require several years lead time. That usually is inconsistent with the test development timetable.

Another problem with tests covering proposed rather than established curricula is the matter of reliability and validity. In the development of experimental tests intended to measure new variables, validity is not assumed and a low degree of reliability can be expected. Tests that are expected to have a high degree of validity and reliability, however, must deal with well-defined domains of behaviors. This suggests that the domains have already been exposed to substantial amounts of scholarship and have had ample periods of institutionalization. That will generally be true of academic subjects, although one can imagine exceptions. Such exceptions, however, are likely to be experimental in nature.

From these considerations, it should be clear that test development should recognize institutionalized change; but test development should not be required to lead it.

Specifying the Instructional Objectives

If we recognize that test development must relate to the established curriculum, how does a test developer go about defining the curriculum for purposes of test development? First, one must decide what the answer to the question must look like. In general, what is needed is a description that will be specific enough to permit item writers to write questions, or behavior samples, to represent all of the questions that could be asked about that topic.

By far the most useful general foundation for test development is the concept of an objective-referenced curriculum. In this frame of reference, an instructional objective becomes the smallest organically-intact unit of instruction. Sub-objectives represent fragments of an objective, but are incomplete concepts when standing alone. Goals are aggregations of logically-related objectives, but lack the cohesiveness and internal consistency of objectives.

An instructional objective is frequently the topic of a unit of study. In classroom testing, the teacher may administer an end-of-unit test of as many as 25 to 30 items. If the test is well-prepared, it is likely to have a useful degree of reliability and validity. Most standardized testing programs cannot yet afford the luxury of administering survey tests covering, say, 100 objectives with 25 questions each, for a total of 2500 test items. Even the classroom teacher may not ask that many questions over the course of a year. So the test developer is constantly in the position of crafting a test that will yield the most information for the defined testing period. Many compromises are necessary, but they need not vitiate the primary goals of test development.

By viewing objectives as the basic building blocks of the curriculum, we can lay a firm foundation for test construction. The basic test development task is to create items to measure the acquisition of knowledge concerning instructional objectives. Each objective is basic; each objective should be measured. Thus each objective becomes the focus of what amounts to a separate test. The test may be short—even as short as a single test item. This does not release the test developer, however, from having to answer all of the usual questions about the test—however short: reliability, validity, population difficulty (percentile), its relationship to other tests. That this is always neglected to some degree does not change the nature of the responsibility.

Defining the Objectives

It is essential that the instructional objectives be written in such a manner that item-writers can use them as guides in creating test items. In recent years, many prescriptions have been offered as guides to item-writing. Most have helped to explicate the issues, but no one procedure has gained general acceptance. In most subjects, the cues necessary for item writing need be only minimal, since the subjects most likely to be tested are ones that have been widely taught for many years. Expositions will be widely available in textbooks and other teaching materials. Item writers will have become thoroughly acquainted with the the subjects through teaching experience. Teaching experience also will be a good resource in the essential task of generating logical but incorrect multiple choices for items (if the test takes that form). What is needed most is rules to keep item writers from duplicating their efforts and to prevent an item writer from becoming stale through an exhaustive effort to produce more and more items to measure a single objective.

Curriculum descriptions written at the goals and objectives level of discourse will be satisfactory for most subjects (e.g., "add one-digit numbers"). Some detail is required if the level of exposition of the objectives is so abstract that item writers have insufficient cues for their work (e.g., "assess strengths and weaknesses in a practical way"). Once test developers have had experience with insufficient cues, their tendency is to over-elaborate the item specifications. But after providing detailed examples that are cloned ad infinitum by item writers, their tendency is to be general. The appropriate level is the level that is appropriate for a particular topic. Each case tends to require a somewhat different solution.

Instructional validity

The definition of goals and objectives proceeds from several directions. In many states, curriculum specialists have defined, or are in the process of defining, goals and objectives for mandated courses. These are usually presented in some formal way: a report; a handbook. The level of specificity of the objectives differs by subject area. Also, the degree to which the objectives represent the institutionalized curriculum may vary. Local validation of these curriculum specifications is required. This should be done through a survey of local school district curriculum administrators and a representative sample of teachers, who need to examine the goals and objectives and identify the ones now taught in their schools.

The goals of reliability and validity always remain paramount. For validity, we can go directly to the point in asking a teacher if a certain objective was taught in his or her classroom this year. The teacher is highly likely to know and, if approached correctly, will have little reason to misrepresent the fact. Reliability questions enter the picture only if an inadequate sample of teachers is queried. Some professional judgment is needed in making a sampling decision, since a reliable sample depends upon the frequency of 'yes' responses as well as other factors, including a lack of ambiguity in the questions.

The questionnaire, if left open-ended, would be a very inefficient tool for gathering reliability and validity information. An extensive list of objectives is desirable, so that teacher responses can be reduced to category judgments ("yes," this is taught; "yes," it is essential). This list can start with the state's list of objectives, assumings its technical adequacy has been endorsed by curriculum specialists. Teachers can be invited to add important objectives covered in their subject-area classes. The exact routine for the survey is given as part of the overall test development procedures detailed in Appendix A. An example of a curriculum survey form is given in Appendix B.

After all data have been analyzed, discrepancies can be eliminated through expert analysis by curriculum specialists and others responsible at a state level for the content of classroom instruction. No required level for teacher endorsement of an objective as 'basic' can be given. If the objectives are written in general terms, endorsement percentages in the 90's may be appropriate; if the objectives are highly specific, it could go either way: comparatively high, as in mathematics, or comparatively low, as in some social studies areas. In practice, however, it usually is easy to separate by inspection those commonly-taught objectives from those rarely-taught objectives. One should not, finally, try to abstract general principles of endorsement where no principles exist. 'One size' does not 'fit all.' Too much depends upon the homogeneity of the curriculum (algebra is very homogeneous; U.S. History is not) and the level of generality with which the objectives are stated.

Finally, the objectives may be clustered under goals; they may be ordered along some continuum such as difficulty; or other meaningful relationships for test development conceptualization may be discovered. (Some Social Studies objectives are History objectives, some are Sociology objectives, etc.) It is at this point, however, that a final conceptualization of the total body of objectives should be made. Once item writing has begun, the objectives should not be subject to further revision.

The objectives so determined constitute a domain of objectives or expected behaviors. They are, within practical limits, an exhaustive list of all objectives described in a specific manner: deemed to be essential to a basic definition of the subject area; taught by *x*% or more of all teachers; regarded as important to a determined degree; etc.

Test Specifications

The second task, subordinate in importance only to the definition of the domain of objectives, is to specify the characteristics the test is to have.

The first consideration in test specifications is that of what purpose the test is to serve. The test may be intended to provide information about the extent to which the test taker has mastered the subject area. That information may be expressed in two forms: as a proportion of all that should be learned, or as a score ranking among the scores of all students who would be eligible to take the test because of their age or level of advancement in the curriculum. The former type of question is associated with the term "criterion-referenced test," the latter with the term "norm-referenced test." Both of these conceptions are summative in intent: they are designed to describe how well the student has done in his or her studies. The idea that tests can be divided into classes—either criterion-referenced or norm-referenced—is more suited for professorial exposition than for practical test development. The test developer is better served by sticking to the concept of objective-referenced test development. The same test can always be treated as either criterion or norm referenced, and indeed the information supporting both views should be available as a result of thorough test development.

A second test purpose may be that of diagnosis. Neither the criterion-referenced nor the norm-referenced test is inherently diagnostic. A criterion, for example, may be that a student must get 75% of the questions correct on a mathematics test, an outcome that is hardly diagnostic in the traditional sense. The distinguishing feature of a diagnostic test is the extent to which the scoring process can give reliable information about the performance of an individual with respect to each objective. The test developer concentrating on an objective-referenced framework must continually be concerned with the diagnostic value of each objective-referenced test, whether the test consists of one item or more.

A third test purpose may be that of curriculum assessment. What is the adequacy of the course of study offered to students? Are students learning the subject matter to a sufficient degree? Are there differences in performance among localities? Are portions of the curriculum being neglected in some classes, schools, local districts, or in the state as a whole? The attempt to provide valid answers to these questions has strong implications for test specifications.

A fourth test purpose may deal with what are popularly called cutoff scores. A test designed solely to provide measurement at a single cutoff score is constructed quite differently from other tests.

Once the purposes of the test have been determined, questions of practicality and efficiency arise. Should all objectives be tested? What mode of questioning will be employed? Do practical considerations dictate that some objectives not be covered, or that the coverage of objectives need not be extensive enough to provide information on each objective? Can the coverage of objectives be obtained in one year, or could it be spread over several years? What are the minimum requirements with respect to these issues? Can the conflicts be compromised satisfactorily, or should the testing program be abandoned as impractical?

If the answer is that the testing program should proceed, then specific answers need to be given regarding acceptable levels of reliability, validity, and generality. The answer to those questions will determine how many items are to go into the test, how they are to be distributed with respect to difficulty, when they are to be administered, and who is to take the test for what purposes. (For example, the characteristics of the items themselves must usually be determined before the test is assembled from the items. This requires that the items be assembled into test forms and given tryouts among representative groups of students. How representative the students must be has generated several answers, depending upon whether traditional or latent trait theory is used in test development. In practice, however, the answer seems to converge on established rules: the only safe sample is a representative sample.)

Once the test specifications have been defined, the costs must be estimated. If the costs exceed the money budgeted for the project, more money must be requested of the funding source or sources, the scope of the project must be reduced and the test specifications reconsidered, other kinds of resources must be substituted for money, the project must be spread over a longer period of time, or some combination of these actions must be decided upon.

Developing the Test

When test specifications and budgets are complete, it is time to start staffing and planning the day-to-day details of the test development program. An overview contains elements for

- Creating Items
  - Writing Items
  - Conducting Item Content Validity and Content Bias Analyses
- Examining Item Psychometrics
  - Testing Items
  - Analyzing Items and Making Revisions
- Constructing Tests
  - Assembling the Tests
  - Conducting Consensual Validity and Standards Analyses
  - Testing the Test: Pilot Test Administration
- Analyzing Usage Results—and Beyond
  - Administering the Test
  - Analyzing Posttest Data
  - Planning for the Next Edition

The first topic to be discussed in the following sections is that of creating the items.

Creating Items

Writing Items

As a result of previous considerations, it is now clear just exactly what objectives will be tested and how many items of particular kinds must be developed to measure each objective. Using this information, it is possible to plan and carry out an item-development program having the following elements:

1. Survey existing tests and item pools as sources

Many sources for items exist, some of which will supply items at no cost, some for a one-time cost, and some for a cost per use. Frequently, it is more cost effective to develop new items. The statistics associated with the pre-developed items may be of little use in practice and the questions may seem just to miss the mark set up by your objectives. In addition, the possibility of a lapse in security must be considered: if the items are available to you, they may be available in the areas where you had hoped to have secure tests.

In addition, an existing item has the status for your test development program that a newly written item has: it must go through the same field-testing procedures. Furthermore, it must be matched to your objectives, the total objective coverage must be inspected for some hidden bias, the item must be edited to your format, and the form and content closely inspected to make sure it does not contain some unwanted alien feature or odd terminology. Conforming the existing item to your structure may take longer than writing a new, entirely appropriate item.

2. Write item specifications

If sources for items already exist, the items will likely be associated with objectives, and, to obtain the items, only an objectives match must be made. If the items must be written, it becomes highly desirable to have some specifications for what they must be like.

Unfortunately, it is easier to conceptualize item specifications than to write specifications that work. Popham's discussion (in Berk, R. A., Ed., *A Guide to Criterion-Referenced Test Construction*, Johns Hopkins, 1984) of this is enlightening. After several false starts, he came up with a procedure that has four components: a general description (basically, the objective); a sample item; stimulus attributes; and response attributes. Stimulus attributes consist primarily in indicating what material is relevant and where it may be obtained. For multiple choice items, which constitute the bulk of all item responses developed in achievement testing, the response attributes could very well be covered in general training of the item writers, with only an occasional nod to some particular needs for that objective.

One does not come away from the topic with any confidence that an easy general solution exists. Experience underscores that feeling. Rather than spend large sums of money in writing elaborate item specifications that will never be read or understood by item writers, it seems the more efficient method is to be satisfied with a modest set of item specifications and an item attrition rate of from 10% to 50%. Under these circumstances, a second wave of item writing may be needed, particularly if the attrition is not spread evenly among objectives.

A point frequently overlooked in item specifications is an instruction concerning the location of correct answers. Under each objective, it is traditional to use each of the (usually four) choices equally as the location of the correct answer. This will reduce the need to change the locations of correct answers in the final test, and will save a substantial amount of editing, reworking, and proofing of answer keys.

An example of an item specification form is given in Appendix C.

3. Select item writers

The characteristics of a good item writer are those you would expect to find in any genius: high intelligence; excellent education in the area of interest; creativity of a high order; and the ability to conceptualize broadly. And, of course, the writer must have an excellent command of language with a special empathy for the level of discourse suitable for the persons taking the test. Finally, the writer should have had teaching experience with the type of persons taking the test. The wages are modest: by the hour or by the accepted item.

Most items writers for elementary and secondary education subjects come from the ranks of teachers or administrators who formerly were teachers. Some become excellent item writers, some do not.

No one has discovered a failsafe method for selecting and training item writers. Work samples before selection, however, may be a good indication of potential success. Payment per accepted item limits the financial risk associated with bad choices and discourages the poor writer's continued participation. A combination per hour/per item payment could be considered.

4. Train writers

A test developer must enter into the training of item-writers with practical expectations. First, it is desirable to have broad representation among item-writers. This lays a foundation for instructional validity—the item-writers will not represent only a small, perhaps biased section of the teaching profession. At the same time, the test developer must realize that few of the item-writers will be excellent writers, and that the item-writing training is unlikely to do more than take some of the rough edges off whatever the item-writer has accomplished over the years in achieving polished literacy. Immaculate prose, fortunately, is not the main thing one hopes to get from item-writers. What the test developer is buying is ideas for questions, furnished with a down-home knowledge of what makes up workable distracters or item foils. If the heart of the item is there, editors can shape it to fit the format.

Although what the test developer is buying from item-developers is primarily ideas, this does not mean that the work is not facilitated greatly by giving the item-writer some formal training in item-writing. This can be done through a home-study course that has been developed for just that purpose. The heart of the course is a videocassette of one hour total duration—not counting three interruptions for work with supplementary materials. This videocassette is accompanied by a written script of the videocassette, which can be used as stand-alone instruction or as a reference for occasional consultation. Three Work Packets accompany the materials.

The item-writing program proceeds as follows:

a. Potential item-writers are nominated by local curriculum consultants from among exemplary teachers who have some writing ability. The teachers are then contracted to write up to a given number of items.

b. The item-writers are sent a package of training materials consisting of:

- the videocassette, *How to Write Multiple-Choice Achievement Test Items*
- the script of the videocassette
- the three work packets
- a copy of McGraw-Hill's *Guidelines for Bias-Free Publishing*
- ten item specifications
- names and telephone numbers of resource persons who will answer curricular or psychometric questions

The item-writers are asked to study the material, write the ten requested items, and return the drafted items to the test developer. These items will then be edited and any necessary suggestions made. The edited items will then be returned for final revision and, barring some intractable problem with the item-writer, the remaining item specifications forwarded. From that point on, the item-writer is working independently, except perhaps for telephone calls to resource persons.

This system replaces a system whereby item-writers were called in to central locations for training. Two days of training was quite inadequate, in some cases leaving the item-writer feeling less competent at the end than at the beginning. This more leisurely home-study course has produced good results and enables the test developer to employ item-writers at any location at any time at far less cost than before.

A copy of the materials used, except for the videocassette and the McGraw-Hill publication, is attached to this report as Appendix D. Copies of the videocassette can be had at cost from Media Consultants Inc., 125 West Chatham Street, Cary, NC 27511, telephone 919/467-3588. You can make your own copies, but the quality deteriorates with each step down from the master tape.

5. Produce items

The planned production schedule should be monitored closely to make sure that deadlines are met. Where necessary, corrective feedback should be given to writers as their material becomes available for inspection. Individual objectives should be monitored to make sure that the quality across the objective domain is uniform. It may be necessary to contract for additional item-writing if it becomes clear that one or more areas are not being covered adequately. At this stage, however, it is very difficult, because of organizational considerations, to terminate existing contracts. The test developer does not need to create a group of disaffected item-writers who will later on be test users.

6. Assemble item pool

At the end of the production period, the item pool should be assembled, checked for comprehensiveness, and given a brief inspection for overall quality. It should then be transferred to electronic storage for further processing. (Later on, it may be possible with scanner/digitizers to move the item-writer's work directly into electronic storage without the necessity of typing it—assuming the item-writer can produce typewritten rather than handwritten copy.) This electronic file, in addition to a draft of the final form of the item, should include the essential information about the item with respect to grade level, subject, objective, key, and any other information that will be of future use (such as the name of the item writer, the date of creation).

7. Edit items (In house)

From the electronic files, an edit copy of each item should be created. This should consist of single sheets of paper, one item per page, with all supporting information regarding the item. After editing and review, this hard copy file will be used to update the electronic file. (The necessity of monitoring the editors' work—which sometimes proceeds with a scapel, sometimes with a meat-ax, depending upon the quality of the text—makes it undesirable to edit directly onto the electronic file. Dual before-and-after electronic files could be created, of course, but the location of a change becomes problematic.)

One of the most critical and demanding tasks is the editing of the item pool. First, the reading level of the material must be checked if there is any suspicion that its difficulty is greater than the reading level of the test takers. Convenient microcomputer programs exist that perform multiple tests at one time on typed material. Words can be checked against word lists (although some programs perform that function automatically).

Second, the items should be read for spelling, syntax, and grammar. A surprising number of inconsistencies will manage to survive this stage unless adequate time and resources are applied.

Third, the items should be checked for biases that, while they might not influence the ability of a student to answer an item, might be offensive to some persons. These usually fall under the heading of stereotyping and fair representation. The examination of items for biases at this stage of development, however, is only for identifying the more obvious transgressions. A full-scale analysis will be conducted at a later stage.

Fourth, the items should be checked for foils that are obviously weak. This must be done, of course, by someone with a background in the subject matter and a knowledge of the characteristics of the persons who will be taking the test.

Fifth, a check should be made to see that the answer key and the item-objective match are correct.

The importance of this step of the item development cannot be overstressed. More substantive improvement in the quality of items can be made at this time than at any later date (except perhaps for the following review by curriculum specialists, which is in fact a continuation of the item edit).

8. Review Items (Curriculum Specialists)

After the item edit is complete, the items should be reviewed by curriculum specialists, who will apply their particular expertise to an examination of the adequacy of the items individually and as a total package. They should also verify the answer key and the item-objective match, since their word should be final on those subjects.

Sometimes the item edit and item review are conducted concurrently from duplicate copies of the items. This frequently leads to problems that can be avoided by having all changes made in sequence on a single hard copy of the items.

9. Review Items (College professors)

It frequently helps to have college professors take a look at the items. If an English professor who is a good editor can be contracted, he or she can check the items for grammar. Frequently, this consultant can recommend another professor who is a specialist in the subject area of interest who can be persuaded for a reasonable fee to check the curriculum validity of the items. This is a powerful editing tool that can be employed in a number of situations.

10. Edit (final)

After the editing and review changes have been made, the hard copy file should be returned to production, where the electronic file should be updated for the changes made. When this is complete, a second hard copy file should be produced. This file will become the basis for the next step. Its contents should be carefully proofed against the editing and review changes made on the first hard copy file.

Conducting Item Content Validity and Bias Analyses

Summary validation of item content and a check for linguistic or pictorial bias (as contrasted with statistical bias) should be timed to begin as soon as the second hard copy file of the items is available. The purpose of this analysis is to expose the completed work to a broad cross section of exemplary teachers within the environment in which the test will ultimately be used, and to make further editorial changes in response to well-reasoned comments. The success of this effort will be judged by the general acceptance the test items have when they reach their intended audiences. What is sought is professional agreement that the items appear to measure what they are supposed to measure and that there is no important linguistic bias present.

This analysis proceeds as follows:

1. Plan procedures

Since statewide consensus is sought, considerable planning must be carried out to see that the teachers who review the test items are representative. Since the amount of data to be handled is large, plans must be made to acquire it in a systematic manner. Otherwise, much time and data may be lost through failure to ask the right questions or obtain answers in a usable form.

The problem can be envisioned by considering that a pool of 1200 test items is not unusual. For ten individuals to examine each test item for multiple characteristics, it would take 80 people to perform the task, each person examining 150 items. That many items is a large order for an individual, who must spend a considerable amount of time getting ready for the task. Since these individuals will be spread among many locations across the state, management, logistics and expense present formidable problems. This problem can be solved satisfactorily, however, by applying the solution applied to item-writing: home study and work.

2. Select item reviewers

The selection of item reviewers follows generally the same procedure as that for the selection of item-writers: nomination by knowledgeable field consultants. The item reviewers—as many as 100—are contracted for the work following their acceptance of the task. They should be representative of all teachers who are expected to administer the final form of the test.

3. Train item reviewers

The item reviewer is supplied with the script for item-writing that was given to the item-writers, the McGraw-Hill pamphlet on bias, a booklet entitled *Instructions for Item Review of Achievement Test Items*, and item records for each item to be reviewed. The item record presents the item, lists its characteristics as known at this time, and displays a decision matrix for checking by the item reviewer. The characteristics to be judged by the reviewer were compiled from the literature and from local experience. Here as elsewhere, test item security is stressed. Reviewers are made aware that insecure items are of no use to the contractor. No copies are to be made, no item content discussions are to be held with others, and no notes are to be taken and kept.

4. Produce the item records

Updated items are available from the final edit described earlier. From eight to ten copies of each item should be reproduced on the overprinted form to create the proper item record for review. (If 1,400 items are reviewed, 14,000 item records must be produced, which points up the desirability, if not necessity, of computerizing the data analysis.)

5. Handle the data

When the completed item records are returned, the information is recorded in electronic form, checks being recorded in a fixed block record, comments copied verbatim and coded for aggregation by item number.

The summarized data very quickly bring the information into a decision mode, where it is examined from psychometric and curricular points of view and decisions made about what items to leave intact, what items to edit further, and what items to discard or re-write completely.

The procedures booklet and a sample item record are given in Appendix E.

6. Further item development or redevelopment

Because item tryouts have not yet occurred, it still is possible to redevelop items or add items to the pool as needed. Very frequently, faulty items can be repaired. When an item is changed by anyone, however, it must go through the item edit and item review process. It is not unusual for a revised item to come out of revision in worse shape than it entered.

7. Revision of electronic files

Approved changes in items or item data (such as the changing of an objective for an item) should have been noted on a single set of item review records. Changes should be prominently marked to avoid confusion or oversights. They should then be given to the item production group for revision of the electronic file. A new hard copy file should then be printed and carefully proofed against the changes.

Examining Item Psychometrics

Testing Items

The best technical efforts have now gone into the construction of the item pool. Yet so far all is theory. It still remains to test the items with student. The plan must deal with a number of issues:

- How many students must take each item? (The answer: from 600 to 1,000, depending upon the technique of analysis adopted and the way the sample is selected.)

- Who should these students be? (The answer: a random sample of eligible students; or a larger random sample of classes; with perhaps oversampling of minority groups in order to satisfy the statistical requirement for item bias analysis.)

- How should the items be presented? (The answer: in much the same way they will appear in the final test—in test booklets, in the same number as planned for the final test; in the same general order; with the same distribution of answer choices; with the same general instructions and the same time parameters. In short, they should appear to the test taker to be conventional tests of the type that will ultimately be offered.)

The first step, then, is:

1. Develop a plan

The plan must include a determination of the number of booklets needed to carry the items. If the number of items per booklet (and in the final test) are to exceed 40 to 60, the test may need to be given in sections.

Also, some method must be devised to place the items from various booklets on the same measurement scale. Usually, this involves the necessity of the booklets having some items in common. These items could be chosen from the item pool, but, if the choices were not very lucky, some or all of the common items might fail to have measurement characteristics that would be adequate for the purpose at hand. Therefore, it is not unusual to select items from some other source for this purpose, assuming that the psychometric characteristics of those items are already known. This also opens the possibility of tying the current item pool into some other, perhaps better developed, item pool. Sometimes, items are purchased from major testing companies in order to reference the test to national norms. The use of common items enlarges the number of items to be administered by at least 10 times the number of booklets needed to carry the items.

For administration, a whole administrative apparatus must be brought into motion to deliver, administer, and retrieve the test booklets.

2. Select items for field test booklets

The selection of items for field test booklets is not as technologically complex as selecting items for the final test. Nevertheless, care must be taken to see that items are systematically distributed by objectives per booklet, that the multiple choice selections do not pile up in one spot (five A's in succession), and that the booklet does not start with a series of very difficult items.

Others things even, it is desirable to have the test start with the ten common items. In any event, it is expedient to have them appear in the same place in all booklets.

In summary, to select items for the item tryouts, it is necessary only to rotate randomly through all of the items by objectives until the items are all gone. Hopefully, one has been able to look far enough ahead to choose a booklet size that will not leave a booklet only half full of items.

3. Revise location of foils (choice responses) as needed

If correct answers were distributed proportionately among item choices when the items were written, little work should be required to prevent distracting runs of correct answers at one particular foil. A rule of thumb is to prevent a run of more than three identical choices in succession. If possible, move the item rather than the foil.

4. Prepare item examples, booklet instructions, and answer sheets

A certain amount of administrative work and final editing is required for each booklet. Some considerations are:

- Present an example for each significantly different item form in the test.
- Indicate START and STOP locations, and identify sections.
- Give the booklet a test name and, if needed, a form number.
- Make provisions on the booklet for student name and any other information needed to locate the test taker. Include items to elicit gender, ethnic origin, and any other factor you may wish to include in an analysis of statistical bias. Attempting to obtain this information from some auxiliary source (e.g., another set of test data) can be very time-consuming and ineffective.
- If necessary, prepare a student background data sheet to be incorporated in the test booklet. This may include teacher information concerning, for example, the length of time students needed to complete the test, or the course grades the students have earned so far.

- Prepare answer sheets or arrange copy for in-booklet answer coding (required at the primary grades).

5. Provide final copy for booklets

When the rough copy has been completed and approved, it should be turned over for production of a final copy suitable for reproduction. The final copy should be carefully proofed with the rough copy.

New material will require the most careful proofing. Most new material will be instructions and data sections or changes in foil locations to assure equal representation among foils and lack of runs of the same response choice. Otherwise, the items will appear as encoded in the electronic file. Additional checking is needed, however, when art work must be added from a different source.

6. Draft administrative manuals

The field administrations of the tests are controlled by administrative manuals. The manuals contain the authoritative routine for administering the tests, from distribution to collection of answer sheets and booklets. Administrative manuals for tests are widely available as models, but each testing situation differs in detail and requires separate explanations.

Once drafted, the administrative manuals should be approved by the various interests in the test development process before they are given to production for final copy.

7. Produce xerox copies of tests and manuals and test the feasibility of administrative procedures

Tentative decisions have been made about a desirable test length and the time required to administer the test. Rough guesses have been made about the general difficulty level of the items. These assumptions should be given a quick check by producing a few copies of the test and administering them to a small group of students typical of the ones for whom the test is intended. The number of students to test in this procedure should not be less than 10. One class with a heterogeneous ability mix will suffice for the purpose. Usually, the assumptions will be found to be tenable. Occasionally, a major rethinking of the item tryouts will be required.

8. Select field sample

The goal of selecting a sample of students to take the item tryout tests is to duplicate the conditions under which the final test will be administered. The field sample should allow for a minimum of 600 students for each test form (1200 items/50 items per form × 600 students = 14,400 students). Sometimes it is desirable to have students take two forms, which reduces the number of students by half, but requires two school periods rather than one to administer.

It is this process that keeps casual test development to a minimum. Attempts to deal with it take various forms: selecting classes rather than individual students; selecting schools rather than individual classes; and even selecting local districts rather than schools. The problem with these expediencies is that, as studies have shown, the item variance is reduced as a consequence of each of these decisions. Class scores are not as heterogeneous as individual scores, which affirms that students are not randomly selected into classes. The same relation exists between classes and schools.

This rather abstruse statistical point means that, under those circumstances, one cannot project the results to the total population with any definable degree of certainty.

Several methods have been offered to make up for deficiencies in sample selection. One, mentioned above, is the use of latent trait theory (the one-parameter Rasch model, the two- and three-parameter Birnbaum models), which claims to measure person ability independently of items and item difficulty independently of persons. Do not run a Rasch program for test *X* on two independent samples of students having different mean abilities, however, and expect to obtain the same Rasch difficulty values. What the Rasch statistics do is provide a rationale for adjusting item difficulty, assuming person ability is known. One cannot do this directly with p values, although the same scheme can be worked by converting the p values to z scores using the Thurstone judgment model and making similar adjustments for standardized ability. Like covariance adjustments, though, both of these adjustments probably work best when they are least needed. A sufficient number of experiments has not yet been reported to make a conclusive judgment.

A second method for adjusting a poor sample is to interleave all booklets within a single class. Thus, in the instance where 24 booklets were prepared to carry 1200 items, the booklets would be shipped in successive series of Forms 1 through 24, so that no two students in a class of 24 students would have the same booklet. This is an excellent way of cancelling local effects and is to be recommended in principle, but it does not solve the basic problem: namely, that the sample does not represent the population.

Another way of dealing with the problem is to weight the sample—something item analysis statistics and programs are ill equipped to do—or oversample and draw a nominally representative sample from the larger sample. The advantages of these tactics may not be worth the effort.

What if the sample is substantially defective? First, the Rasch model may still be a successful solution. Failing that, however, it is likely that even a poor sample can be used to construct equivalent tests if the tests have been interleaved at distribution. In this latter case, subsequent standardization of the tests is essential if one wishes to predict what the population will score. Inferring population statistics from even good samples of item tryouts is chancy, however. Some sort of standardization is needed.

The objective-referenced model is robust in the face of these problems. When that model is used, item statistics are used primarily to cleanse the item pool of undesirable items.

A second sampling consideration must be taken into account: the necessity of collecting enough data to conduct item bias studies relating to sex, race, socioeconomic status, geographic location, and anything else serving to disturb a common culture. In general, representation of important minorities of whatever kind will be adequately represented in a moderately large sample, because their numbers in the population is usually the characteristic that makes them important. This is not always true, however. For example, if information is desired on handicapped students of specific types, it may be necessary to employ some degree of oversampling to provide statistically comparative data.

Sometimes it becomes necessary for political reasons to base a sample on volunteers. This brings an old Italian expression to mind: *en boca lupo* (into the mouth of the wolf). In volunteer sampling, one can only pray for the best; matters are largely out of control. The worst danger in the use of "grab" samples is in the unreliable definition of difficult and easy items—those at the tails of the p value distribution.

In summary, the selection of the sample may be the task most critical to successful item development. Too much effort cannot be put into its successful completion.

9. Obtain cooperation of participants

It is highly desirable to present a broad range of testing activities to potential participants, explain the overall testing burden, and ask each to shoulder his share. That way, one need not start from scratch in obtaining cooperation in developing a specific test. In any event, it is necessary to contact participants in order to obtain concurrence in the testing activity. Inevitably, some defections will occur, some for good reasons. These must be allowed for in the sampling plan.

10. Organize procedures

In established school systems, standardized testing has been carried on for years, although in any particular case the responsibilities may have just shifted to individuals with no previous experience in testing at all. It is to the advantage of everyone to tap into this existing system to administer the item tryouts, first, because less start up and training is required, and second, because this is the system that will be administering the completed test, and is more likely to produce representative results than other administrative systems.

Ordinarily, statewide test administration proceeds hierarchically: state personnel train regional personnel; regional personnel train local school system test coordinators; test coordinators train test administrators (usually teachers); and test administrators administer the tests. The school system superintendent is a controlling link between the state and the local test coordinator; the principal is a link between the test coordinator and the teacher/administrator. In some instances, it is expedient to train regional personnel and test coordinators together.

Materials and instructions flow through the hierarchy from the apex to the base and back again. The result is that students all over the state contribute data to the test development program with hardly a ripple in the usual administrative flow of the educational system. It is an analog of the busy duck, however: calm on top of the water, paddling madly below the surface, and watching anxiously for crocodiles all the while.

After the tests have been given, they must be scored. This does not mean that they must be corrected for right and wrong answers, but rather that the data encoded on the answer sheets must be transferred to electronic form. This can be done at a central scoring center, or it can be done locally or regionally and transferred via on-line procedures or floppy disks to a central point. In either case, the object is to obtain a single clean data tape of all test results. Careful organization of the scoring program is necessary to prevent irrecoverable loss of data.

11. Print booklets and manuals

Once the size of the sample and the number of testing locations are known, arrangements can be made to print up copies of the materials from the final drafts. An oversupply should be planned in order to be able to replace lost materials, supply needed file copies and working copies, and so on. Provisions also must be made for the requisite number of answer sheets. If special answer sheets are required, lead time must be planned well in advance.

12. Package material

When the printed material is available, it should be packaged for distribution. Some pre-packaging may be available from the printer. If the booklets are to be interleaved, the best time to do it is now. Once they leave the central location, interleaving can become a difficult task to control.

13. Distribute materials

Several methods are employed for distributing materials, depending upon the amount and the timing of their distribution. In some cases, they can go hand by hand through the administrative hierarchy. In other cases, they may be sent by express to various locations.

Accounting for the arrival of materials at the test site is important to the test development program. The burden is on the state staff to confirm the arrival of materials. Failure at this step could mean the loss of a portion of the item pool, which could be an expensive mistake.

14. Administer tests

It helps to have some local acknowledgment that tests are to be given on a certain day to certain students. Occasionally, some principal will forget to test. If the school is a secondary school, it can have serious consequences for item development.

Once administration begins, questions always arise. This is where the adequacy of the supervisory training comes into play. It is not a good time to go on vacation. The critical work has just begun.

15. Collect/score tests

The collection and scoring of tests must be followed closely: missing materials should be traced promptly; scoring problems should be solved at once.

Once the tests have been scored, the data tape should be prepared and a backup file generated. Also, some disposition must be made of used materials. They contain items that must be kept secure, perhaps for years.

16. Note problems mentioned by test administrators

Some systematic method should be employed to gather and evaluate comments made by teachers. This may take the form of a questionnaire, coded spaces on a header sheet, or word of mouth through the administrative hierarchy. Reactions of students to the test may be important, although most relevant student behavior can be deduced through their answer sheets.

The test developer must recognize that the final decision-making about test content must be of two kinds: first, a qualitative decision about the item individually, with all information presented in a clear, readable form; and, second, a quantitative decision where all items are somehow considered as a whole. This need is served by two complimentary data sets:

- First is an item record in hard-copy form that contains everything that is known about the item: statistics, the item itself, any failures the item registers, and all other characteristics (objective, thinking level, etc.). Attached to this is the earlier item record and the item specification sheet. Together, three or four sheets of paper, with the item record on top, tell all there is to know about the history of the item.

- Second is an electronic file which mirrors the item record, except that it does not contain the item itself and it can only be read electronically through the data list (which can be as explicit as one desires for purpose of hard-copy printouts).

The analysis of the field test data, which is basic to the item record and electronic file, is carried out in the following manner.

The tape holding the field test results must be accessed and tested for accuracy of placement of data and for the presence of out-of-range data. After these needs are met, the data are ready to be analyzed. One procedure will be detailed here. Its purpose is to produce summary statistics for each item.

a. Data sets for test forms (usually from 10 to 15 files) are taken one by one from the data file and analyzed by a Rasch program (BICAL), modified to produce classical as well as Rasch item statistics. The program is also modified to produce a summary of all of the desired statistics (see the sample item record in Appendix E to identify the statistics produced). Once clean data sets are available, the 10 to 15 programs can be run quickly by batch on the large mainframe. The entire analysis can be executed in a matter of minutes. If desired, the entire Rasch analysis can be printed out for a permanent file. The main information of interest, however, is the summary statistics files. These can be "snapped" off the files—they all appear on the same pages, usually—by the TSO convention, cleaned of extraneous labels, and put in one file by the TSO command "include." This is hardly more than a day's work.

b. Other data not on the file (e.g., item objective numbers) are entered into a separate file for merging with the data set. This includes a bias analysis of race and sex based on partials (Item Score = *f*(Total Score, Race, Sex)). Decision parameters can be placed on the bias partials, just as on the other statistics.

c. An analysis of common items is made and constants originated to adjust each Rasch item difficulty score for student ability, thus placing the difficulty scores on a common scale. This information, however, has not been as useful as originally expected, due to the general adequacy of the samples.

d. The summary data, the objective numbers, the bias partials, and the adjusted Rasch scores are merged into a case file for items. This is the initial objective of the data analysis: an items × characteristics file.

e. When the item case file becomes available, the item statistics are analyzed by a program especially designed to make decisions concerning the acceptability of their values. For example, the program is conditioned to print out a comment, "too hard," if the item p value is under some given value, "inverted ICC" if the item characteristic curve is not monotonically increasing, or "white" if the bias statistic favors white students.

As a result of these analyses, the data are now available to complete the item record for creation of the item domain and the electronic file for item selection.

The final item record is created by starting with the item record as revised to reflect all previous item editing. To this is added two lines of item statistics and one line of decision exceptions. These data are printed on strips of contact paper directly from the computer and are transferred to the item record by clerical staff. Note that this avoids the time and errors associated with typing. The content of the item record can be conveyed quickly by the sample given in Appendix F.

When the item records are complete, the psychometric comments are noted and a decision recorded concerning the psychometric adequacy of each item.

Following the psychometric analysis, the item record goes over to curriculum specialists, who will verify the answer key and the objective assignment, note the psychometric recommendation, and make an independent recommendation. (The curriculum specialist considers factors not considered in the psychometric analysis. For example, the curriculum specialist may want to keep an item that performs no current measurement function because it is an important part of the existing curriculum and should be taught successfully.)

Following these analyses, the items are divided into those that are approved for the item pool and those for which the recommendation is negative. The revised item pool is then re-examined as a totality, to see whether the losses are serious and whether additional item development is required. The greatest danger is that losses will not be symmetrical; that they will be concentrated in certain objectives. (Careful supervision during item writing and a decision not to concentrate all of the items for one objective in the hands of a single item writer can forestall this problem to some degree. The curriculum specialist review during item edit should identify and provide means for eliminating other problems, such as a mediocre group of items or items that fail to respond exactly to the objective. This does not leave much to worry about except for clusters of items that are so unexpectedly difficult that no one can answer them or so easy that everyone can.)

The reduced item pool is then placed before the oversight management and a decision is made to go ahead with the existing item pool or to recycle into a supplementary item writing program to create additional items. If the latter is the case, the entire process must be repeated. Generally, the initial item-writing program should be able to sustain a 50% loss.

Following the definitions of the item domain, the electronic file is created for item selection. This file contains some but not all of the information from the electronic files used to create the item records. The item identification, objective number, p values or adjusted Rasch difficulty values, corrected item-total correlation, and a few other key statistics, such as level of thinking skills, are included in the working file. This information is usually downloaded from the mainframe to a Lotus file for description and graphics, although the mainframe may still be used for statistical analyses.

Constructing the Tests

Assembling the test

If the item pool is satisfactory, the assembly of the test can begin.

1. Plan for test assembly

Early in the process of test development, decisions were made about what purposes the test would serve. These guide the method of test assembly. For example:

- The test will be used only to establish a cutoff point: items must be concentrated in difficulty around the cutoff point.
- The test will be used for diagnostics: plan for at least 10 items per goal and 4-6 items per objective. Item distribution will be symmetrical by objective unless objectives are weighte
- The test will be a general survey test: plan for a distribution of difficulties (equal reliabilities at all levels of ability); or employ a domain-referenced model that weights the items by objectives; or concentrate the items at the greatest concentration of ability (a typical norm-referenced tactic, operationalized through use of the Rasch model); or employ some combination of tactics, such as:

  Prepare five test editions for simultaneous administration. Each test will contain 60 items in common across all objectives. This will provide comparative data for students. Add 200 items, 40 per test, that will not be in common across the five editions. These will be used to measure curriculum coverage (a total of 260 items). Test forms are to be interleaved within the classroom.

Usually, an objective-referenced model provides a good starting point. The domain is not of items, however, but of objectives. As Nunnally points out, a domain-sampling model holds reasonably well for multi-factor conditions, although the precision of reliability estimates may be lower. A domain-referenced model is especially appropriate for achievement tests, where the focus is on the degree to which the curriculum has been learned, irrespective of whatever mental traits are required to learn it.

Any model of test development suffers from compromises with what are known to be the facts. The types of compromises a model tolerates reveals its basic values. For example, the IRT model permits a calculation of a standard error at each point on the score scale. Test construction directing item selection toward minimizing those standard errors reveals itself as a traditional model in which validity takes a back seat to reliability. For the model used here, validity is the first concern. The assumptions are:

a. The instructional objective is the basic concept being measured. Each instructional objective is to be attained by the student; therefore, the weight of each objective is one.

b. The test should represent each instructional objective equally. In general, this means that each objective should have the same chance of being represented by the same number of items.

c. The domain of test items representing the objectives exists in the mental records of exemplary teachers who have been teaching the objectives for years. This domain can be sampled by asking a representative sample of exemplary teachers to write items to measure the objectives.

d. The domain of items is also represented in the mental records of students, but we regard the students as being a less accessible resource for converting mental records to physical records. (This assumption would bear further investigation.) When we ask students to answer the teachers' questions in item tryouts, however, we reveal the fact that we expect the item domain to have a joint representation in the mental records of teachers and students.

e. Because the domain of items is expected to have a joint representation among teachers and students, teacher-written items found to be dysfunctional in field tryouts with students must either be discarded or rewritten.

f. It is assumed that an infinite number of items can be written for each objective (except in the case of some skills objectives), and that the number of items representing a particular objective is not important as long as the number is adequate. Selection of an item is a random choice among available items.

g. Because item selection is at random from the domain of items representing an objective, the variance associated with items is neither inflated nor restricted, but, as Goldilocks said, "Just right!"

h. Once a sample of the domain of items is defined (meets minimum criteria), no item is to be preferred because of its level of difficulty or its coherence with other items in the item domain. Although exercising such preferences might contribute to a greater test reliability, it would detract from test validity. A satisfactory level of reliability has been assured by the definition of the sample of items. Nothing is gained by trying to create an appearance of a level of precision that is greater than the working materials can support, and much is lost.

Other considerations have an effect on test assembly. For example, at least one parallel test should be constructed for each test needed. This provides an indispensible estimate of reliability and extends the usefulness of the test to purposes of evaluation. One must decide whether the parallel test should be an exact clone of the first test, in which case the exact objectives and difficulties should be replicated; or whether the parallel test should be constructed under the same conditions as the first. The latter procedure serves most needs better. The reliability, although it will not be as high, will be a more accurate reflection of measurements in the objective domain and the items that represent it. The former method, in common use 50 years ago, is now regarded as somewhat naive.

2. Assemble the items for the test

The items selected for the test should mirror the objective structure. If a domain of objectives contains 25 objectives, a 100-item test can contain up to four items for each objective. If the objective domain contains 200 objectives, then only half of the objectives can be represented, and by only one item each. It is these irregularities that make test development an art rather than a science. Each condition requires the test developer to bring his or her experience to bear to make the best decision for the conditions under consideration. This involves looking ahead to the use of the data, the number of forms that can be constructed, and so on.

A domain-referenced model calls for the use of a random selection process for item selection. The basic structural unit is the objective. In short, one makes random selections for each objective. If there are more objectives than items on the test, one makes random selections of objectives, then random selections of items. The process can get to be very complicated when attempts are made to satisfy several criteria at once.

The electronic file, which can be sorted along several dimensions, is used as a basis for making the initial item selections. When the items have been selected, the estimated characteristics of the test are approximated through use of the electronic file. Adjustments are made as needed.

When an item configuration produces the desired test characteristics, the item records of those items are pulled from the item record file. They are then visually inspected by the test developer and the curriculum specialist for unexpected sequential effects.

3. Compose the test

The selected test items must now be formatted into a test form. This task will be aided greatly by the formatting designed for the item tryouts, which can be adapted to the final test, with any changes dictated by the earlier field experience. The result should be camera-ready copy.

4. Prepare administrative manuals

Again, item tryout manuals can be adapted for use with the final test. Camera-ready copy is required.

5. Conduct final edit

Xerox copies of the final products should be made and a final edit conducted.

6. Obtain managerial approval of form

Simulations of the final test should be produced by xeroxing. These, together with an estimate of the measurement characteristics of the test, should be presented to oversight management for consensual review.

Conducting Study of Consensual Validity and Standards Analyses

So far, all of the right things have been done according to prevailing technological customs. But how well, as they say, will it play in Peoria? Is there a possibility that something has been overlooked, that our test, like a newborn baby, has some flaw our parental eyes cannot discern? To answer this question, it is necessary to expose the test to a sample of people who in some way represent those who will ultimately critique the test in the field: school administrators and teachers; college professors; community leaders with an interest in education; and perhaps even a few students.

Since it also may be necessary to set cutoff scores or establish other performance standards for the test, the people who are selected to determine consensual validity for the test should be picked with an eye to the second purpose. That way, both tasks can be served at one time.

To satisfy these two issues, we must:

1. Plan for a field review of the test

The number of persons required, and their qualifications, must be decided on the basis of needed representation: geographical, cultural, educational. If cutoff scores are involved, a rationale (of which there are several) must be settled upon. For this task, community representation (parents, politicians, business people) should be strong. If a cutoff score is not needed, representation can be more clearly educational and cultural in composition. (It should be noted that the items have already undergone linguistic and statistical bias analyses—that should not be a serious problem at this stage.)

The data-gathering system must be planned, together with the types of analyses desired and the technical means of carrying out the analyses. Finally, a budget must be worked out and meeting locations identified.

2. Select group leaders and reviewers

For a typical test review in which cut scores are not at issue, eight LEA subject matter specialists are identified to manage a review—one in each educational region. Each consultant is asked to identify two exemplary teachers in the subject area of concern. These three people meet to conduct the review under the direction of the consultant, who has been provided the necessary materials by the test developer. Thus the test itself is reviewed by eight curriculum specialists and sixteen teachers.

As this activity falls under the heading of staff development, no consulting fees are paid, but payment is made for substitute teachers for the day allocated to the test review. This activity has had a high degree of acceptance, with volunteers offering to do more of the same.

3. Conduct reviews

Each of the eight LEA consultants receives three drafts of the final test, a booklet of instruction, and forms on which to record comments. Although the focus is on the test as a test, not as individual items, the desire to examine each item again has always proven irresistible to the review teams, much to the advantage of the test developer, who is treated to perhaps the most searching analysis of the items by non-psychometric professionals.

A copy of the instruction booklet and the reply forms is given in Appendix G.

4. Analyze data

When the forms (and tests drafts) have been returned, the scaled results and comments are all encoded in an electronic data base where scaled results can be summarized and comments collected by item number or a general heading. Comments can also be aggregated by the names of reviewers, thus allowing the test developer to see if a reviewer is contributing more that his or her share of comments. Again, this computerized assembling of data greatly facilitates decision-making, both of a psychometric and a curricular type.

5. Conduct final review

When the results of the test reviews are available, the final go-no go decision can be made about the test. This usually calls for a formal meeting of the interested parties (test development and curriculum), and an up or down vote. Oversight management should then be briefed concerning the decision.

Comments about items at this stage of development usually fall in the trivial class, although occasionally someone will turn up something that requires action. The scaled comments have proven to reflect later opinion accurately.

Pilot Test Administration

This is the phase of test development that major publishers use to standardize a test on a norm group. In the pilot test administration, one administers the test to a group of students who are selected to represent the entire population of students eligible to take the test. (In this respect, the goal does not differ from the best sampling solution to the item tryout problem.) The purpose, however, is to establish the basic psychometric characteristics of the test: item p values; validity and reliability estimates; percentile scores; and so on.

Getting acquiescence in the standardization of a test has proven to be the most difficult step to accomplish. Of particular concern is that the procedure adds a year to test development. For that reason, other ways were devised to handle the problem. Standardization can be done as part of the first regular administration of the test. For grading purposes, teachers would have only the raw scores at the end of this first administration. But if other test forms are administered to samples of students at the same time, the state norms can be set up for later use over a period of years.

Standardization from the first regular administration has greater feasibility for states than for commercial publishers. The students receiving the first regular administration of the commercially-developed test will not be representative of all students in the United States; they will represent the school systems electing to buy the commercial tests. In the state administration, the students comprise the total population of eligible students. Thus the state test developer has an advantage not available to commercial publishers developing nationally-normed tests.

Test Redesign versus Multiple Norms

The special advantage a state has in giving the test to the entire population also permits some other developments not usually available to a commercial publisher. If five forms of a test are developed, one form can be chosen as the first form of record. (Using the domain-referenced model, all forms are statistically equivalent except for chance differences.) The other four forms can be administered to representative samples of the student population, say a week before the "test of record" (i.e., the fifth form). The four forms can then be equated to the fifth form—re-engineered to whatever degree necessary to make them psychometrically equivalent to the fifth form, all up and down the scoring scale. Then, the table of norms worked out for the fifth form as the test of record can be used for all five forms. (Note that the scores on the four forms are not equated to the fifth-form scores of the students who had just taken one of the other four forms, but to the scores of students who did not take one of the four forms. This gets around the practice or familiarity effect of taking two forms.)

The redesigned forms are subject to the study of consensual validity and standards analysis described earlier. It usually is expedient to review a test in the year it is to be used, rather than attempting to review as many as 25 test forms simultaneously.

Test development is about over at this point, except for one essential component: a test technical manual. A technical manual is essential for the proper use of a test. A technical manual will contain the item characteristics, test characteristics, information on instructional objectives measured by the test, and norms tables. This information has three uses: It estimates the performance level for the state along all relevant dimensions; it provides a record of the test characteristics, from which judgments can be made about how scores relate, and how stable they may be; and it provides the parameters for scoring programs that produce standard scores, percentiles, and other norm scores for individual students. The technical manual should be made accessible to those who will be using the tests. An example is given as Appendix H.

Analyzing Usage Results—and Beyond

Administering the test

The test development is now over at this point, except for some closing analyses. To administer the test, the usual format must be followed:

- Plan administration
- Communicate procedures to field
- Train administrators
- Prepare scoring procedures
- Print tests, manuals, and administrative materials
- Package and distribute tests
- Administer tests
- Score tests and prepare tape
- Publish results

The possibilities of catering results to particular needs is endless. Cheap computer storage and processing has created a new age in the possible utilization of test results—for curriculum analysis; for diagnosing of instructional needs; for evaluation; and for accountability.

Analyzing Posttest Data

To establish and maintain the quality of the operation, the test developer must analyze the administrative results by doing selective studies of the test results—the accuracy of prediction of item difficulties; the equivalence of alternate forms; and so on.

Planning for the Next Edition

The operation of a testing program requires constant attention to test development. Tests quickly become insecure if the same form is given over and over. Also, instructional objectives change and tests must be changed to reflect that. Furthermore, tests will need to be expanded as users find more applications for them.

The test developer must be alert to the opportunity to administer new trial items in regular test administrations in order to keep the item pool current. This can save both time and money in item generation and provide an ideal tryout environment.

Appendices

Appendix A

Test Development Procedures

(Prepared by Division of Research Staff)

TEST DEVELOPMENT PROCEDURES

Note: Items with double star (**) mean that the final product or decision should be checked with Director before moving ahead.

GENERAL INFORMATION

Director	General Correspondence
Director	Background Information
Admin. Asst.	Archives files contains: – a general correspondence file under general files and – a file labeled "Background Information".

Specifying the Instructional Objectives

Director	1. Obtain a list of curriculum objectives.
Director	2. Determine which objectives are proposed for testing.
Research Asst.	3. Type a list of these objectives in a word processing file.

Curriculum Survey Procedures

Research Asst.	**1. Prepare survey according to content specifications.
Director/Staff	**2. Decide which teachers to sample.
Research Asst.	3. Identify a statewide list of teachers (printout).
Admin. Asst.	4. Convert list of teachers into mailing labels.
Research Asst.	5. Count and note the number of teachers to be sampled in each LEA.
Director/Staff	6. Prepare the contents of a cover letter.
Admin. Asst.	**7. Type and mail a cover letter and a survey to each teacher.

When surveys are returned,

Consultant	1. note the rate of return overall by LEA to ensure respondents are representative of the State as a whole,

Consultant

2. aggregate the results and compile the comments; use NMI.prog.survey,

Director

3. provide the Instructional Services Consultant with the results and comments, and

Director/ Instructional Services Consultant/ (Assistant State Supt.)

4. make a final list of the objectives to be measured. (If objectives do not agree with the objectives previously approved by Board, consider desirability of obtaining Test Commission and Board concurrence.)

Admin. Asst.

Archive file contains:

- a copy of the survey with goals and objectives
- a copy of the cover letter
- a statewide list of teachers
- a copy of the aggregated results and compiled comments
- any general or supplemental information including the number of surveys sent to each LEA and the rate of return
- a record of the decisions made jointly by the Division of Research and the appropriate Instructional Services division

Test Specifications

Procedures

Director/Staff

1. Through consultations with staff, test commission, Board members, assistant superintendent, Instructional Services, etc., determine what characteristics the test is to have (length, number of forms, objectives, etc.).

Admin. Asst.

Archive file contains:

- working file (transferred from general files--see General Information section on page A-1)

CREATING ITEMS

Writing Items

Director 1. Determine the number of items to be written for each objective.

Director 2. Determine the number of teachers needed to write items.

Director 3. Contact the designated Instructional Services Consultant for names, addresses (home and school), phone numbers (home and school), and social security number of teachers who will write items.

Consultant/ Research Asst. **4. If necessary, contact teachers to

(a) confirm their willingness to participate in the item-writing process and inform them that they will receive an honorarium for their work;

(b) check to be sure that we have their correct addresses and phone numbers;

(c) explain the process, tell them how many items they will be expected to write, and provide them with a brief description of the materials they will receive and how those materials should be used (may want to tell them that they will be expected to return the video tape along with the items);

(d) inform them that they will receive a **Personal Services Contract** form and a **Personal Services Contract Certification** form which must be completed, signed, and returned before materials can be mailed;

(e) determine whether the item writer would like to have the materials mailed to his/her home address or school address; and

(f) tell them when they should expect to receive materials and the deadline date for returning the items.

Research Asst. 5. Assign each item writer a unique identification number.

Research Asst. 6. Create a database file containing each item writer's name, address (home and school), phone number (home and school), identification number, and region. Print a list of the item writers that includes all of the information entered into the file.

Admin. Asst. **7. For each item writer, prepare and mail a **Personal Services Contract** form, a **Personal Services Contract Certification** form, and a cover letter with instructions. **(Monitor closely the return rate of forms.)**

Research Asst. 8. Prepare a list of goals and objectives.

Research Asst. **9. Prepare **Item Specifications Form** using the following procedures.

(a) Make sure there is one printed form for each objective.

(b) Get appropriate number of objective forms duplicated.

(c) Reassemble all printed forms in objective order.

(d) Stamp each form with a unique item number.

(e) While items are in objective order, decide how many items will represent a difficulty level of easy, medium, or hard and circle 1 (easy), 2 (medium), or 3 (hard) on each form.

(f) While items are in objective order, decide how many items will represent a thinking skill level of low or high and circle 1 (low) or 2 (high).

SPECIAL NOTE: For any future development of items for U.S. History, Biology, Algebra II, Science, or Social Studies, "1" should indicate "High", and "2" should indicate "Low" to be consistent with forms used in previous years.

(g) While items are in objective order, specify the placement of the correct response by writing an "A", "B", "C", or "D" on the blank next to "Correct Answer".

Director and Research Asst.

10. Determine how items will be distributed among teachers and then reassemble items.

Research Asst.

11. Write each item writer's identification number in the appropriate space on each of the writer's **Item Specification Forms.**

Research Asst.

12. Prepare an **Item Development Log.** (This form is used as a tracking device for all items.)

13. Prepare the instructional materials that will accompany the **Item Specification Forms** and edit these materials to reflect the appropriate subject area.

Editor

(a) **Packet A: Edit to include goals, objectives and item samples for the subject area.

Editor

(b) **Packet B: Edit to include appropriate item samples for the subject area.

Editor

(c) **Packet C: Edit to reflect the subject area.

Director

(d) Script for How To Write Multiple-Choice Achievement Test Items

Director

(e) videocassette

Director

(f) cover letters, one for contracts, one for materials

Admin. Asst.

14. Prepare a brown envelope (addressed to our office or contractor's office) to be used by item writers for returning items and videocassette. Stamp "First Class" on each envelope.

Admin. Asst.

15. Prepare an address label and cover letter for each item writer.

16. When a signed copy of both the **Personal Services Contract** and **Personal Services Contract Certification** form has been received from each item writer,

Admin. Asst.

(a) check to be sure both forms have the required information and signatures,

Admin. Asst.

(b) xerox one copy of each form and forward originals to accounting office, and

Admin. Asst./
Research Asst.

**(c) assemble and mail a packet of materials to each writer. Each packet should contain

(1) a cover letter,
(2) the item-writing forms,
(3) a copy of Packet A,
(4) a copy of Packet B,
(5) a copy of Packet C,
(6) a copy of the Script,
(7) Guidelines for Bias-Free Publishing,
(8) a copy of the videocassette, and
(9) a return envelope

Consultant/
Research Asst.

**17. As the deadline date for return of items approaches, contact those teachers who have not returned their items to verify that the task will be completed on time.

18. As items are returned,

Consultant/
Research Asst.

(a) indicate the date received on the item-writer list and check items in on the Item Development Log. Be sure to indicate, by the item number, whether or not artwork is needed—(1) for "Yes", (2) for "No".

Admin. Asst.
(upon
notification
from
Consultant)

**(b) Submit authorization for payment (indicating completion of task) to the accounting office.

Director/Staff

19. Forward returned/logged items to the designated college professor/English editor for review and necessary revisions. Prepare any contract forms required for their services. (**Special Note:** Ask the editor to use a different colored pen/pencil from the one used by the specialist, and ask both to initial each form they review.)

Director

20. Inspect editing changes.

Director

21. Approve consultant fees and pass for payment to contracted accounting.

Director

22. Arrange for the Instructional Services Consultant to review all items. (The Consultant should use a different colored pen/pencil than those used by the specialist or editor, and should initial each form reviewed.)

Director

23. Inspect changes made by Instructional Services Consultant and OK reviews.

Research Asst.

24. Following this review,

(a) note any deletions on **Item Development Log**,

(b) enter all accepted items into a database file (Filemaker Plus or 5520 file, whichever is appropriate), and

(c) create all accompanying artwork and label artwork documents using the original item number.

Research Asst.

25. After all accepted items have been entered into the database file and all artwork has been created,

(a) print a list containing the original item number, the objective number, and the artwork information;

(b) print a list of all artwork files;

(c) compare the database list with the **Item Development Log** to be sure all items have been entered, that the item record contains the correct objective number, and that each record correctly specifies whether or not there is accompanying artwork (make necessary corrections); and

(d) compare database list with list of artwork files to be sure that all artwork has been created. (Make necessary corrections and print new lists.)

Research Asst.

The item development process is now complete. **Test Specification Forms for all accepted items should be reassembled in item number/objective number order. **Test Specification Forms** for all deleted items should be grouped together and reassembled in item number/objective number order.

Admin. Asst. Archives should contain

(1) a file labeled "(Content Area) Item-Writing Information". This file will contain

(a) a list of the item writers, including names, identification numbers, addresses, phone numbers, a list of the items written by each, and the date items were returned;

(b) a signed copy of each item writer's **Personal Services Contract** and **Personal Services Contract Certification** and a copy of the cover letter mailed with these forms;

(c) a copy of the cover letter and all instructional materials;

(d) a copy of the **Item Development Log**; and,

(e) a list of artwork files.

(2) file folders, labeled by objective number, which contain all accepted items for those objectives.

(3) a file folder labeled "Rejects", which contains all rejected items.

Conducting Item Content Analyses (Item Review)

Director 1. Determine the number of teachers needed for item review.

Director 2. Determine the number of items each teacher will review.

Director 3. Contact the Instructional Services Consultant for a list of reviewers. **(There should be an equal number of teachers representing each of the eight educational regions. Each item will be reviewed by one teacher in each of the eight regions.)**

Consultant **4. If necessary, contact each reviewer to

(a) confirm their willingness to participate in the item review process and to inform them that they will receive an honorarium for their work;

(b) check to be sure we have correct addresses, phone numbers, and social security numbers;

(c) explain the process, tell them how many items they will be expected to review, and provide them with a brief description of the materials they will receive and how those materials should be used;

(d) inform them that they will receive a **Personal Services Contract** form and a **Personal Services Contract Certification** form which must be completed, signed, and returned before materials can be mailed;

(e) determine whether the reviewer would like to have the materials mailed to his/her home or school ad dress; and

(f) tell them when they should expect to receive materials and the deadline for returning the materials.

Research Asst. — 5. Assign each reviewer a unique identification number.

Research Asst. — 6. Create a database file containing each reviewer's name. address (home and school), phone number (home and school, if available), region, and identification number, and print, by region, a list of reviewers and all related information.

Director — 7. Write a cover letter; confirm that **Instructions for Item Review** is appropriate in this context—otherwise, see to edit and reprinting.

Admin. Asst. — **8. For each reviewer, prepare and mail a **Personal Services Contract** form, a **Personal Services Contract Certification** form, and a cover letter with instructions.

Admin. Asst/ Research Asst. — **9. Revise Item Review form according to content area and have appropriate number of copies duplicated.

Research Asst.

10. Prepare item review sheets by following the procedures listed below. (These procedures are for files created with **Filemaker Plus**. Item review sheets created from files on the 5520 would be produced using a merge control document.)

    (a) Use the item review sheet layout in the database file to select items by objective.

    (b) As each set of items is selected, sort those items by original item number (if necessary) and output to a text file.

    (c) Use **File Converter** to convert

      (1) tabs (\t) a space,

      (2) "&" to tabs (\t), and

      (3) "@" to returns (\r)

    in each text file. The result will be a new text file for each set of items.

    (d) Open new text files as MacWrite documents.

    (e) Use **Tempo** to insert a page break after each item so that each item is on a separate page.

    (f) Use the "Find and Replace" feature to change the "1", "2", or "3" by **Difficulty Level** to "Easy", "Medium", or "Hard", and the "1" or "2" next to **Thinking Skill Level** to "Low" or "High". **(Remember to use "1" for "High" and "2" for "Low" for future review of additional items for U.S. History, Biology, Algebra II, Science, and Social Studies.)**

    (g) Print, from database file, a list of items for each objective. (This list should show the original item number, the objective number, and artwork information.)

(h) Use this list to

(1) confirm that all items for a particular objective were retrieved from the database, and

(2) identify items with artwork.

(i) Place all artwork with appropriate item.

(j) Print each document of items on item review forms. **Remember to keep printed forms in objective order.**

Admin. Asst.

**11. When all item review sheets have been printed, xerox (or duplicate) eight copies of each sheet and reassemble (in objective order) into eight complete sets, one for each of the eight regions.

Research Asst.

12. Determine how items will be distributed among the teachers in each region an... reassemble accordingly.

Research Asst.
/Admin. Asst.

13. Stamp each reviewers identification number on the item review sheets he/she will receive.

Research Asst.
& Consultant

14. Prepare an **Item Review Log** which contains each reviewer's name, identification number, the original item number for each item he/she will review, and the objective number for those items. **This process could be streamlined if the first reviewer in each region received the same items, the second reviewer in each region received the same items, etc.**

Admin. Asst.

**15. Assemble, for each item reviewer, a set of the instructional materials which will accompany the item review forms.

(a) **Instructions for Item Review of Achievement Test Items**

(b) **Script for How to Writer Multiple-Choice Achievement Test Items**

(c) **McGraw-Hill's <u>Guidelines for Bias-Free Publishing</u>**

Admin. Asst. 16. Prepare a brown envelope (addressed to our office or our contractor's office) to be used by reviewers for returning items, and stamp each envelope "First Class".

Director 17. Prepare a cover letter.

Admin. Asst. 18. Prepare an **address label** and type **cover letter** for each reviewer.

Admin. Asst. **19. When a signed copy of both the **Personal Services Contract** form and **Personal Services Contract Certification** form has been received from each item writer,

(a) check to be sure both forms have the required information and appropriate signatures and xerox one copy of each form;

(b) send the original of both forms to the accounting office; and,

(c) assemble and mail a packet of all materials to each reviewer. Each packet should contain

(1) a cover letter,
(2) item review sheets,
(3) **Instructions for Item Review of Achievement Test Items,**
(4) **Script for How to Write Multiple-Choice Achievement Test Items,**
(5) **McGraw-Hill's Guidelines for Bias-Free Publishing**, and
(6) a return envelope.

Consultant **20. As the deadline date for return of items approaches, contact those reviewers who have not returned items to verify that the task will be completed on time.

21. As items are returned,

Consultant (a) indicate the date received on the **Item Review Log**;

Consultant (b) check items in on the Item Review Log, noting, in the right margin, the total number of review sheets received for each item; and

Admin. Asst. (upon notification from Consultant)

**(c) submit authorization for payment (indicating completion of task) to the accounting office.

Director/ Consultant

**22. Forward item review sheets to contracted personnel for aggregation of data in the approval/problems section, summary of correct responses, and compilation of comments.

Director/Staff

23. When item review data analyses are complete, a meeting will be held between Research and the appropriate Instructional Services division to go over the results for each item. Decisions will be recorded on the item review record.

Director/ Editor (overload to Peace or Meredith)

24. All items that were modified should then go to an English editor for review. He/she should also initial each item reviewed. **(This editor should be asked to use a different colored pen/pencil than the one used to record the results of the item review.)

Research Asst.

25. Once an item has been logged in and reviewed by an English editor (if needed), all requested changes to item content, key, or artwork should be made to the database or artwork file.

Admin. Asst.

26. When al' corrections have been made, reassemble all approved item records into original item number/ objective number order. (The multiple copies of item review sheets can now be shredded.)

Research Asst.

27. Reassemble all rejected items into the same order, check them off on the **Item Development Log** as being **Rejected after Item Review**, delete each rejected item from the database file, and then file these items as "Rejects". Keep the artwork files for deleted documents for possible use in the future. Type "Reject" after the document name, copy these files to a floppy disk, and label the floppy disk appropriately. (The multiple copies of these item review sheets can also be shredded.)

Director/Staff

28. Review results of analyses.

The item review process is now completed.

Admin. Asst.

Archives should contain:

(1) a file labeled "(Content Area) Item Review Information". This file will contain

(a) a list of the item reviewers, including names, identification numbers, addresses, phone numbers, a list of the items reviewed by each, and the date items were returned by each reviewer;

(b) a signed copy of each item reviewer's **Personal Services Contract** and **Personal Services Contract Certification** and a copy of the cover letter mailed with these forms;

(c) a copy of all instructional materials (mailed with item review sheets) and the cover letter mailed with those materials; and

(d) a copy of the **Item Review Log**.

(2) the file folders, labeled by objective number, which contain the original item writer's form and, stapled on top, an item review sheet containing decision data for each approved item; and,

(3) the file folder, labeled "**Rejects**", which contains all rejected items.

(NOTE: Remove the original item-writer forms for items rejected after the item review from the objective folders and attach them to the appropriate item review sheet in the "Rejects" file.)

EXAMINING ITEM PSYCHOMETRICS

<u>Testing Items (Field Testing Procedures)</u>

(Steps 1–3 can be revised to accommodate the needs of a particular content area.)

Director/Staff

1. Review test specifications and extent of item pool.

Consultant **2. From the total number of items to appear on each field test, subtract the number of common items (see below). Divide the remainder into the total number of items in the item pool. **The result will be the number of field tests to be administered.**

Consultant 3. In selecting the ten common items (if needed),

(a) the Instructional Services Consultant recommends which objectives will be represented in the common items, and

(b) a research assistant selects one item from each of those objectives (approximately 25% of the items should be easy, 50% should be medium, and 25% should be hard).

Consultant **4. The location of the common items on the test form is decided.

Research Asst. 5. Arrange the hard copies of the remaining items in stacks by objective.

Research Asst. 6. Allowing for the common items (if any), collate from the stacks of objectives until you have enough items for one test. Repeat this process for all tests.

Research Asst. 7. Name field tests “1” through “?” (depending on the number of tests) and number each hard copy in the upper right-hand corner (example: 1-11, 1-12; 2-11, 2-12, etc.).

Research Asst. 8. Prepare a manila folder for each form of the test and insert the hard copy of test items for that form.

Research Asst. 9. Enter the form and form item number for each item into the database file. **Enter “Z” - the form for common items.**

Research Asst. 10. Print a list of the items in each form (presort file by form and form item number). **This list should include the form, the form item number, the original item number, the objective number, the difficulty level, the thinking skill level, the information related to artwork, and the correct answer.**

Research Asst.

11. Make sure the proportion of correct responses ("A's", "B's", "C's", and "D's") are approximately equal on each test, adjust as needed, and indicate any changes (adjustments) on the printed list. Make all changes to

    (a) the item record and

    (b) the database file.

    Be careful with math items where answer choices are in numerical order.

Research Asst.

12. Compare all of the data on each list with the data on the hard copy of item records, note any corrections on each list, and then make those corrections

    (a) on the item record and

    (b) in the database file.

Research Asst.

13. Print new lists of items and compare the new lists with (1) the old lists and (2) the hard copy of item records to be sure all changes were made in both places. Reprint a list if any corrections were made to the database file after these comparisons.

Research Asst.

14. Provide the Educational Research Consultant with a copy of each list.

Research Asst.

15. Pull the items for each test form from the database file into Pagemaker and place all artwork with appropriate item. **Each test should be in two parts and correctly labeled Part 1 or Part 2. There should be a **"The End"** at the end of each test.

Research Asst.

**16. Prepare a cover page for each form of the test.

Consultant/
Research Asst.

17. Prepare a page of two (2) sample items, one copy for each form of the test. **(Sample items should be representative of items found in the test.)**

Research Asst.

18. "**North Carolina Field Test of ________, Form __**" should appear at the top of each page of the test, including the page of sample items.

Research Asst.

19. Page numbers should appear at the bottom of each page (except for the cover page and page of sample items). **Begin numbering with Page 1.**

Editor/
Research Asst.

20. Once a test has been printed, compare each item on the test with the hard copy of the item record to be sure wording and art is consistent. **Check, again, to make sure the order of foils reflects any changes made to the answer key.** Make any necessary adjustments and/or corrections to

(a) the hard copy of item records,

(b) the database file, and

(c) the Pagemaker document.

Reprint any corrected pages of the test.

Editor/
Research Asst.

21. Check each test to be sure that

(a) it **does not** contain duplicate items,

(b) the length of the foil is not a clue to the correct response,

(c) the test contains the correct number of items,

(d) each item contains the correct number of foils,

(e) **"Part 1"**, **"Part 2"**, and **"The End"** are in place,

(f) page numbers and test/form identification (see #17) are correct and correctly placed,

(g) the sample items are in good array and consistent with items in the test,

(h) the cover is correct, and

(i) the last page contains copyright information.

Research Asst.

22. Attach a **correct** answer key to the camera-ready copy of each form of the test.

Director

23. Ask Instructional Services Consultant to make final review.

Director/ Editor

24. Have tests read one more time for grammar.

Director

25. Preparation of field tests is complete. Pass on to Division of Testing for reproduction and administration.

Director/Staff

26. Consult on selection of student samples.

Admin. Asst.

Archives contains

(a) files for final copies (test booklets) of each form of the test,

(b) a file for the hard copy of item records for the common items,

(c) a file for the answer keys for each form of the test, and

(d) manila folders containing the hard copy of item records for each form of the test. **(Hanging folders do not need to be made for these item records since they will be either returned to the objective folders or placed in the "Rejects" file after field test analyses are run.)** Include in each folder a list of the items in that form of the test. **This is the list that includes the form, the form item number, the original item number, the objective number, the difficulty level, the thinking skill level, the information related to artwork, and the correct answer.**

Analyzing Items and Making Revisions

Field Test Analysis Procedures

Research Asst.

1. Prepare hard copy of final item records. These records

   (a) specify the content area and the objective,

   (b) contain a cut and taped copy of the item as it appeared in the field test,

   (c) contain space for labels with the item statistics

(d) contain space to indicate whether the item received **Psychometric Approval, Edit Approval, Curriculum Approval, Committee Approval, and Final Approval.** Items receiving Final Approval will remain in the item pool.

Consultant

2. Run analyses on field test results.

(a) Obtain hard copy of field test data file and key file designations from the Division of Testing.

(b) Run Rasch reliability analysis for each field test form.

(1) Use the Program.writdata to read in data, place items starting column 1 through column ? [corresponding to the number of items on the test form, and write data to a dataset (*.data.form?)].

(2) Create file for keys with headers for Rasch analysis.

(3) Combine steps (1) and (2), delete unused keys, attach footers, and print the resulting data set.

(4) Modify and run Program.Rasch. This program invokes NTIME.BICAL.Loadlib.

(5) In Level 2 IOF, snap last two tables (pages) of Rasch analysis and separate into two files (*.tab1.form? and *.tab2.form?). Remove headers and footers acquired during Rasch analysis.

**(c) Execute gender and ethnic bias anlaysis procedures for each field-tested item and output results (partial correlation) to a dataset.

(d) Print item statistics labels for all items.

Research Asst. /Consultant

(1) Obtain item information from item files on the 5520 or the Mac—original item number, field test form, item number on form, difficulty level, thinking-skill level, and objective number. Change format of core item numbers (0#). Change @, ob, difficulty level: to blanks, and tabs to spaces. Upload information to a TUCC file (*.obj) using MacLink and/or Kermit.

Consultant

(2) Output Rasch difficulty scores for common items into a data set (use Program.rawread to read scores from *.tab1.* for all forms) and output to *.common.rasch. Analyze common item difficulty scores (use Program.rawline) and adjust the scores for equal ability samples for each field test form.

(3) Modify and run Program.labels. Use the two tables from the Rasch analysis (*.tab1.form? and *.tab2.form?), the item information (*.obj), the item bias results *.bias.form*?), and the adjustment to the Rasch difficulty score to print labels. Also, output the item statistics to a file (*.itemstat.form?).

(4) Combine the item label information (*.itemstat.form?) for each field-tested form into one file for all forms (*.itemstat.all).

(e) Run analyses to obtain descriptive statistics and distributions of the item statistics (using *.itemstat.all file) for the overall item pool and by objective.

(f) Print item warning labels for item pool (all forms combined) (use *.itemstat.form? files).

**(1) Determine cut-off values for item statistics from the distributions of the item statistics.

(2) Modify and run Program.cutoffs.

(3) Print warning labels for items within each form.

(g) Run analyses to obtain descriptive statistics and distributions of the item statistics (using *.itemstat.all file) for the overall item pool and by objective for retained and dropped items.

Research Asst.

3. Place item statistics labels and warning labels on hard copy of final item records.

Research Asst.

4. Attach final item record to previous item records (in field test folders) and file in objective folders. Place staple in the upper left-hand corner.

Director/
Designated
Instructional
Services
Consultant

5. Review items for curricular and psychometric adequacy.

Research Asst.

6. File rejected items in "Rejects" folder.

Consultant

7. Input decision into the *.itemstat.all file from the *.baditem.all file. Run analyses to obtain descriptive statistics and distributions of the item statistics of the accepted items (using *.itemstat.all file) for the overall item pool and by objective.

Consultant

8. Create item pool by downloading *.itemstat.all file to Lotus on the PC. Create the following worksheets using the xtract command:

(a) itempool.wk1 with all of the item statistics for all items available for use in test development,

(b) itemusag.wk1 with the basic information (original item number, field test form, and field test item number, and objective) and field test statistics (p-value and adj Rasch difficulty) and columns for usage of item during each year of item pool use (form, item number, and key), and

(c) rejects.wk1 with the item statistics for all rejected items.

Admin. Asst.

Archives contains:

1. Complete sets of item records, refiled into the objective folders. (There is no longer a need for field test folders.)

2. Reject file for the item records of rejected items.

3. Analyses Files:

   (a) Raw data used in Rasch analysis with headers, footer, and keys for each field test form.

   (b) Item statistics analyses (programs and documented results and decisions).

   (c) Printout of Lotus worksheets (itempool.wk1, itemusag.wk1, and rejects.wk1).

CONSTRUCTING TESTS

Assembling the Tests

1st-Year Statewide Testing Procedures

Director/
Consultant

1. Decide how many forms of the test there will be and how many items will be on each form.

Director/
Consultant

2. Decide how many of the items on each form will be core items and how many will be variable items.

Director/
Consultant

3. Decide whether additional tests of core items should be field tested during the statewide administration.

Director/
Consultant

4. Select statewide core and variable items and determine placement of items within the test. (If tests of core items will be field tested, these items should also be selected.) The first item on each form of the test should be a variable and have a high p-value.

Consultant

5. Set up Lotus files of item statistics for each form and determine equivalency.

Research Asst. 6. Arrange the hard copy of item records into core and variable sets. Name the test forms using letters of the alphabet ("A" through "?"—depending on the number of tests) and number each item record in the upper right-hand corner (example: A1, A2; B1, B2). Use the letter "Z" to name the statewide core items (Z1, Z2, etc.). Name the field tests of core items using numbers. Begin naming these tests with the number following the last one used for the original field tests

Research Asst. 7. Prepare a manila folder for each test form and insert the item records for that form.

Research Asst. 8. Enter the form and form item number for each item into the database file. **Remember to enter "Z" as the form for statewide core items.**

Research Asst. 9. Print a list of the items in each form, including the core items (presort file by form and form item number). This list should include the form, the form item number, the original item number, the objective number, the correct answer, and artwork information.

Research Asst. 10. Make sure the proportion of correct responses ("A's", "B's", "C's", and "D's") are approximately equal on each test, adjust as needed, and indicate any changes (adjustments on the printed list). Make all changes to

(a) the item record and

(b) the database file.

Be careful with math items where answer choices are in numerical order.

Research Asst. 11. Compare all of the data on each list with the data on the hard copy of items records, note any corrections on each list, and then make those corrections

(a) on the item record and

(b) in the database file.

Research Asst.

12. Print new lists of items and compare the new lists with (1) the old lists and (2) the hard copy of item records to be sure all changes were made in both places. Reprint a list if any corrections were made to the database file after these comparisons.

Research Asst.

13. Provide the Educational Research Consultant with a copy of each list.

Research Asst.

14. Pull the items for each test form from the database file into Pagemaker and place all artwork with appropriate item. **Each test should be in two parts and correctly labeled Part 1 or Part 2. There should be a "The End" at the end of each test.

Research Asst.

**15. Prepare a cover page for each form of the test and place copyright.

Consultant/
Research Asst.

16. Prepare a page of two (2) sample items, one copy for each form of the test. **(Sample items should be representative of items found in the test. May be able to use the same sample items as those used with the field tests.)**

Research Asst.

17. **"North Carolina Test of ________, Form ___"** should appear at the top of each page of the test, including the page of sample items.

Research Asst.

18. Page numbers should appear at the bottom of each page (except for the cover page and page of sample items). **Begin numbering with Page 1.**

Editor/
Research Asst.

19. Once a test has been printed on the laser printer, compare each item on the test with the hard copy of the item record to be sure wording and art is consistent. **Check, again, to make sure the order of foils reflects any changes made to the answer key.** Make any necessary adjustments and/ or corrections to

(a) the hard copy of item records,

(b) the database file, and

(c) the Pagemaker document.

Reprint any corrected pages of the test.

Editor/
Research Asst.

20. Check each test to be sure that

(a) it does not contain duplicate items,

(b) the length of the foil is not a clue to the correct response,

(c) the test contains the correct number of items,

(d) each item contains the correct number of foils,

(e) "Part 1", "Part 2", and "The End" are in place,

(f) page numbers and test/form identification (see #17) are correct and correctly placed,

(g) the sample items are in good array and consistent with items in the test,

(h) the cover is correct, and

(i) the last page contains copyright information.

Research Asst.

21. Attach a correct answer key (with historical/statistical data) to the camera-ready copy of each form of the test.

Director

22. Provide a copy to the Instructional Services consultant for final review (if desired).

Preparation of statewide tests is complete.

Admin. Asst.

Archives contains

(a) files for final copies (test booklets) of each form of the test,

(b) a file for the hard copy of item records for the common items,

(c) a file for the answer keys for each form of the test, and

(d) files for the hard copy of item records for each form of the test, including a list of the items in that form. **(This list, provided by the Educational Research Consultant, includes the form, the form item number, the original item number, the objective number, the correct answer, and statistical/ historical data.)**

Conducting Consensual Validity and Standards Analyses

Test Review Procedures

Director — 1. Determine the number of teachers and, if standards are involved, other persons needed for test review.

Director — 2. Contact the Instructional Services Consultant for a list of reviewers (LEA supervisors). **(There should be one LEA curriculum supervisor and two teachers (of the supervisor's choice) representing each of the eight educational regions.)**

Consultant — 3. If necessary, contact each supervisor to

(a) confirm willingness to direct and participate in the test review process;

(b) check to be sure we have correct addresses;

(c) explain the process, tell the supervisor how many tests the team will be expected to review, and provide a brief description of the materials we will send and how those materials should be used;

(d) inform the supervisor that designated teachers will receive a **Request for Reimbursement of Substitutes** form to be used for reimbursing substitute teachers needed during the day they are reviewing tests; and

(e) tell the supervisor when to expect materials and the deadline for returning these materials.

Research Asst.

4. Assign each reviewer a unique identification number. (Example: The supervisor in Region 1 will be "1.0", and the two teachers in that region will be "1.1" and "1.2".)

Research Asst.

5. Create a database file containing each supervisor's name, address, phone number, and region; and print, by region, a list of the supervisors that includes all of the information entered into the file.

Admin. Asst.

6. Xerox (or duplicate) 26 copies of each test form.

Director

7. Write a cover letter. Check other materials fcr needed editing.

Admin. Asst./
Research Asst.

**8. Prepare for each supervisor a set of the instructional materials, questionnaires, and summary forms that will accompany the tests. (Some of these materials will be color coded to distinguish between supervisor's materials and tcachers' materials.

(a) **Procedures for Instructional Review**

(b) **Curriculum Goals and Objectives**

(c) **Notes to the Curriculum Supervisor**

(d) **Questionnaire for Curriculum Supervisor Evaluation**

(e) **Curriculum Supervisor's Summary of Teacher Reviews**

(f) **Questionnaire for the Teacher Evaluation**

Admin. Asst.

9. Prepare a brown envelope (addressed to our office or contractor's office) to be used by supervisors for returning tests, questionnaires, and summary forms. Stamp each envelope "First Class".

Admin. Asst.

10. Prepare an address label and the cover letter for each supervisor.

Admin. Asst./
Research Asst.

**11. Assemble a packet of all materials for each supervisor. Each packet should contain, in predetermined order,

(a) a cover letter,

(b) **Notes to the Curriculum Supervisor,**

(c) **Questionnaire for Curriculum Supervisor Evaluation,**

(d) **Curriculum Supervisor's Summary of Teacher Reviews,**

(e) 3 copies of **Procedures for Instructional Review,**

(f) 2 copies of the **Questionnaire for the Teacher Evaluation,**

(g) three sets of each form of the test

(h) three copies of the **Curriculum Goals and Objectives,**

(i) **Request for Reimbursement of Substitutes** forms (2 copies), and

(j) a self-addressed envelope for returning documents.

Admin. Asst./
Research Asst.

12. Write each reviewer's identification number on the appropriate questionnaire. [Example: supervisor (1.0), teachers (1.1, 1.2)]

Admin. Asst.

13. Mail materials to supervisors.

Research Asst.

14. Prepare a **Test Review Log** which contains each supervisor's name, region, and a list of the materials he/she must return.

Director

15. Provide the designated content professor with a xerox copy of each form of the test. Ask him/her to (1) review each core and variable item, (2) note any changes directly on the test form, and (3) circle the correct answer.

Director	16. Provide the designated Instructional Services Consultant with a xerox copy of each form of the test. Ask him/her to (1) review each core and variable item, (2) note any changes directly on the test form, and (3) circle the correct answer.
Consultant	**17. As the deadline date for return of tests and materials from supervisors approaches, contact those supervisors who have not returned materials to verify that the task will be completed on time.
	18. As materials are returned by supervisors,
Consultant	(a) indicate the date received on the **Test Review Log**;
Consultant	(b) check materials in on the **Test Review Log**, and
Admin. Asst.	**(c) submit **Request for Reimbursement of Substitute** forms to the accounting office.
Consultant	19. Forward questionnaires and summary sheets to contracted personnel for aggregation of data, summary of answer keys, and compilation of comments.
Director/ Designated Instructional Services Consultant	20. When the content professor's review and the results of the test review by supervisors and teachers have been received in our office, the appropriate Instructional Services Consultant will meet with the Director and designated staff to compare and consolidate all necessary changes to test items and answer keys. All designated changes will be noted on the Instructional Services Consultant's copy of each test form.
Editor	21. All items that were modified should then go to an English editor for review. **(The editor should use a different colored pen/pencil than the one used by the Instructional Services Consultant.)**
Director	22. Review all changes.

Research Asst.

23. Make final changes to test items or keys

    (a) on the hard copy of the item record;

    (b) in the database file, the artwork file (if needed), and in the test statistics file; and

    (c) in the Pagemaker document.

Research Asst.

24. Once all changes have been made, reprint any necessary pages of the test.

Research Asst.

25. Assemble a final camera-ready copy of each form of the test, and attach a **correct** answer key to each form. (These keys will be printed from the statistics file.) Double check answer key with item records.

Director

26. Final copy to Instructional Services Consultant for last look.

Director

27. Clean forms for printing.

The test review process is now completed.

Admin. Asst.

Archives should contain

A file labeled "(Content Area) Test Review —(Year)". This file will contain

(a) a list of the supervisors, including names, addresses, phone numbers, regions, and the date materials were returned by each supervisor;

(b) a copy of all instructional materials (including questionnaires and summary form) and the cover letter mailed with these materials; and

(c) a copy of the Test Review Log;

(d) a copy of the aggregated results and compiled comments; and

(e) the xerox copy of each test form containing all final content and editing changes.

Two final copies of each form of the test and the test statistics summary should be placed in the appropriate (existing) files.

Manila folders containing the hard copy of item records for each form of the test should be returned to the appropriate (existing) files.

Test Redesign Procedures

Consultant 1. Match student scores on the first test of record and the field test (*.data.matched.form?).

Consultant 2. Determine number of students per semidecile class and semidecile class means on the first test of record and the field test (use program.sdclass.means) using matched data.

Consultant 3. Obtain item p-values within each semidecile class for first test of record and the field test (use program.itempval.bysdclas) using matched data.

Consultant 4. Determine the amount of discrepancy within each semidecile class (difference between first test of record and field test times the number of core items on the test).

Consultant 5. Select items that match needed improvement within each semidecile class.

Consultant 6. Evaluate item substitutions by:

(a) determining new semidecile class means on the field test (use program.sdclass.means.withsubs)

(b) plotting (in Lotus) statewide*statewide semidecile class means (for perfect agreement) and overlaying statewide*fieldtest semidecile class means.

	7.	When discrepancies between the field test of record and the field test semidecile classes have been accounted for, make item substitutions in electronic files:
Consultant		(a) Lotus files: testyear.wk1 and itemusag.wk1
Research Asst.		(b) Mac or 5520 files
Director/ Consultant	8.	Select variable items and determine placement of items within the test. The first item on each form of the test should be a variable and have a high p-value.
	9.	Repeat "Assembling the Tests" Procedures beginning with #5 (p. A-22).
	10.	Repeat "Test Review Procedures" (p. A-26).

Appendix B

Curriculum Survey Form

NORTH CAROLINA STATE DEPARTMENT OF PUBLIC INSTRUCTION
CURRICULUM SURVEY FOR END-OF-COURSE TESTING IN PHYSICS

(Office Use Only)
keypunch
column

Preliminary Information

1. School System (LEA)__________________ LEA Code______________ 1-3
(Please refer to the list of Code Numbers on the last page)

Please Circle One Number

	Yes	No	
2. In your school, are physics students assigned to classes on the basis of ability?	1	2	4

IMPORTANT: You may teach several physics classes. Do not try to answer the following questions with respect to all of your classes. Please choose only one class — THE FIRST PHYSICS CLASS YOU TEACH DURING THE DAY — as the class you have in mind as you answer the questions.

3. Which of the following best describes the ability level of most students in your first physics class? (Please circle one number)

advanced, honors	1	
above average	2	
average	3	5
below average	4	

4. Circle the grade in which most of the students in your first physics class are enrolled. (Please circle one number.)

ninth grade	1	
tenth grade	2	
eleventh grade	3	6
twelfth grade	4	

INSTRUCTIONS FOR EVALUATING THE COMPETENCY GOALS AND OBJECTIVES

On the following pages are listed competency goals and objectives for physics as given in the Teacher Handbook. Please evaluate the objectives by circling the appropriate number on the form. At the back of the questionnaire are several general questions plus space for making general comments. Use the ability level of your first physics class, as you described it above, as a reference in answering the questions.

Please use the envelope provided to return the completed forms through your school mail by **September 30, 1988.**

End-of-Course Testing
Physics
Page 2

DIRECTIONS FOR COMPLETING THE CURRICULUM REVIEW FORM FOR PHYSICS

Before completing the form, please study the rating scale. It is used to answer the question, **Do you teach this competency objective every year?** By circling one number, you provide two answers (as explained below):

Do you teach this competency objective every year?

If yes, circle either number (1) or (2); if no, circle either number (3), (4), (5), or (6). Circle only one number in each row.

If yes, do you consider this objective as

basic for all students in this class?	circle (1)
enrichment for only the top students in this class?	circle (2)

If no, why isn't this objective taught? (Circle the most appropriate number.)

It is too advanced for this class.	circle (3)
It is in the curriculum but is not covered due to time constraints	circle (4)
It is in the curriculum but is not essential.	circle (5)
Other	circle (6)

EXAMPLE: In the example below, the teacher reports that the objective is taught every year, and is basic for all students in the class.

CURRICULUM REVIEW FORM FOR END-OF-COURSE TESTING: PHYSICS COMPETENCY GOALS AND OBJECTIVES

	Do you teach this objective every year? (Please circle one of the six numbers)						
	Yes		No				
Competency Goal: 1. The learner will understand the science of physics and how it affects our lives.	Basic for Everyone	Enrichment Only	Too Advanced	Not Enough Time	Not Essential	Other (write in margin)	(Office Use Only) keypunch column
Competency Objectives:							
1.1 Know how to solve (physics) problems using basic algebra and trigonometry.	[✓ 1]	2	3	4	5	6	7
1.2 Know how to use measuring devices and scalar numbers (to solve physics problems).	[✓ 1]	2	3	4	5	6	8

CURRICULUM REVIEW FORM FOR END-OF-COURSE TESTING: PHYSICS COMPETENCY GOALS AND OBJECTIVES

	Do you teach this objective every year? (Please circle one of the six numbers)						
	Yes		No				
Competency Goal: 1. The learner will understand the science of physics and how it affects our lives. **Competency Objectives:**	Basic for Everyone	Enrichment Only	Too Advanced	Not Enough Time	Not Essential	Other (write in margin)	(Office Use Only) keypunch column
1.1 Know how to solve (physics) problems using basic algebra and trigonometry.	1	2	3	4	5	6	7
1.2 Know how to use measuring devices and scalar numbers (to solve physics problems).	1	2	3	4	5	6	8
1.3 Understand the elements of graph construction (used to solve physics problems).	1	2	3	4	5	6	9
1.4 Know of recent advances, career potentials, and current societal issues in physics.	1	2	3	4	5	6	10
Competency Goal: 2. The learner will understand basic mechanics.							
Competency Objectives:							
2.1 Know laws, mathematical expressions, and factors which represent and affect various types of motion.	1	2	3	4	5	6	11
2.2 Know how to analyze systems involving vector quantities and component forces.	1	2	3	4	5	6	12
2.3 Understand the behavior of gravitational forces.	1	2	3	4	5	6	13
2.4 Know how to quantify work, power, and mechanical energy.	1	2	3	4	5	6	14

CURRICULUM REVIEW FORM FOR END-OF-COURSE TESTING: PHYSICS COMPETENCY GOALS AND OBJECTIVES

	Do you teach this objective every year? (Please circle one of the six numbers)						
	Yes		No				
Competency Goal: 3. The learner will understand kinetic theory and have a general knowledge of properties of matter.	Basic for Everyone	Enrichment Only	Too Advanced	Not Enough Time	Not Essential	Other (write in margin)	(Office Use Only) keypunch column
Competency Objectives:							
3.1 Understand phases of matter in terms of the kinetic molecular theory.	1	2	3	4	5	6	15
3.2 Know how to use indexed information relating to physical constants.	1	2	3	4	5	6	16
Competency Goal: 4. The learner will understand elementary principles of thermodynamics.							
Competency Objectives:							
4.1 Understand factors associated with the characteristics of heat.	1	2	3	4	5	6	17
4.2 Know how to quantify the conservation of heat.	1	2	3	4	5	6	18
4.3 Know how to make determinations of the heat equivalent of work.	1	2	3	4	5	6	19

CURRICULUM REVIEW FORM FOR END-OF-COURSE TESTING: PHYSICS COMPETENCY GOALS AND OBJECTIVES

	Do you teach this objective every year? (Please circle one of the six numbers)						
	Yes		No				
Competency Goal: 5. The learner will understand basic elements of wave mechanics.	Basic for Everyone	Enrichment Only	Too Advanced	Not Enough Time	Not Essential	Other (write in margin)	(Office Use Only) keypunch column
Competency Objectives:							
5.1 Understand the general properties of wave phenomena.	1	2	3	4	5	6	20
5.2 Know how to investigate and describe sound in a quantified manner.	1	2	3	4	5	6	21
5.3 Know how to investigate and describe light in a quantified manner.	1	2	3	4	5	6	22
Competency Goal: 6. The learner will understand basic principles of electricity and magnetism.							
Competency Objectives:							
6.1 Know how to identify and quantify certain electrostatic phenomena.	1	2	3	4	5	6	23
6.2 Understand basic quantities and components associated with direct current circuits.	1	2	3	4	5	6	24
6.3 Understand basic terms and phenomena associated with magnetism.	1	2	3	4	5	6	25
6.4 Understand basic descriptive material pertaining to alternating current circuits.	1	2	3	4	5	6	26

CURRICULUM REVIEW FORM FOR END-OF-COURSE TESTING: PHYSICS COMPETENCY GOALS AND OBJECTIVES

	Do you teach this objective every year? (Please circle one of the six numbers)						
	Yes		No				
Competency Goal: 7. The learner will understand some fundamental concepts of particle physics.	Basic for Everyone	Enrichment Only	Too Advanced	Not Enough Time	Not Essential	Other (write in margin)	(Office Use Only) keypunch column
Competency Objectives:							
7.1 Know about atomic theory and its development.	1	2	3	4	5	6	27
7.2 Know about major aspects of quantum theory.	1	2	3	4	5	6	28
7.3 Understand the properties and functions of major atomic particles.	1	2	3	4	5	6	29
7.4 Know about practical applications of nuclear energy.	1	2	3	4	5	6	30

Appendix C

Sample Item Specification Form

Item Specifications Sheet

Curriculum Objective:	
Difficulty Level: 1 = easy 2 = medium 3 = hard	Level of thinking skills: 1 = lower 2 = higher
Artwork Required: 1 = yes (If yes, please attach) 2 = no	Curriculum Source:
Item Writer Number:	

Physics Test Item *(final draft)*

Correct Answer ______

Edit ______ ______ ______

Did You...

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

Appendix D

Script for How to Write Multiple-Choice Achievement Test Items,three Work Packets, and a sample Item Specification

Script for

How to Write Multiple-Choice Achievement Test Items

Available in:

- Sound/slide for large group presentation
- Videotape for individual or small group presentation

Work Packets (A, B, and C) are also available. These can be modified to reflect special item-writing topics: e.g., content areas, skills areas.

NC TESTS

For further information, call your Regional Research Consultant
(videotapes and work packets are available there,
and may be copied by you), or contact:

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332
(919) 733-0809

Published 1988 (revised)

(Music with voice over)

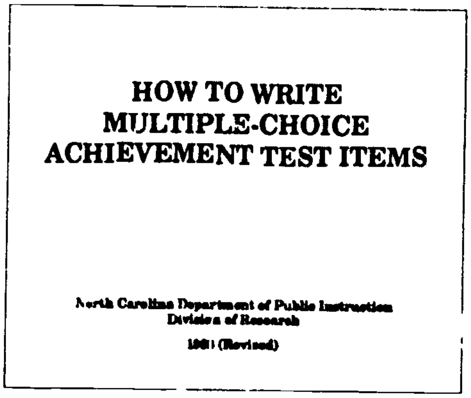
(Young man) I really didn't know I was smart enough for college until I took the SAT.

(Mature woman) These diagnostic test results tell me something I had not realized before. The students' errors all tend to be of one type.

(Female with trained speaking voice) The CAT achievement scores assure us that our appropriations for education are effective.

(Man with business-like voice) These employment test results confirm my belief that you are the sort of person we want to employ.

(Mature woman) In my senior year they refocused me: more science, more math. But if I hadn't taken the test, would they have gone to bat for me? (Mary Futrell, president of NEA)



HOW TO WRITE
MULTIPLE-CHOICE
ACHIEVEMENT TEST ITEMS

North Carolina Department of Public Instruction
Division of Research

1981 (Revised)

Tests have been used since ancient times to develop systematic knowledge about people: what they know, what they can do, what they think. One of the common forms of testing today is the multiple-choice pencil and paper test. This type of test came into use before the turn of this century. Through the years, it has been developed to the point of a high technology.

Item-writing:

The heart of test development

ITEM = STEM + FOILS

In this session, we plan to explore one critical part of multiple-choice test development: that of item-writing. An item comprises a question and a set of response choices. In test development, the question is called the stem; the response choices are called the foils. After the items are written, they will undergo editing for grammar and syntax, curricular relevance, and form. Then they will be field tested and checked against other criteria, mainly statistical. The items that survive these procedures become the raw materials from which a test will be constructed. Our concern here, though, is only with item-writing, and specifically the writing of multiple-choice items.

Part 1. Test Content: What kind of test?

Part 2. Item Content: What kind of items?

Part 3. Item Format: What makes a good item?

Part 4. Item Context: What makes a good item bad?

The subject of item-writing will be discussed in four parts:

- The first part deals with test content, and how that content is determined by curricular validity, instructional validity, and examinee validity.
- The second part is directed at item content. In this part, the logical categories into which items can be classified are identified and described. Also, other more specific questions about item content are discussed.
- The third part describes the characteristic form of a good item—what it is and what it is not.
- The fourth part deals with item context. In this final part, the discussion covers questions of bias, stereotypes, and fair representation.

Part 1 Test Content

The first of the four topics, test content, will now be discussed in more detail. The discussion will begin with the concept of CURRICULAR VALIDITY.

Part 1 Test Content

A Curricular Validity

A curriculum refers to a course of study. A curriculum description consists of a set of statements describing a course of study. Usually, an important part of a curriculum description is a list of goals and objectives. The goals and objectives normally refer to educational outcomes for students.

Test items are intended to reflect the students' degree of success in attaining the educational outcomes described by the curricular goals and objectives. Items that succeed in measuring the educational outcomes are said to have CURRICULAR VALIDITY: that is, they measure what they were meant to measure; namely, whether the students have learned the curriculum.

Part 1 Test Content

A Curricular Validity

- **Verify that objectives are clearly stated for purposes of item-writing.**

Obviously, the curricular validity will depend in part on how well the educational goals and objectives are defined. Rewriting weak objectives is not part of the item-writer's task. Before accepting the task of item-writing, however, the item-writer should make sure that the goals and objectives are intelligible. If they are not, clarification should be requested.

Part 1 Test Content

A Curricular Validity

- **Verify that objectives are clearly stated for purposes of item-writing.**
- **Write items that directly represent the educational objectives.**

Items that do not reflect the curricular objective for which they are intended cannot be used to support that objective. This means that the objective will be short one or more of its planned items. Item-writers should write for the intended objective. Then they should check their final work against the objective to make sure its content has not strayed from its original intention.

Part 1 Test Content

B Instructional Validity

The next topic is that of INSTRUCTIONAL VALIDITY.

An item has instructional validity if it measures the educational outcomes of instruction. It is possible for a test to have **curricular** validity and not have **instructional** validity; which is another way of saying that the official curriculum may not be what is taught in the classroom.

Part 1 Test Content

B Instructional Validity

- **Write items that represent what is being taught in the classroom.**

Sometimes tests are developed to see what is being learned without considering what is being taught. The subsequent test results are ambiguous. Did the students fail an item because it was too difficult, or because it was not taught? And what does one do with the students' test scores in that case? It is obviously unfair to record scores where they may be used as evidence of low academic ability when in fact they reflect only an inadequate instructional program.

In general, test developers try to insure that their tests will have both curricular and instructional validity; that is, that the tests will measure the joint outcome of what **should** be and what **is** being taught in the classroom.

Part 1 Test Content

B Instructional Validity

- **Write items that represent what is being taught in the classroom.**
- **Stick with mainstream of instruction.**

Curricular and instructional validity depend strongly on work done by test developers before item-writing is scheduled. But item-writers still play an important part in establishing validity. Item-writers have an effect on validity when they select one approach to creating an item from many possible approaches. In making these selections, item-writers should ask themselves the questions, "Does this approach I am about to take deal with the mainstream of the curriculum? Is the item something that one would expect all teachers to cover? Or is it an interesting but obscure and unimportant side issue that only a few teachers will choose to teach?" Item-writers should bear in mind that space on a test form is expensive. Items, to be effective, must follow the mainstream of instruction.

Part 1 Test Content

B Instructional Validity

- **Write items that represent what is being taught in the classroom.**
- **Stick with mainstream of instruction.**
- **Avoid tricky questions.**

Another aspect of this problem is the introduction of tricky or irrelevant cues. Items containing such cues introduce an unwanted degree of intelligence testing into what should be achievement testing. The intent is to measure what has been learned—not how much native intelligence the student may possess. Also, tricky questions usually do not survive field testing, and therefore represent a substantial waste of money and other resources. Item-writers should be straightforward and stay with the important issues.

Part 1 Test Content

C Examinee Validity

In addition to curricular and instructional validity, the items should have EXAMINEE VALIDITY. This means simply that the items should be suitable for use with the target students.

Part 1 Test Content

C Examinee Validity

- Use suitable language and illustrations.

First of all, the language and illustrations employed in writing the items should be suitable for the students who will take the test. Language difficulty, or readability, as it is called, is one aspect of language suitability. Teachers have a good grasp of readability for students at the grade level they teach. That is one reason they are often chosen as item-writers. Item editors, however, can check the final readability level through the use of desktop-computer programs. For example, one program can be used to calculate readability as measured by seven commonly-used formulas.

Part 1 Test Content

C Examinee Validity

- Use suitable language and illustrations.

- Keep language simple.

Unless the test is a test of language, the readability of items should not be more difficult than the language level of most of the students who will take the test. This restriction does not apply, of course, to technical terms that are an integral part of the subject being tested.

Part 1 Test Content

C Examinee Validity

- Use suitable language and illustrations

- Keep language simple.

- Avoid culture-specific references.

Also, language suitability requires that no culture-specific items be written unless culture is the topic being tested. Item-writers should avoid unusual terminology, references to obscure localities, or customs experienced by only a few examinees. This restriction applies not only to the stem and the correct answer, but also to the incorrect foils.

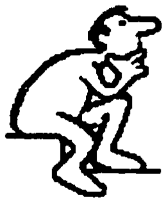
Part 1 Test Content

C Examinee Validity

- **Use suitable language and illustrations.**
- **Keep language simple.**
- **Avoid culture-specific references.**
- **Practice fair representation in sex and race.**

Another topic related to language suitability is fair representation. The premise here is that a test taken by boys and girls, for example, should not contain only girls' names or boys' names, but some of both. Illustrations showing sports being played should include both sexes. Students of minority races should be represented proportionately to their numbers in the student population. The absence of fair representation is sometimes referred to as linguistic or pictorial bias. Good practice requires that fair representation be taken into account in the interests of face validity and general test acceptance. Fair representation will be discussed again in Part 4.

Part 2 Item Content



So far, we have discussed curricular, instructional, and examinee validity. We now turn to Part 2, which concerns item content.

In developing tests, test developers realize that thousands of items can be written on any curricular subject. They must decide on some procedure that will result in items that are representative of the curriculum, but are limited in number. That is the concern of item content.

Part 2 Item Content

What types of items to write:

A different question for content subjects and skills subjects

The problem is partially solved by asking item-writers to write a certain number of items for each curricular objective. But what type of items should item-writers write? The question can be examined in two parts. One part concerns **content** subjects such as biology, history, and chemistry. The other part concerns **skills** subjects such as mathematics, language, and reading. Defining item types for content subjects will be discussed first.

Part 2 Item Content

Four Ways of Knowing:

A Who, which, where, and when?

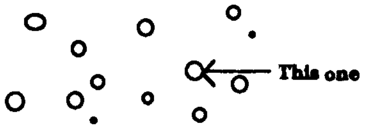
The *denotive* category

What is needed for content subjects is a logical system to suggest what can be known about a person, topic, event, thing or idea, since we intend to measure academic knowledge. One such system, which uses four logical categories, is described in the following section.

The first category covers the questions of who, which, where, and when. This category is denotive. It simply identifies a thing.

"I am John."
Which John? Answer:
"The John who signed the Magna Carta."
Who are you? Answer:
"I'm the man who discovered America in 1492."
Who is that? Answer:
"Christopher Columbus."

Part 2 Item Content



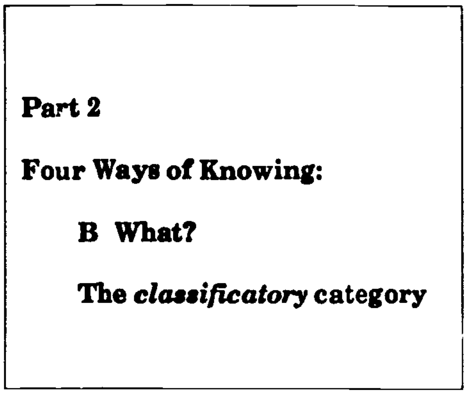
The denotive category of knowledge "points"

In grasping the meaning of this category, it is important to realize that the intentions are only to point to a thing, to separate it from others, to discriminate it from the whole. Sometimes pointing is enough. Which man? "The one standing over there." If doubt still exists, we can tell what the object of discussion is made of, what its physical composition is. Which one of those standing over there? "The one in the blue suit, with the mustache and the long nose; the short one."

Examples can be taken from any context. For instance:

> Which form of business organization allowed directors of one company to serve as directors of several other companies?
>
> Answer: "Interlocking directorates."
>
> Which type of symbiosis is beneficial to both species?
>
> Answer: "Mutualism."
>
> Or:
>
> When is symbiosis called mutualism?
>
> Answer: "When it is beneficial to both species."

Note that one need not know what symbiosis is to answer these questions. The intention is not to understand the thing, person, concept, or idea, but simply to identify it, to place some sign on it that enables the examiner to say, "Yes, I know which one you mean."



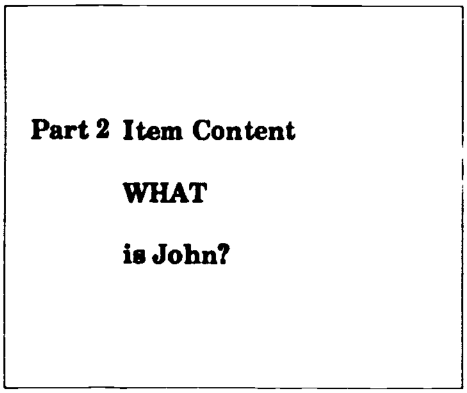
Part 2

Four Ways of Knowing:

B What?

The *classificatory* category

The second category answers the question, "What?" It is **classificatory**. It tells **what** a thing is. This category sounds a little like the first one, but its meaning is different. Here the idea is to place the thing in relation to others of its kind . . . to make a connection.

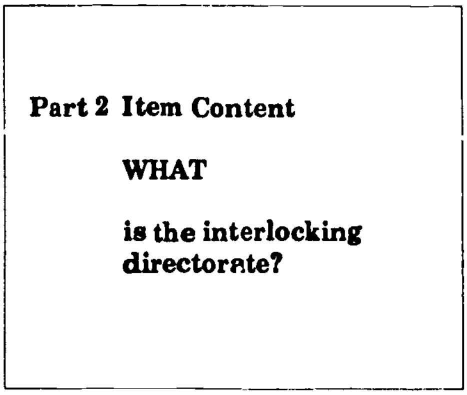


Part 2 Item Content

WHAT

is John?

John is not only the man standing there, the one who just came in. He is married; rich; a lawyer; a Yuppie; a pathological thief. As you can see, we soon begin to know quite a bit about the man.

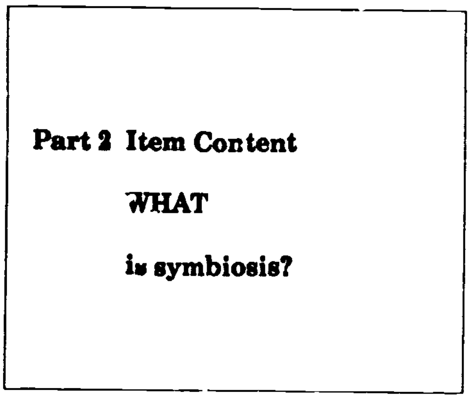


Part 2 Item Content

WHAT

is the interlocking directorate?

What was the interlocking directorate? That business organization was a means of consolidating the activities of separate companies. Other means were holding companies and trusts. That is what the interlocking directorate was.

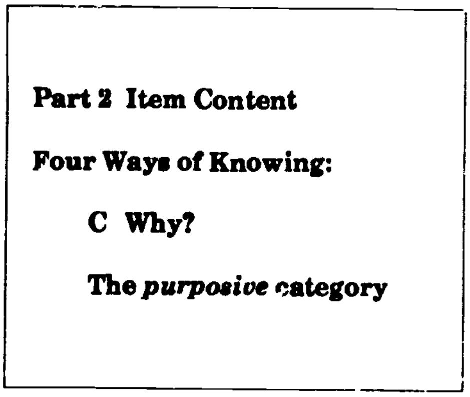


Part 2 Item Content

WHAT

is symbiosis?

What is symbiosis? It is an intimate relationship between two different species. It contrasts with simple coexistence and total isolation.



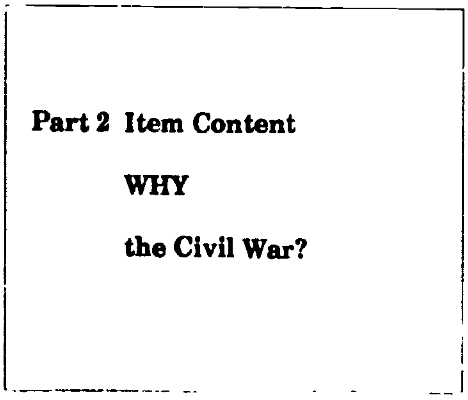
Part 2 Item Content

Four Ways of Knowing:

C Why?

The *purposive* category

The third category is the purpose, the "why," of things.



Part 2 Item Content

WHY

the Civil War?

In brief, the purpose of the Civil War was to settle on the battlefield what could not be settled across the table. Economic interests were a key factor, and they could not be resolved peacefully.

Part 2 Item Content

WHY

the interlocking directorates?

The why of interlocking directorates had to do with the way industry had grown up in the United States: essentially as the competitive efforts of small entrepreneurs. This led to intense competition and duplicate efforts in manufacturing, merchandising, distributing, and transportation. Interlocking directorates reduced the competition and duplication. Thus the "why" of the interlocking directorate was to bring some order into a chaotic entreprenurial environment.

Part 2 Item Content

WHY

symbiosis?

The why of symbiosis has to do with the sensitivity of life forms to environmental stimuli, the tendency of life forms to expand their territory until they contact other life forms, the competition for territory, and the nature of the fortuitous encounters that occur during this process. We have symbiosis because that is one of the possible adjustments occurring during competition for territory.

Characteristic of the "why" question is the tendency to go outside the "thing" to explain it, to relate it to other things not of its own kind. The explanation calls for a larger scheme of discussion than the denotive or classificatory categories.

Part 2 Item Content

Four Ways of Knowing:

D How?

The *functional* category

The fourth category tells how the thing works, how it "lives and moves and has its being," so to speak.

Part 2 Item Content

HOW

the Civil War worked.

The Civil War worked this way: a state of war was declared; finances were arranged; soldiers snd sailors were recruited and trained; strategies and tactics were devised; battles were fought on land and sea; successes and failures were experienced; and surrender brought it to a close. That is the way most wars work.

Part 2 Item Content

HOW

the interlocking directorate worked.

The way interlocking directorates worked was that the same people were appointed as directors of the companies that wished to reduce competition and duplication of effort. In their various roles, the directors could decide that company A would specialize in one aspect of the business and companies B, C, D, and E would specialize in other aspects. Also, the directors could decide that all companies would share in, for example, a common purchasing or advertising activity, thus reducing costs.

Part 2 Item Content

HOW

symbiosis works.

In another example, symbiosis works in various ways, depending upon the interacting pair of species. For instance, some common forest mushrooms absorb minerals from the soil and pass them on to plants, allowing the plants to grow much faster than they would in the absence of the mushrooms. It is not known whether the mushrooms benefit from their association with the plants. This is how symbiosis works.

Experience has shown that item-writers, if left uninstructed, are very likely to write too many Type 1, denotive, items. It is important to have a proper mixture of item types when content is the main material to be learned.

Part 2 Item Content

THINKING LEVELS of the four types of knowledge.

Consider these four categories in the context of their intellectual complexity and the rather ambiguous term "thinking levels." It is usually true that the Type 1 **who-where-when** question represents the lowest of thinking levels. The Type 2 **what** category is higher. And the Type 3 **why** and Type 4 **how** categories vie for top spot, depending upon their specific content. For instance, the "how" of a black hole in astronomy may be more difficult to understand than the "why" of the black hole.

Content item-writers, after writing an item, should try to categorize it according to these four categories. If they are unable to do so, that is a good sign the item is poorly written. Watch for "Why" stems that have "How" foils, and similar mismatches.

Part 2 Item Content

Thinking levels in skills subjects:

Do item-writers need them?

We now turn to the problem of writing items for the **skills** subjects. The skills most often tested by multiple-choice questions concern the manipulation of symbols—either verbal or mathematical in form. The manipulation of symbols is addressed in language and mathematics courses, and in reading. These subjects follow formal rules and can be described in detail in the objectives. Add two-digit numbers. Solve a quadratic equation. Write a complete sentence. Know the correct use of end punctuation. Items written for most skills objectives will not have specifications for thinking levels. It is far-fetched to try to classify 2 + 2 as one thinking level and the addition of any other two one-digit numbers as a different thinking level. The same is true of writing items to test mastery of end punctuation.

Part 2 Item Content

Thinking levels vs. item difficulty:

Two Different Concepts

The so-called thinking level of an item should not be confused with item difficulty, however. In principle, thinking level and item difficulty are entirely independent of each other. The item-writer may or may not need to know whether to write a difficult or an easy item. Even with the addition of two one-digit numbers, some combinations are more difficult than others. A few people carry an uneasiness concerning 8 + 7 right on into adulthood. Item-writers will usually be given some instructions about how many easy, average, and difficult items to write. This gives the item-writer something to aim at, although the true difficulty of an item will not be known until it has been field-tested.

Part 2 Item Content

Beyond content subjects and skills subjects:

psychomotor skills

A third category of subjects is not included in this discussion cf the writing of multiple-choice items. The third catcgory deals with psychomotor skills, such as the act of riding a bicycle or using a typewriter. Multiple-choice items may be useful in finding out if the examinee knows how a bicycle is ridden, but they cannot be used to determine whether the examinee can actually ride a bicycle. Psychomotor skills are best tested by activity samples, not multiple-choice items.

Part 2 Item Content

Specific Concerns:

A Use artwork only as required.

Having considered these basic item content problems, we now turn to some more specific concerns of item content.

First, about the use of pictures, graphs, charts, and the like. No hard and fast rule exists about the number of illustrations to include in a test. The material usually dictates the number. Occasionally—in map reading, for example—every question may require an illustration. On aesthetic and logistic grounds, however, the number of items requiring illustrations should not exceed 25% of the number of items on the test. This factor may be covered in the item specifications, in which case the item-writer need only follow instructions.

Part 2 Item Content

Specific Concerns:

A Use artwork only as required.

B Do not quote from textbooks.

Second, item text should not be taken word for word from a textbook. This prevents rote memory from being a direct aid in answering the question.

Part 2 Item Content

Specific Concerns:

A Use artwork only as required.

B Do not quote from textbooks.

C Avoid duplicate questions.

A third consideration is the relationship among items. In writing items, useful information about one item should not be contained in another item. This rule may be violated in several ways, of which two of the most common will be described. In the first way, one question is only a reworded duplicate of another question. For example, What is important to rising to power in Russia? And, What associations should one have to obtain office in the USSR?

The second way can be called the Janus question. This two-faced question is first asked in one direction (What do Mayors do? Run a town.) and then in the other (Who runs a town? A Mayor.).

This problem runs contrary to the principle of "local independence," which is a condition usually assumed to be present in test construction. Broadly speaking, local independence means that every item is independent of every other item in the test. Obviously, that is not the case if two of the items cover the same content in the same way. That violates "local independence."

Part 2 Item Content

Specific Concerns:

A Use artwork only as required.

B Do not quote from textbooks.

C Avoid duplicate questions.

D Make every word count.

A fourth consideration is content necessity. Everything in the item should play its part and should be necessary to an understanding of the item. In published tests, it is not unusual to find instances where text or illustrations are superfluous to meaning. In tests of reading skills, too, the items can sometimes be answered without reference to the reading text. In the matter of illustrations, some artwork may be entirely irrelevant. Occasionally, one finds a set of foils out of which one can pick the right answer without even reading the question.

Waiting...

We now take a break from the presentation to allow you to work with Packet A of the supplementary materials. In those materials, you will find some item content considerations that are specific to your item-writing needs. After you have completed the exercises, return to the presentation. If you are watching this program in videotape, please place the VCR on STOP.

If you are watching this program in slide format, please stop the audio cassette and change the slide tray.

Part 3 Item Format—Stems

One **Write a complete sentence; make it a complete statement of the question.**

We turn now from item content to item format, which is the subject of Part 3.

In item format, the form of the stem is the first consideration. The stem should be a complete question. It should contain all of the information that an informed examinee would need to answer it, even without the foils. Grammatically, the stem should be a complete sentence. Exceptions to these guidelines are sometimes needed, but should be stated explicitly in the item specifications. Otherwise, the complete sentence guideline for stems should be followed. Consider this example. *(Music)*

Problem: stem not a complete sentence

The Apache Indians lived in the

A **Northeast.**

B **Northwest.**

C **Southeast.**

D **Southwest.**

Solution: stem a complete sentence

The Apache Indians lived in which region?

A **Northeast**

B **Northwest**

C **Southeast**

D **Southwest**

Part 3 Item Format—Stems

Two **Focus directly on the objective.**

Besides being a complete statement of the question, the stem should focus directly on the instructional objective. Often, in an attempt to make a stem more difficult, the item-writer will write indirectly about the topic, tackling some issue that has only a marginal relationship to the instructional objective. This is playing a game with the examinee. Games are more suitable for an intelligence test than for an achievement test. Consider this example. *(Music)*

Problem: item not responsive to objective

State causes and results of the French and Indian War.

Who was a young Virginia militia officer during the French and Indian War?

A Sieur de La Salle

B William Pitt

C Louis Joliet

D George Washington

Solution: item responsive to objective

The French and Indian War was made more likely by which of these events?

A struggle between French and Indians for trading rights

B efforts of French to link their northern and southern territories

C massacre of settlers at Fort Ticonderoga and Crown Point

D failure of the French to live up to their financial obligations

Part 3 Item Format—Stems

Three Use proper grammar and syntax.

Naturally, the item writer should use proper grammar and syntax. Unfortunately, this reminder is necessary from time to time. The item-writer should be wary of misplaced modifiers, dangling participles, and other threats to intelligibility. These are more likely to appear in the stem than in the foils. Consider this example. *(Music)*

Problem: organs and tissues caused by diseases?

Changes in organs and tissues caused by diseases are studied by which of these scientists?

A behaviorist

B herpetologist

C pathologist

D pharmacist

Solution: "caused by diseases" follows "changes"

Change:

Changes in organs and tissues caused by diseases are studied by which of these scientists?

To:

Changes caused by diseases in organs and tissues are studied by which of these scientists?

Part 3 Item Format—Stems

***Four* Avoid negatives in writing items.**

The item-writer should not use negatives or double negatives in writing items. For example, do not construct stems of the form, "Which of these is not a lily?" That type of item introduces non-achievement factors into the process of testing. The same is true of a stem that, for example, reads, "Why did Smith not fail to yield to the other car?" Consider this example. *(Music)*

Problem: use of "not" in stem

Of the following functions, which is not performed by the blood?

A digesting food

B transporting food

C regulating body temperature

D resisting disease

Solution: "not" removed from stem

Of the following functions, which is performed by the blood?

A digesting food

B transporting food

C excreting wastes

D bathing tissue

Writing the foils is the most difficult part of item-writing. A list of do's and don't's about writing foils will be given next.

Part 3 Item Format—Foils

Five **Write four foils.**

Four foils should be written for each item unless the item specifications say otherwise. Consider this example. *(Music)*

Problem: only three foils

Pavlov is best known for his conditioning experiments with what animals?

A birds

B cats

C dogs

Solution: four foils

Pavlov is best known for his conditioning experiments with what animals?

A birds

B cats

C dogs

D apes

Part 3 Item Format—Foils

Six **Locate correct answer equally among foils.**

The right answer should be placed in each of the four choice positions an equal number of times, in random order from one item to the next. Consider this situation. *(Music)*

Problem: uneven distribution by foil

Tally of Correct Response Choices
60-Item Test

A	卌 卌 卌 卌 ||	= 22
B	卌 卌 卌	= 15
C	卌 卌 卌 |||	= 18
D	卌	= 5
		60

Solution: even distribution by foil

Tally of Correct Response Choices
60-Item Test

A	卌 卌 ||||	= 14
B	卌 卌 卌	= 15
C	卌 卌 卌 |	= 16
D	卌 卌 卌	= 15
		60

Sometimes, the item specifications give the location of the correct answer; that is, whether it is to be response choice A, B, C, or D. If they do not, the item-writer can save time by putting the locations on the specification sheets (to be discussed later) before starting to write items. Having done this, the item-writer will not be interrupted in the creative process by the necessity of deciding whether the correct foil is to be choice A, B, C, or D.

Part 3 Item Format—Foils

Seven **Make grammatical structure of foils the same.**

The grammatical structure of the four foils should be similar. Parallel construction is almost always desirable. Parallel construction is not difficult to achieve. Yet it is a common failing of untrained writers, however intelligent and informed they may otherwise be. Consider this example. *(Music)*

***Problem:* varied grammatical structure**

The 1840 potato famine in Ireland resulted from which of these?

A spraying of potatoes

B Only corn and wheat were planted.

C Emigration to other countries increased.

D Ireland engaged in manufacturing.

***Solution:* similar grammatical structure**

Irish farmers responded to the 1840 potato famine in which manner?

A sprayed their crops

B planted other crops

C emigrated elsewhere

D went into industry

Part 3 Item Format—Foils

Eight **Write foils of equal length.**

If practical, the length of the foils should be roughly the same. Consider this example. *(Music)*

Problem: foils unequal length

Electricity is produced on a large scale by which means?

A burning wood

B running water over a dam

C capturing and storing underground steam

D fusion

Solution: foils nearer the same length

Electricity is produced on a large scale by which means?

A burning wood

B falling water

C geothermal steam

D nuclear fusion

Part 3 Item Format—Foils

Nine **Correct answer should be lo~gest answer no more than 25% of the time.**

The correct answer should be the longest answer no more than one-fourth of the time. Item foils frequently fail this criterion. In fact, books on "How to Take a Test Successfully" feature this as one of the main points: **When in doubt, check the longest foil.** Consider this example. *(Music)*

Problem: longest foil correct too often

Correct foil is longest foil.

Yes	**60%**
No	**40%**

Solutio. : longest foil correct propor·ionately

Correct foil is longest foil.

Yes	**25%**
No	**75%**

The item-writer should make periodic checks of foil lengths every 20 to 30 items. If the correct foil is the longest foil more than 25% of the time, the item-writer should make adjustments in the length of the foils. Conversely, the correct answer should not systematically be the shortest foil. This is rarely a problem, however.

Part 3 Item Format—Foils

Ten **Use similar data and ideas in foils.**

Data and ideas used in the foils should be similar, so that one foil cannot be preferred to another solely on the basis of general content. Consider this example. *(Music)*

Problem: "igneous" stands out

When lava erupts from a volcano, what kind of rock is formed?

A igneous

B asphalt

C liquid

D fossils

Solution: "igneous" appears similar

When lava erupts from a volcano, what kind of rock is formed?

A igneous

B metamorphic

C sedimentary

D granodoritic

Part 3 Item Format—Foi!s

Eleven **Use same context for foils.**

The context of the foils should be the same. For example, if the subject is tropical plants, the incorrect foils should not give examples of plants found in a temperate zone. In another instance, if the item stem asks for an advantage of some sort, all foils should present potential advantages; and so on. Consider this example. *(Music)*

Problem: contexts differ—volcanoes, wind

Existing rock changes into metamorphic rock under which of these conditions?

A pressure

B wind

C weather patterns

D volcanoes

Solution: all processes in similar context

Existing rock changes into metamorphic rock under which of these conditions?

A compression

B erosion

C glaciation

D dehydration

Part 3 Item Format—Foils

Twelve **Make the foils explicit; no fuzziness.**

The foils should be explicit. Examinees will ignore foils that are vague or seem to fall short of being a reasonable answer. Each incorrect foil should draw its share of incorrect answers. A foil should not be a wrong answer on the basis of some half-hidden contradiction, however. Tricky foils do not make good item response choices. Consider this example. *(Music)*

Problem: foils fuzzy; not credible

For late nineteenth-century trusts, what was a major goal?

A creating better relations among new immigrants

B improving science

C controlling the production of a single commodity

D not allowing debtors to default

Solution: choices clear; credible

For late nineteenth-century trusts, what was a major goal?

A getting government contracts for goods

B producing goods for wholesale export

C controlling production of goods

D providing regional goods and services

Part 3 Item Format—Foils

Thirteen If correct answer falls on a continuum, write incorrect foils on same continuum.

If the correct foil deals with a dimension or a continuum, then the other foils should also. For example, if the answer is 6, other possible answers may be expected to be numbers. Consider this example. *(Music)*

Problem: mixture of feet and inches

The width of a rectangle is 4 inches less than half its length. If the perimeter of the rectangle is 64 inches, what is its width?

A 8 inches

B 2 feet

C 20 inches

D 2 feet, 2 inches

Solution: all inches

The width of a rectangle is 4 inches less than half its length. If the perimeter of the rectangle is 64 inches, what is its width?

A 8 inches

B 20 inches

C 24 inches

D 26 inches

Part 3 Item Format—Foils

Fourteen **Arrange continuous foils in logical order.**

If the foils fall along a continuum or a dimension, arrange them in logical sequence. If this is not practical because of some other consideration, randomize their presentation. Consider this example. *(Music)*

Problem: dates not in numerical order

Which tariff is also known as the Trade Expansion Act?

A Tariff of 1962

B Tariff of 1828

C Tariff of 1909

D Tariff of 1930

Solution: dates in numerical order

Which tariff is also known as the Trade Expansion Act?

A Tariff of 1828

B Tariff of 1909

C Tariff of 1930

D Tariff of 1962

Part 3 Item Format—Foils

Fifteen **If foils are not all continuous, make them all discrete.**

If a dimensional sequence does not seem possiole, make each foil different in its own plausible way. Avoid making two of the foils dimensional and the other two something entirely different. Here is an example. *(Music)*

Problem: mixture of dates and events

Germany gained control of the Sudetenland at what time?

A 1935

B 1938

C after the Nazi-Soviet Pact

D with the signing of the Locarno Pact

Solution: all events

Germany gained control of the Sudetenland at which period in history?

A before the Anglo-German naval agreement

B following the Munich Pact

C after the Nazi-Soviet Pact

D with the signing of the Locarno Pact

Part 3 **Item Format—Foils**

Sixteen **In some cases, write incorrect foils as diagnostic tools.**

Incorrect foils can be written as diagnostic tools. In this method, each incorrect foil is written to signal some specific misunderstanding the examinee may have about the subject. This technique requires that wrong answers be scored individually for diagnostic purposes. This technique is now mostly restricted to skill areas such as mathematics. Consider this example. *(Music)*

Problems: no systemic errors

Solve: 248 + 123 =

A 372 (wrong addition)

B 371

C 258 (wrong addition)

D 283 (wrong addition)

Solution: systemic errors

Solve: 248 + 123 =

A 471 (carried incorrectly)

B 375 (added incorrectly)

C 371

D 361 (failed to carry)

Part 3 Item Format—Foils

Seventeen **Avoid echo or clang associations.**

Echo or clang associations between the stem and the correct foil should be avoided. If the stem and the correct foil share the same technical term, or if one term in the stem is similar to one in the correct foil, then the correct answer may be inadvertently disclosed. Here is an example. *(Music)*

Problem: echo association

↓

The adrenal medulla secretes which of the following?

A androgens

↓

B adrenalin

C prostaglandins

D steroids

Solution: no echo association

The adrenal medulla mediates which reaction?

A metablic rate

B lactation

C stress

D urine secretion

Part 3 Item Format—Foils

Eighteen **Use same amount of descriptive detail in all foils.**

The amount of descriptive detail and logical complexity should be the same in the correct foil and in the incorrect foils. Consider this example. *(Music)*

Problem: differing amounts of description

Which product was invented for the space program?

A disposable clothing

B freeze-dried foods and beverages

C radar

D canned foods and beverages

Solution: same amount of description

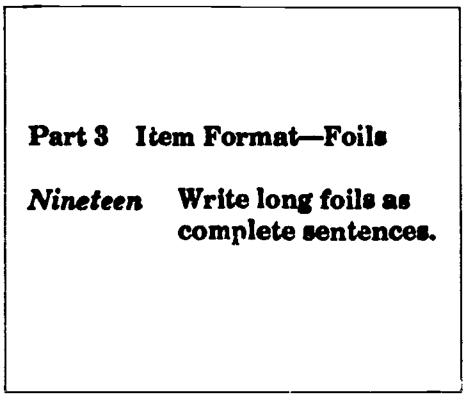
Which product was invented for the space program?

A disposable clothing

B freeze-dried food

C recycled water

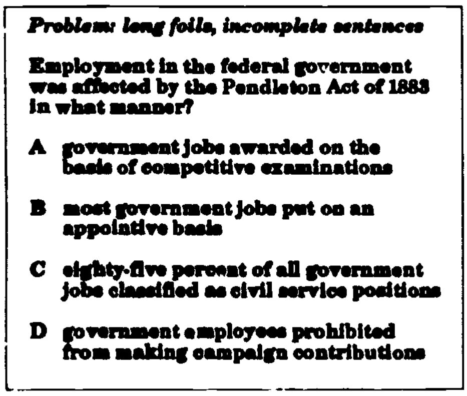
D liquid oxygen



Part 3 Item Format—Foils

Nineteen Write long foils as complete sentences.

When the foils are long, they may cause less confusion if they are written in the form of complete sentences. Consider this example. *(Music)*



Problem: long foils, incomplete sentences

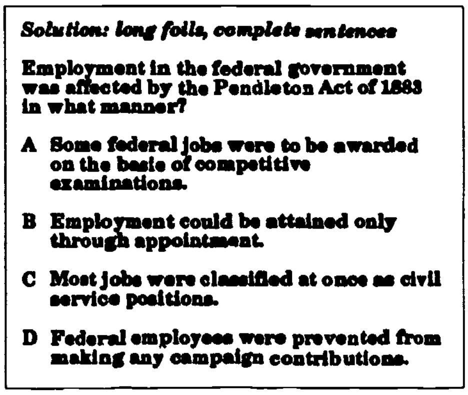
Employment in the federal government was affected by the Pendleton Act of 1883 in what manner?

A government jobs awarded on the basis of competitive examinations

B most government jobs put on an appointive basis

C eighty-five percent of all government jobs classified as civil service positions

D government employees prohibited from making campaign contributions



Solution: long foils, complete sentences

Employment in the federal government was affected by the Pendleton Act of 1883 in what manner?

A Some federal jobs were to be awarded on the basis of competitive examinations.

B Employment could be attained only through appointment.

C Most jobs were classified at once as civil service positions.

D Federal employees were prevented from making any campaign contributions.

Part 3 Item Format—Foils

Twenty Write one, and only one, correct answer.

Make sure there is one, and only one, correct answer. Try to avoid correct foils that are only the "best" answer, leaving some incorrect foils to be partially correct. Here is an example. *(Music)*

Problem: two possible answers

Which is the same as 1 + 1 + 1?

A 6 + 2 (right total)

B 8 × 4

C 3 × 1 (correct form)

D 7 – 5

Solution: only one possible answer

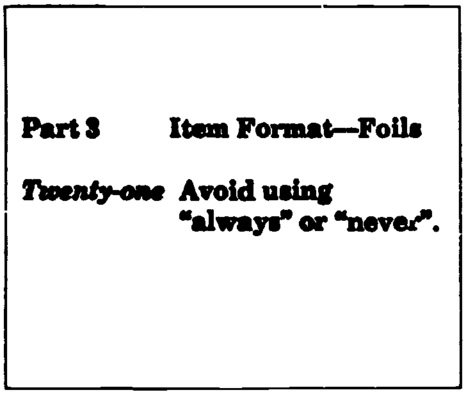
Which is the same as 1 + 1 + 1?

A 6 + 3

B 8 × 4

C 3 × 1

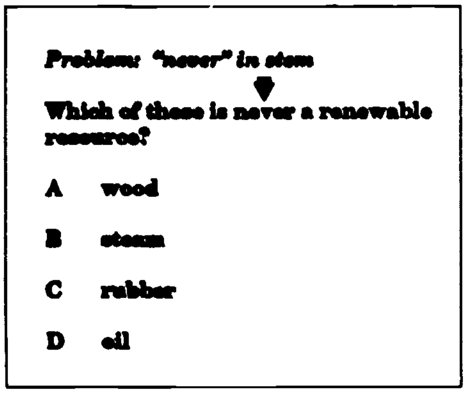
D 7 – 5



Part 3 **Item Format—Foils**

Twenty-one **Avoid using "always" or "never".**

"Always" and "never" should be avoided in writing foils, because examinees may be able to construct some plausible instance that is an exception to that rule. The same is true of other wording that asks the examinee to identify the wrong answer. Consider this example. *(Music)*



Problem: "never" in stem

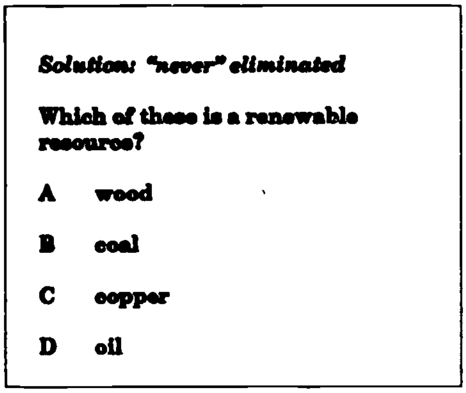
Which of these is never a renewable resource?

A wood

B steam

C rubber

D oil



Solution: "never" eliminated

Which of these is a renewable resource?

A wood

B coal

C copper

D oil

Part 3 Item Format—Foils

Twenty-two **Avoid using "all of the above"or "none of the above".**

"All of the above" or "None of the above" should not be used as foils. If a good fourth alternative cannot be constructed, cut your losses and write a different item. An example: *(Music)*

Problem: "all of the above"

A paleontologist studies which of these?

A fossil remains

B geological periods

C ancient life

D all of the above

Solution: "all of the above" eliminated

A paleontologist studies which of these?

A animals

B plants

C fossils

D stars

Part 3 Item Format—Foils

Twenty-three **Make every word count.**

All words and phrases in the stems and foils should serve a purpose. Frequently, the temptation is to set a stage for a question or to add modifiers because they are a traditional way of speaking about the matter. Make every word count. Exceptions may be made when words are added to lengthen a foil or to make some other change that will lend credibility to the foil. Here is an example. *(Music)*

Problem: unnecessarily wordy

Which type of pollution is a major problem for Venice, Italy?

A pollution of air by toxic substances

B pollution of land by insufficient dumpsites and poor garbage collection

C pollution from insufficient noise regulation

D pollution of water by canal boats

Solution: concise responses

Which type of pollution is a major problem for Venice, Italy?

A air

B land

C noise

D water

Part 3 Item Format—Foils

***Twenty-four* Follow the usual rules of punctuation for complete sentences.**

Follow the usual rules of punctuation with foils that are complete sentences. The same is true in the use of proper nouns. An example: *(Music)*

Problem: punctuation lacking

Refrigeration prevents food from spoiling for what reason?

A bacteria reproduce more slowly

B bacteria cannot get into a refrigerator

C toxins are made harmless by the cold

D bacteria are killed by the cold

Solution: complete punctuation

Refrigeration prevents food from spoiling for what reason?

A Bacteria reproduce more slowly.

B Bacteria cannot get into a refrigerator.

C Toxins are made harmless by the cold.

D Bacteria are killed by the cold.

Part 3 Item Format—Foils

Twenty-five **For lists, do not capitalize first word (unless a proper noun); no period at end.**

The initial word of lists, however, should not be capitalized in any of the foils (unless it is a proper noun, of course), and the ending period should be omitted. Consider this example. *(Music)*

Problems: punctuation incorrect

Which of the following best describes a social change?

A The completion of the Trans-Siberian Railroad.

B The increase in the birthrate following World War II.

C The growth of the wine industry in France.

D The division of Germany following World War II.

Solution: punctuation right for list

Which of the following best describes a social change?

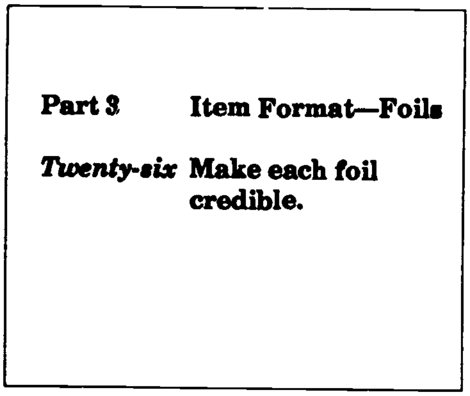
A the completion of the Trans-Siberian railroad

B the increase in the birthrate following World War II

C the growth of the wine industry in France

D the division of Germany following World War II

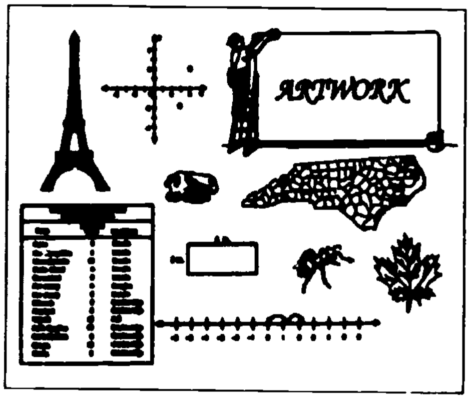
Since each item is written as a complete question, no follow-on punctuation from stem to foil is anticipated.



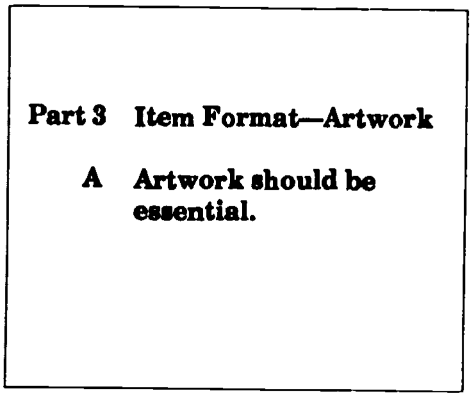
Part 3 Item Format—Foils

Twenty-six Make each foil credible.

Here is a final thought. In this extensive list of guidelines for construction of stems and foils, one idea must be paramount: try to make each foil an answer that would be plausible to an uninformed examinee. This is at the heart of the writing of good multiple-choice items.



The last topic of item form is artwork, which includes drawings, graphs, maps, diagrams, and other inclusions needed to complete the verbal content of the item. The item specifications (to be discussed at the end of this program) will set some guidelines for the extent to which artwork is to be used and the circumstances under which it will be appropriate. Once it is determined that an item will have artwork, the following guidelines apply.



Part 3 Item Format—Artwork

A Artwork should be essential.

After the verbal portion of the item has been written, it should still be evident that the artwork is needed to make the meaning of the item clear.

Part 3 Item Format—Artwork

B Artwork should be appropriate in context.

The artwork should be appropriate. Is it related to the age and educational level of the examinees? Does it focus clearly on the topic being treated? Is it artistically faithful to time and place? Does it present no distracting influences?

Part 3 Item Format—Artwork

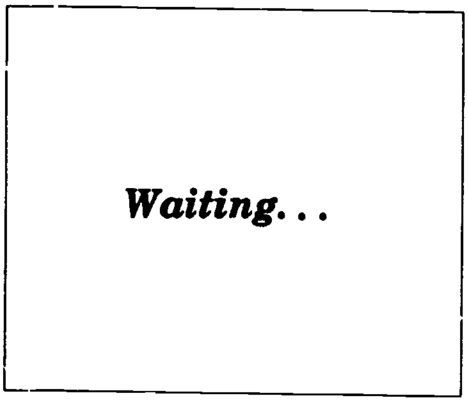
C Artwork should be clean.

The artwork should be clean. Are the forms easily identifiable, numbers legible, proportions correct, grids properly placed, points accurately located on the graph, the activities readily understandable to the examinees? If the item-writer is not expected to produce the artwork, but only to prescribe the form it should take, is the prescription given in enough detail that the artist can achieve these criteria?

Part 3 Item Format—Artwork

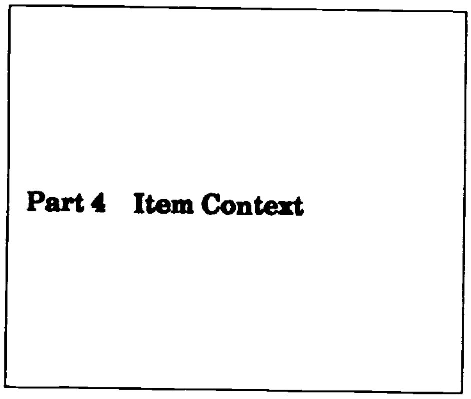
D Artwork should be uncluttered.

The artwork should be uncluttered; it should have no meaningless shadings, lines, labelings, figures, and so on.



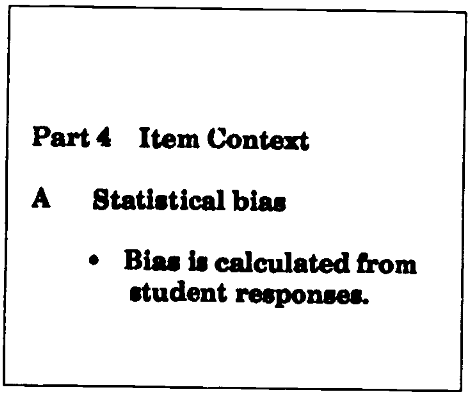
Waiting...

This completes the section on item format. The presentation will stop now while you give your attention to Packet B. The materials in Packet B will give you some further opportunity to get acquainted with the principles discussed under the topic of item format. If you are watching this program in videotape, please place your VCR on STOP. If you are watching this program in slide format, please stop the audio cassette.



Part 4 Item Context

The final section, Part 4, covers the topic of context, in which bias, stereotypes, and fair representation are discussed again.

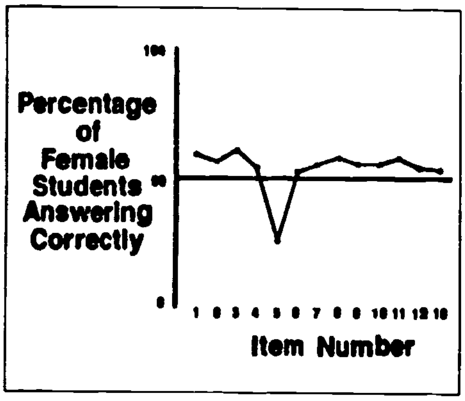


Part 4 Item Context

A Statistical bias

- **Bias is calculated from student responses.**

An item used in an academic achievement test is meant to discriminate between those who know the answer and those who do not. The distinction should be the result of what the two groups have learned, or failed to learn, in a formal instructional program. The group that gives the wrong answer to one question is more likely to give the wrong answer to other questions. Granting that the group is giving the test its best effort, we may assume that it learned less in the instructional program than the other, more successful, group.



This general level of performance across all items provides a baseline for examining each item on the test. If we have a group, say group X, that scores slightly above the average on most items, what do we make of an item on which it scores 30% below the average? We would suspect that the item may cont^in some type of bias. If the item has instructional and curricular validity, it may be desirable to keep the item in the test and say, "This is something on which group X needs remedial instruction." Other things equal, however, it would be preferable to replace that item with an unbiased item.

Part 4 Item Context

A Statistical bias

- **Bias is calculated from student responses.**
- **Item-writer can help screen content for potential bias.**

Several computer programs exist to detect statistical bias. The item-writer can help avoid some item loss through statistical bias, however, by asking of each item whether some group, because of experiences not related to school and homework, would be more likely than another to know the answer to a question. The groups most commonly screened for item bias in academic achievement tests are those based on gender, ethnic origin, and geographical location. The final arbiter of statistical bias is the analysis based on statistical tests and their relation to item content.

Part 4 Item Context

A Statistical bias
- Bias is calculated from student responses.
- Item· vriter can help screen content for potential bias.

B Linguistic (and pictorial) bias
- Bias is found in the eye of the beholder.

Another kind of bias is sometimes called face bias, through an analogy with face validity, or linguistic (and pictorial) bias. Frequently, linguistic bias bears no relation to statistical bias; the examinee correctly answers the linguistically biased question as readily as the unbiased questions. What, then, is the need to consider bias that appears only on the surface and does not affect the measurement process? Linguistic bias should be avoided in a test because it may lower public acceptance of test results, thereby reducing their ultimate usefulness.

Waiting. . .

The presentation will stop now while you give your attention to Packet C of your work materials. In this set of materials, we have attempted to give item-writers some exposure to the more common types of linguistic bias. After you have worked through the material, we will return with some final words on item specifications. If you are watching this program in videotape, please place your VCR on STOP. If you are watching this program in slide format, please stop the audio cassette.

ITEM SPECIFICATIONS

A Goals and objectives

You have completed the fourth and last part of the discussion of item-writing. The topics have been wide-ranging and detailed. They may seem difficult to consider all at once when you are asked to write an item. Some help will be given, however, by the item specifications. An item specification is written by the test developers for each item—or each type of item. These are given to the item-writers when they are ready to begin item-writing.

Item specifications are short descriptions of the characteristics an item is to have. Item specs should cover at least four topics:

- the goal and objective for which one or more items are to be written;
- the location of relevant source material;
- the general characteristics of the examinees who will be asked to answer the item; and
- some technical characteristics that the item must have to meet specific needs of the test development program.

It is assumed otherwise that the item-writers will be familiar with the state-of-the-art technology in item-writing and will employ it in their work.

The goal and objective is a common means of indicating which part of the curriculum the item is to address. Sometimes a third level of specificity will be given, but experience indicates that greater specificity should be used only when it is feasible to specify all possible curriculum branches. For instance, it may be possible, as mentioned earlier, to specify all possibilities for adding one-digit numbers. The danger in greater specificity is that it may fall short of enumerating all curriculum possibilities. This shortfall may artificially restrict the range of items the item-writers will choose to write.

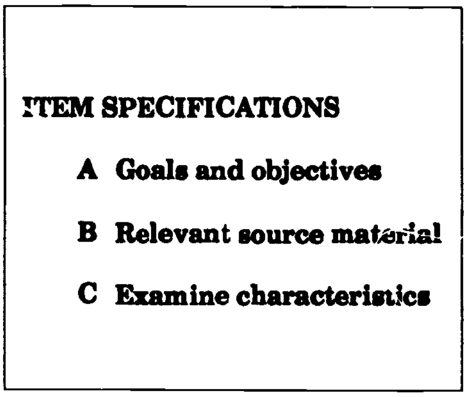
Also given as a curriculum feature may be some indication about the types of items to be written: whether they should be easy or difficult, and whether they should embody "thinking levels." If used, these terms should be given operational meanings that will enable the item-writers to categorize their efforts correctly. The specifications should also indicate how many items should be written, and by whom.

ITEM SPECIFICATIONS

A Goals and objectives

B Relevant source material

The complete item objective should indicate the source of relevant material. This can be as explicit as book, section, and page. Frequently, however, the source may be very general: textbooks approved by the State for that subject or materials available from some other source. Failure to provide source direction may result in items that lack instructional validity.



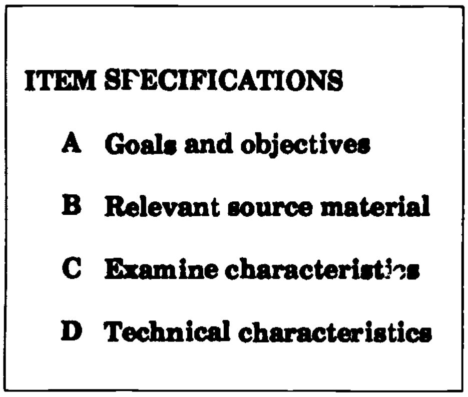
ITEM SPECIFICATIONS

A Goals and objectives

B Relevant source material

C Examine characteristics

Also, the item specifications will give some indication of the characteristics of the examinees. Sometimes, the specifications will be explicit in terms of how many items are to be directed to students scoring above grade level, at grade level, and below grade level. Specificity is especially important when the test is intended to have individual diagnostic characteristics. Frequently, though, the directions may be very general: items suitable for third-grade students; or for any student that has completed Algebra I.



ITEM SPECIFICATIONS

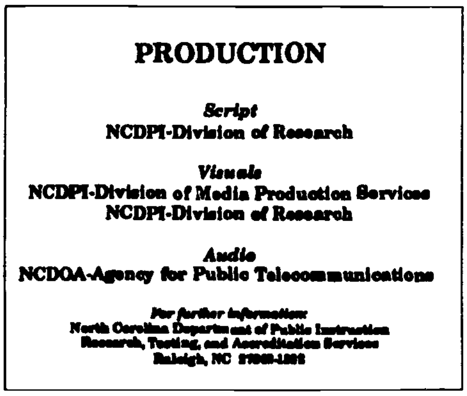
A Goals and objectives

B Relevant source material

C Examine characteristics

D Technical characteristics

Finally, something will be indicated about the technical characteristics the items should have: for example, whether they are to be multiple-choice with four foils, and where the correct foils are to be located. When few details are given, it is assumed that the item-writer will have been trained in item-writing, and will apply that knowledge to the task.



PRODUCTION

Script
NCDPI-Division of Research

Visuals
NCDPI-Division of Media Production Services
NCDPI-Division of Research

Audio
NCDOA-Agency for Public Telecommunications

For further information:
North Carolina Department of Public Instruction
Research, Testing, and Accreditation Services
Raleigh, NC [illegible]

Item-writing can be learned only through practice. But this presentation will give the serious student of item-writing an idea of where to start. Good luck.

Packet A

Work Materials for:

How to Write Multiple-Choice Achievement Test Items

1. List of goals and objectives
2. Item examples for first study section
3. Discussion of classification of item types

NC TESTS

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332

Published 1988 (revised)

1. List of Goals and Objectives

An achievement test can be no better than the objectives that support its content and structure. For illustrative purposes, we have given samples of goals and objectives for several curricular areas as they are represented in the North Carolina Teacher Handbook.

It is not unusual to find that the pressure of test construction acts to identify weak spots in the curriculum, either in terms of areas that are inadequately represented, or in terms of objective statements that are too obscure in meaning to provide a sound basis for item specifications (or instruction). Frequently, curriculum specialists can repair the deficiencies without altering the original intent of the goal statement. Sometimes this can be done through giving examples, sometimes through breaking the objective into sub-objectives.

Examine the samples of objectives given on the next pages and consider what kinds of items could be written for these objectives.

GOALS AND OBJECTIVES FOR VARIOUS CURRICULAR AREAS

Science

Goal 6: The learner will have a general understanding of nuclear energy.

Objective 6.1: Know that isotopes are forms of elements whose atoms differ only by atomic mass.

Objective 6.2: Know that the nuclei of radioactive elements undergo spontaneous change.

Objective 6.3: Know about the processes of nuclear fission and fusion.

Objective 6.4: Be aware of useful applications of radioactive isotopes.

Objective 6.5: Know the necessity for protection against nuclear radiation.

Social Studies

Goal 4: The learner will know that there are different forms of government and that these forms may change over time.

Objective 4.1: Identify European [illegible] Soviet governmental forms.

Objective 4.2: Identify the reasons for, and the results of a change in government in terms of individual rights.

Objective 4.3: Distinguish differences between revolutionary and evolutionary changes in government.

Communication Skills

Goal 3: The learner will use word analysis to aid in comprehension.

Objective 3.1: Identify words by using prefixes, suffixes, and Greek and Latin roots.

Objective 3.2: Determine the effect of an inflectional ending on a root word.

Objective 3.3: Use possessives and contractions to identify meaning.

Objective 3.4: Interpret abbreviations to comprehend meaning.

Algebra I

Goal 7: The learner will solve linear inequalities.

Objective 7.1: Find the solution set for a linear inequality when replacement values are given fc the variables.

Objective 7.2: Solve a linear inequality by using transformations.

Objective 7.3: Use inequalities to solve verbal problems.

Objective 7.4: Find the solution set of combined inequalities.

U.S. History

Goal 9: The learner will know that the Civil War and the Reconstruction of the Union affirmed the power of the national government.

Objective 9.1: Understand how states divided along sectional lines.

Objective 9.2: Understand the causes of the Civil War as immediate and long-term.

Objective 9.3: Recognize the significance of important political/military events related to the Civil War.

Objective 9.4: Distinguish similarities and differences be. veen presidential and congressional plans for reconstructing the South.

Objective 9.5: Describe the effects of Reconstruction on the South.

2. Item examples for first study section

Please consider the following examples.

Examples of items written for goal and objective

Goal 2: The learner will demonstrate knowledge of factors affecting the health of mother and child.

Objective: Describe the effects of alcohol on the health of the pregnant woman and her child.

A woman who drinks alcohol while pregnant may cause which of these defects in her baby?

A damage to liver

B alcoholism in later life

C intoxication at birth

D brain abnormalities

Goal 1: The learner will know important developments in American History from the pre-Columbian period through the first years of exploration and discovery.

Objective: Identify major artistic, scientific, agricultural, and mathematical contributions of pre-Columbian cultures.

Which technical achievement was common to the Mayas, Aztecs, and Incas?

A hybrid grain
B complex architecture
C basic astronomy
D prefabricated housing

Examples of things to avoid in writing items

Tricky or Misleading Item:		Rewrite:	
What is the name and date of the first permanent European settlement in North America?		Which European nation had permanent settlements in North America in the 1500's?	
A	New Amsterdam, 1512	A	England
B	St. Augustine, 1565	B	Holland
C	Jamestown, 1607	C	Italy
D	Santa Fe, 1608	D	Spain

Explanation: The question is tricky for the following reasons:

The key words in the stem are **first permanent European**, making choice B the correct response. Students may be misled into thinking that the question is very simple and that A is the obvious answer since it has the earliest date. A is incorrect, however, since New Amsterdam was settled in the 1620's, not 1512. Students aware of this may then choose C as the correct answer without realizing that Jamestown was the first English, not the first European, settlement.

Language Difficulty Item:	Rewrite: (simplify language)
An orthodontist specializes in what?	An orthodontist specializes in what?
A promoting oral hygiene	A cleaning teeth
B filling caries	B filling cavities
C correcting malocclusions	C straightening teeth
D preventing gingivitis	D preventing gum disease

Explanation: The assumption here is that the purpose is not to test knowledge of technical terms in dentistry, but to find out if the student has practical knowledge of what an orthodontist does.

Thinking Level vs. Item Difficulty

The following examples illustrate the independence of thinking level and item difficulty.

Low Thinking Level, High Difficulty	Low Thinking Level, Low Difficulty
What is the most common cause of maternal mortality in the United States? A ectopic pregnancy B birth complications C prematurity D post-natal infection	What is a good source of vitamin C? A pork chops B orange juice C candy bar D peanuts

Where to Place the Interrogative

In general, place the interrogative as close to the item foils as possible. For example:

NO **To which group** does Cro-Magnon human belong?

YES Cro-Magnon human belongs **to which group**?

NO Why are human fossils harder to find than fossils of horses?

YES Human fossils are harder to find than fossils of horses **for which of these reasons**?

NO What can be learned about a specimen if the foramen magnum opens at the bottom of the skull?

YES If the foramen magnum opens at the bottom of the skull, we can draw **which of these conclusions**?

In some instances, however, it is better to place the interrogative at the beginning of the stem. This is usually where the foil becomes a predicate adjective of the stem or forms an identity. For example:

NO Lead is which color?

YES Which color is lead?
Lead is gray. (predicate adjective)

NO Which of these is larger in primates as compared to other animals? (original form)

NO When primates are compared with other animals, the larger primate organ is which of these? (revised with interrogative at end)

YES When primates are compared with other animals, which primate organ is larger?
Cerebrum is larger. (predicate adjective)

NO The Trade Expansion Act is also known as which of these tariffs?

YES Which tariff is also known as the Trade Expansion Act?
Trade Expansion Act = Tariff of 19__ (identity)

Do not worry too much about these fine distinctions. All items will be edited before they appear on the final test form.

3. Classification of Item Types

Four accounts can be given of an object (idea, concept, thing). These have been described as denotive, classificatory (relational), purposive, and functional.[a] A further description will be given of them here. Table 1 will provide a point of reference.

Table 1
Structure of the Item Classification System

Type 1	Type 2	Type 3	Type 4
denotive	relational	purposive	functional
who, where, when?	what?	why?	how?
cross-sectional	cross-sectional	sequential	sequential

Types 1 and 2 are similar inasmuch as they are cross-sectional; Types 3 and 4 are similar in that they are sequential. Let us take a closer look at Types 1 and 2 first.

[a] cf., Aristotle's four causal categories: material, formal, purposive, and efficient.

Type 1

Type 1, denotive (who, where, when), permits the concept to be distinguished from other concepts. The content of a Type 1 item acts as an arrow to point to the concept, to discriminate it from others. If the concept is a physical object (a person, a rock, a planet), we can expect denotive questions to deal with names, sizes, colors, characteristics regarding time, and other features that are unique to those objects. For example, the planet is named Mars; it is located in an orbit outside Earth and inside Jupiter; it appears to be red in color; it is 4,215 miles in diameter; etc.

Exactly the same type of information can be given about the person or the rock. The rock is called Gibraltar. It is located at the southern tip of Spain; etc. All together, these denotive data describe an object unique in space and time.

The same Type 1 logic applies when the object is an abstract concept such as war or exploration.

> The war was called the Spanish-American War; it was fought in 1898 in Cuba and the Philippines between Spain and the United States; it involved certain numbers of ships, sailors and troops; etc.
>
> The exploration was called the search for the Northwest Passage; its locale was the northern portion of what is now the United States and the southern portion of what is now Canada; the search lasted for five centuries; etc.

As you can see, the descriptive material for the denotive item is individualistic and particularistic. It isolates the concept from others. Table 2 contains four items that are denotive in nature.

Note that the categories used to denote the concepts are general — e.g., extent, weight, date of origin, or name.

Table 2
Denotive Items

Cortés conquered which Indian group?

A Aztec
B Maya
C Inca
D Sioux

The current laws of planetary motion were discovered by whom?

A Copernicus
B Galileo
C Kepler
D Newton

The first battle of the Civil War took place where?

A Fort Sumter
B Bull Run
C Antietam
D Gettysburg

The Declaration of Independence was signed in what year?

A 1774
B 1776
C 1781
D 1786

Type 2

Type 2, classificatory or relational (what), has much in common with Type 1. It is cross-sectional. It defines the concept by classifying it. But the classificational categories are systematic, apply to a limited set of concepts (objects), and provide an exclusive set of relations. For example, the animal-vegetable-mineral set of categories can provide a basis for a Type 2 item. Compare this set of categories with the infinite set of categories making up the scale for weight. A rock and a dog both have weight, but a rock is exclusively a mineral and a dog is exclusively an animal.

The systematic categories supporting the Type 2 item cannot always be expected to add up to exactly four categories (to provide the four foils of an item). Sometimes there will be only two categories — kinetic and potential energy, for example. Thus the other two categories must be lifted from some other system of logic. Just the opposite also may be the case. The category system may have more than four components, in which case only four can be given. Preferably, the sat of categories will have at least three components; otherwise, the examinee may see through the system and reduce the selection to only two foils.

In general, the Type 1 item will involve quantitative descriptors (extent, weight), the Type 2, qualitative descriptors (ego, id, superego). Some quantitative measures will masquerade as qualitative measures. For example, height may be given in inches or as tall, short, or medium. The latter is hardly the type of qualitative measure we expect in a Type 2 item. A tree can also be tall, short, or medium in height. Table 3 contains four Type 2 items.

Table 3
Classificatory Items

Which term best describes all of the organisms living in a pond?

A population
B ecosystem
C community
D biome

The writer William Wordsworth belonged to which literary period?

A Elizabethan
B Jacobean
C Romantic
D Victorian

Which division of the psyche is associated with instinctual impulses?

A ego
B id
C self
D superego

Which is the least populated biome?

A desert
B polar region
C tropical region
D tundra

Types 3 and 4

Type 3 (purposive) and Type 4 (functional) categories are so closely related that their meaning will be clearer if they are discussed together. Type 3 is said to answer the question *why* and to deal with means and ends. Type 4 answers the question *how* and deals with cause-effect relationships. Sometimes the two types seem merely to be opposite sides of the same coin. For example, "*Why* did Columbus sail west?" "To reach the Indies." But, "*How* did Columbus expect to reach the Indies?" "By sailing west." The facts are: sail west → reach Indies. This is simple cause and effect and is conceptualized as a basic Type 4, *how*, question, even though the word *why* is employed in the item stem.

The use of *why* as an interrogator regarding events close in time can be so inappropriate that it elicits a laugh. Everyone is familiar with two old examples: "Why does a chicken cross the road?" "To get to the other side." and the question asked of Willie Sutton: "Why do you rob banks?" "Because that's where the money is." In both cases, some more remote reply was expected. In Willie Sutton's case, we might have expected him to say, "I never had an opportunity to learn the skills required to obtain money honestly (deprivation)," or, "There is no need to work for a living when you can just help yourself to money (asociality)."

The Type 3 (purposive category) should be used to indicate a remote or final relationship. Why did Columbus sail west? To try to discover a new trade route to the Indies. *Trade* was behind the *why*. Merely reaching the Indies was not the purpose of the voyage. For illustration, four purposive items are given in Table 4.

Table 4
Purposive Items

President Theodore Roosevelt issued the Roosevelt Corollary for which reason?

A to prevent European interference in the Americas
B to block Japanese expansion into the Pacific
C to justify American intervention in the Americas
D to reverse American policy under the Monroe Doctrine

President Franklin Roosevelt declared a bank holiday in 1933 to accomplish what?

A to mobilize support for the FDIC
B to collect overdue government loans
C to keep cash reserves from being exhausted
D to stabilize the price of gold certificates

Americans favored an isolationist policy following World War I for which reason?

A Americans were concerned with restoring the national economy.
B Americans had no interest in the problems of foreign countries.
C Americans were discouraged by the loss of men and materials on foreign soil.
D Americans were discouraged by anti-American sentiments held by Europeans.

The Navigation Acts of 1663 were passed for which purpose?

A to protect New England merchants from [illegible]iots
B to protect the economic interest of England
C to protect colonial trade with the West Indies
D to protect the slave trade with England

Type 4, *how*, deals with causal sequences: one thing leads to another, which leads to another, which leads to another, etc. Questions that probe for all or part of these sequences are *how* questions. The best known examples of how sequences are illustrated in the "how to do it" manuals: how to build a house, how to lose weight, how to sell real estate, etc. Most scientific questions are how questions. The science of chemistry began with a how question, albeit a futile one: how to transmute base metals to gold. In a sense, the predominance of *how* questions over *why* questions separates the modern age from the medieval age. (The *why* question, however, may again be coming into its own in a neo-modern age.) Four functional questions are given in Table 5.

Table 5
Functional Items

In the 1920's, an investor bought stock "on margin" by which method?

A by paying part of the price down and borrowing the rest
B by agreeing to sell the stock after only a few weeks
C by paying for the stock in installments
D by purchasing certificates at a discount interest rate

The passage of food through the digestive tract follows which order?

A esophagus, stomach, small intestine, duodenum
B stomach, duodenum, esophagus, small intestine
C esophagus, small intestine, large intestine, stomach
D stomach, esophagus, large intestine, duodenum

Having adequate fiber in your diet helps your body by what means?

A cleans out materials left after digestion
B maintains fluid balance in the body
C supplies the body with essential vitamins
D builds and repairs body cells

Honey bees communicate the location of food to other bees by which method?

A a dance
B chemical scents
C verbal messages
D trial and error

Higher Order Questions

We have discussed four ways of categorizing items, in brief calling them *which*, *what*, *why*, and *how* items. Frequently, item-writers may be asked to write an easy, or a difficult, item. That is clear enough. Then again, they may be asked to write a high or low-order thinking level item. Efforts to give those terms useful meanings are typically less than satisfactory. No attempt at definition will be made here, except to note that the low-order thinking items seem to be one-step memory items, while the high-order thinking items may require more than one step to arrive at an answer.

One aspect of the problem can be laid bare, however, without offering a clear solution. In dealing with sequential material, the path goes from A to B to C to D in an unbroken historical line of advancement. Most elementary and secondary textbooks concentrate on this unbroken line. Yet when events were at A, B was not a certainty. Some other path could have been taken. Also, we might not have used A to get to B; there usually are competing means available at a historical moment of decision. Or we might have used A, but not to get to B.

A real understanding of events requires not only that one know things proceeded from A to B, but why they took that form and that direction. With that understanding we can truly profit from the past. Note that the student is not being asked what he or she would have done in some historical situation, which is speculative at best, but what pro and con arguments were made by people on the scene at the time historical decisions were made.

This line of reasoning may become clearer with some examples and illustrations. For example, take the question, "Why was the Gadsden Purchase made?" In everyday usage, we might answer, "To get land for a railroad." That suggests a cause-effect relationship, not a probing of means. Why was the Gadsden Purchase made in preference to some other means of acquiring the desired transportation? We could have put the railroad elsewhere; we could have bought just the right-of-way, or leased it; we could have asked the Mexicans to build a railroad that we could use: these are some of the alternatives to the Gadsden Purchase. We made the Gadsden Purchase, in all likelihood, because we wanted full control of the development potential and could force our will on a weak neighbor.

A similar condition is involved in the question, "Why did the Russians blockade Berlin?" They hoped to force the Allies out, of course, but *why* the *blockade*? Because the Russians had the quasi-legal right to stop traffic; the act was not likely to provoke open warfare; and its effects could be unacceptably punishing to Berliners. Some alternatives to the blockade were to bomb or invade the Allied sectors; to make diplomatic demands and threats of punishment; to tax goods entering Berlin; etc. None of the alternatives was as attractive as the blockade. The Russians chose to blockade Berlin because of the blockade's means-end effectiveness.

In another example, take the question, "Why did Roosevelt declare a bank holiday in 1933?" We know he wanted to help banks, which were failing one after another. But why a bank holiday rather than loaning the banks money, or asking for public restraint? Roosevelt declared a bank holiday in 1933 because it was an immediate, legal, and focused way of dealing with the act that was precipitating bank failures: depletion of cash reserves by depositors asking for the return of their deposits

The situation may be diagrammed as in Figure 1.

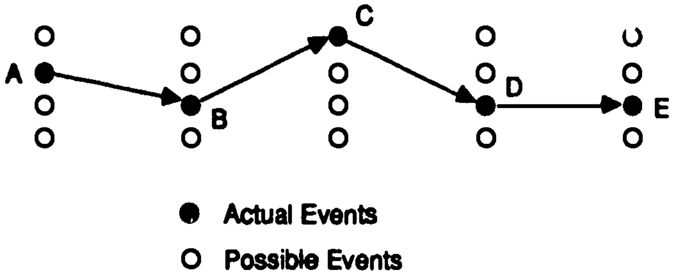


Figure 1 — Path of actual events compared with possible events

The meaning of a decision is given not only by its consequences, but by the matrix of possibilities at the time the decision was made.

Again, though, testing must reflect the curriculum and instruction. Demands should not be made on item writers to lead the curriculum or to test for non-achievement abilities such as general intelligence. The probing of means and ends, if it is to become the topic of items, must be represented in the educational program.

Please return now to the presentation.

Packet B

Work Materials for:

How to Write Multiple-Choice Achievement Test Items

1. A set of incorrectly written items to be edited or rewritten
2. A reference list of item format characteristics
3. A discussion of the nature of the errors contained in the set of items

NC TESTS

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332

Published 1988 (revised)

1. Items For Revision

Please examine the item on the left and revise it to correct any problems it may have, either by editing the item directly in the left-hand column, or rewriting it in the blank right-hand column. If necessary, consult your list of desirable characteristics (page 9).

If rewriting the item requires more technical information than you have at hand, just indicate what is wrong with it.

1. The Apache Indians lived in the

   A Northeast.

   B Northwest.

   C Southeast.

   D Southwest.

State causes and results of the French and Indian War.

2. Who was a young Virginia militia officer during the French and Indian War?

   A Sieur de La Salle

   B William Pitt

   C Louis Joliet

   D George Washington

3. The development of the St. Lawrence Seaway allowing the migration of which of the following into the Upper Great Lakes, which led to the resultant destruction of game fish?

   A lake trout

   B salmon

   C coho

   D lamprey

4. Which of the following functions is not performed by the blood?

A digesting food

B transporting food

C regulating body temperature

D resisting disease

5. What was a major goal of late nineteenth-century trusts?

A creating better relations among new immigrants

B improving science

C controlling the production of a single commodity

D not giving debtors an opportunity to default

6. Which development resulted from the 1840 potato famine in Ireland?

A Irish farmers learned to spray their potatoes to fight disease.

B After this disaster, only corn and wheat were planted.

C Emigration to other countries increased.

D Ireland became a major manufacturing country.

7. How is electricity produced on a large scale?

A burning wood

B running ·/ater over a dam

C capturing and storing underground steam

D fusion

8. What kind of rock is formed when lava erupts from a volcano?

A igneous

B asphalt

C liquid

D fossils

9. Which of these changes existing rock into metamorphic rock?

A pressure

B wind

C weather patterns

D volcanoes

10. What was a major goal of late nineteenth-century trusts?

A producing goods for wholesale export

B controlling production of goods

C providing regional goods and services

11. The width of a rectangle is 4 inches less than half its length. If the perimeter of the rectangle is 64 inches, what is its width?

A 8 inches

B 2 feet

C 20 inches

D 2 feet, 2 inches

12. Which tariff is also known as the Trade Expansion Act?

A Tariff of 1962

B Tariff of 1828

C Tariff of 1909

D Tariff of 1930

13. When did Germany gain control of the Sudetenland?

A 1935

B 1938

C after the Nazi-Soviet Pact

D with the signing of the Locarno Pact

14. Solve: 248 + 123 =

A 372

B 371

C 258

D 125

15. The adrenal medulla secretes which of the following?

A androgens

B adrenalin

C prostaglandins

D steroids

16. Which invention was first developed for the space program?

A disposable clothing

B freeze-dried foods and beverages

C radar

D canned foods and beverages

17. How did the Pendleton Act of 1883 affect government workers?

A awarded government jobs on the basis of competitive examinations

B put most government jobs on an appointive basis

C classified eighty-five percent of all government jobs as civil service positions

D made campaign contributions by federal employees illegal

18. Which is the same as 1 + 1 + 1?

A 6 + 2

B 8 × 4

C 3 × 1

D 7 − 5

19. Which of these is never a renewable resource?

A wood

B steam

C rubber

D oil

20. Which of these does a paleontologist study?

A fossil remains

B geological periods

C ancient life

D all of the above

21. Which type of pollution is a major problem for Venice, Italy?

A pollution of air by toxic substances

B pollution of land by insufficient dump sites and poor garbage collection

C pollution from insufficient noise regulation

D pollution of water by canal boats

22. Why does refrigeration prevent food from spoiling?

A bacteria under refrigeration reproduce more slowly

B bacteria cannot get into food in a refrigerator

C bacteria-producing toxins are made harmless by the cold

D bacteria are killed by the cold

23. Which of the following best describes a social change?

A The completion of the Trans-Siberian Railroad.

B The increase in the birth rate following World War II.

C The growth of the wine industry in France.

D The division of Germany following World War II.

Please return now to the presentation.

Packet C

Work Materials for:

How to Write Multiple-Choice Achievement Test Items

Discussion and supplementary materials

NC TESTS

North Carolina Department of Public Instruction
Division of Research/Raleigh, NC 27603-1332

Published 1988 (revised)

The Roots of Item Bias

As a means of placing item bias in perspective, imagine that you have a computer bank of achievement test scores with all sorts of student background information: sex, race, locality, age, family status, schools attended, teachers, religion, IQ, and so on. Imagine that you calculate the test score averages for the groups suggested by the background variables—male/female, low IQ/high IQ, etc.—and that you find differences in achievement. Is that evidence of some sort of test or test item bias? No—not in itself.

Suppose that you find that some students had early advantages or special experiences that other students did not (such as speaking French in the home). If the items reflect these advantages, are the items biased? No, that is not sufficient evidence of item bias. For an item to be unbiased, it needs to satisfy only two conditions:

1. The item must be a valid measure of a curricular objective.

2. The item must not require *non-curricular* information that places one student at an advantage over another.

A test can be constructed from items that are individually unbiased, but become biased in the aggregate. This bias can occur when one section of the curriculum is favored over another section without a good pedagogic reason. For example, boys may have an overall advantage over girls in problems that deal with space perception. This advantage does not mean that problems requiring space perception should be eliminated from the test; these problems, however, should not be over-represented in relation to their importance in the curriculum. For a test to be unbiased, then, it must satisfy two conditions:

1. The test must be constructed from unbiased items.

2. Each curricular objective must be represented equitably with respect to item coverage.

With reference to the two conditions needing to be satisfied to produce an unbiased item, the first, that the item must be a valid measure of a curricular objective, is reasonably clear and straightforward. The second, that the item should not contain *non-curricular* material that places one student at an advantage over another, is far more complex and is the more common source of bias. Let us consider the second criterion in more detail.

An instructional objective is an abstraction standing in unadorned simplicity. The item must clothe the objective in tangible garments. *Add two one-digit numbers* becomes *2 + 2 = ?*. That seems clear. But if we wish to test the understanding of some general principle regarding family life, can we select an example that requires the examinee to have had personal experience with the nuclear family as a prerequisite to answering the question? This may be unfair because not everyone grew up in a nuclear family. Not everyone has personal experience regarding snow, or tides, or farm life, or sibling competition, or sports. Care must be taken in choosing the garments in which we clothe items testing general principles. We must be certain that we do not assume common experiences where they do not exist.

Two types of assumptions are frequently found:

1. *We err in extending the general case to the particular case.* We have lots of help in making this error. The culture is replete with generalities that help us in understanding the general trend of things: men are stronger and taller than women, for example; men are more interested in sports; women are more emotional; men are not very understanding. Pick the topic—race, region, religion—; the list is endless. Product developers and advertisers use these differences as the basis for identifying markets. Their businesses succeed or fail on their success in making valid generalizations.

These generalities, however, can all fail in the particular case. The strongest and tallest person in the third grade may be a female; the person picking up the baseball and the catcher's glove may be a female; the most understanding person may be a male. Because the generalities may not be true for the person taking the test, they should not be assumed to be true when the item is written. Enough general material exists to find instances to illustrate a curricular principle without drawing on generalities that may not be valid for the individual taking the test.

2. *We err in extending the particular case to the general case.* This problem is particularly elusive, because it may not be as obvious to the item-writer. If, for example, I grew up in a nuclear family with a father, mother, brothers, and sisters, I may forget that everyone did not have that experience. If I played guitar in a rock band in high school, I may forget that rock music was forbidden to some students. If I like art and creative dancing, I may forget that such activities are of no interest to some students and that others may have had no experience with art and dancing at all.

Because of the personal factor, item-writers must be careful not to assume that their particular experiences are general experiences for all students. Watch for the instances where items are clothed with references to the mountains or the ocean; to small towns, farms, or cities; to middle-class housing or transportation; or to special family conditions. Ask of each item: Are the non-curricular references in this item common to all students who will take the test?

Before going ahead, please consider an example. Suppose the objective concerns the environment and deals with tides. A question on tides is perfectly legitimate. All students, whether they come from the coastal plains, piedmont, or mountains, should have studied the subject in the classroom and be prepared to answer it. But if the objective is one concerning arithmetic, questions about the characteristics of tides as measured by numbers may somehow give the coastal student an unfair advantage. In that instance, apples and oranges may be the better substance in which to clothe the item.

McGraw-Hill's *Guidelines*

For further study, we have included a copy of McGraw-Hill's *Guidelines for Bias-Free Publishing*. You will find McGraw-Hill's extensive and detailed *Guidelines* helpful in alerting you to some common forms of bias.

The document is not perfect. McGraw-Hill, however, must be congratulated on its effort to help with this troublesome and comparatively unresearched topic. The booklet should be taken for what it is: a serious effort to identify offensive words and concepts and to create a friendly contextual background for academic material. Please return it to us.

Bias and Test Validity

The biases discussed in McGraw-Hill's *Guidelines* deal mostly with terms and representations that may offend someone. A test may be offensive without being invalid. For example, every illustration in a test may picture only white males, but this may in no way interfere with the ability of a female or black to answer the questions correctly. A more significant threat to test validity is the use in tests of **non-academic** background material that unwittingly presents an advantage or disadvantage to some group whose life experiences outside of school have differed significantly from those of other groups. People who live on the coast know about tides; people who live in the mountains know about ski slopes; people who have money know about investments; people who have unusual musical talent are more likely to be involved in music outside of school: the list of lifestyle advantages and disadvantages is endless. We cannot anticipate every possible threat to validity; you must search in your own mind to avoid these biases when you write a question. Remember that we are seeking to determine the degree of knowledge attained by a student in some **academic** subject. If the academic subject is mathematics and the student's parents are mathematicians and have taught the student at home, that is an advantage, but not a bias. The student knows mathematics because he or she knows mathematics. An example of a bias would be a mathematics test that gave an advantage to students who were familiar with ski slopes, or farms, or city traffic, or brothers and sisters, or any other non-mathematical lifestyle experiences that were essential to an understanding of the test questions, but were not experienced by everyone.

The topic of bias and test validity is endless. McGraw-Hill's 38 page *Guidelines* hardly touches the subject. We have suggested some ways to avoid bias and have enunciated the general principles. But do not become so concerned about this topic that you allow it to inhibit your creative effort. Remember that the items will be edited later by several people and subjected to statistical tests of bias. Through this cooperative effort, we should be able to produce tests in which bias plays no significant role.

ITEM SPECIFICATIONS SHEET

CURRICULUM OBJECTIVE:

1.1 Know how to solve physics problems using basic algebra and trigonometry.

DIFFICULTY LEVEL: [✓ 1 = EASY]
2 = MEDIUM
3 = HARD

LEVEL OF THINKING SKILLS: [✓ 1 = LOWER]
2 = HIGHER

ARTWORK REQUIRED: 1 = YES
(IF YES, PLEASE ATTACH) 2 = NO

CURRICULUM SOURCE:

ITEM WRITER NUMBER: 12

PHYSICS TEST ITEM *(FINAL DRAFT)*

CORRECT ANSWER A

EDIT ____ ____ ____

Did You...

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

Appendix E

Booklet, Instructions for Item Review of Achievement Test Items, and Sample Item Record for Item Review

PROCEDURE FOR ITEM REVIEW

To develop achievement tests that are valid, reliable, educationally appropriate, economical, and administratively manageable, the NCDPI Division of Research staff carries out a two-year series of operations. In a broad overview, the procedures call for (1) curriculum definition; (2) test design; (3) the creation of item specifications; (4) the selection and training of North Carolina teachers in the writing of multiple choice test items; (5) the writing of test items; (6) their initial editing; (7) a review of items by North Carolina teachers; (8) the collection of items into pilot test booklets; (9) their administration to samples of students; (10) analyses of the data; (11) further editing; (12) the selection of items for tests; (13) another review of the test booklets by teachers and curriculum supervisors; (14) the final editing of the test booklets; (15) preparation of test administration manuals; (16) printing of test and manuals; (17) test administration; and (18) preparation of norms and technical manuals. In the above sequence, we are now at stage 7 (a review of items by North Carolina teachers).

To carry out the item review, you first need to be familiar with the directions given to item-writers. That information is contained in the script for *How to Write Multiple Choice Achievement Test Items*, a copy of which is enclosed. Read through it until you have a good grasp of the characteristics an item should have. Also enclosed is a copy of McGraw-Hill's recent publication, *Guidelines for Bias-Free Publishing*, which will alert you to current issues in linguistic bias, although the solutions offered may not always be ideal from everyone's point of view.

Second, you need to know specifically what item the item-writer was trying to write. That information (objective, etc.) is given at the top of each Test Item Review Sheet.

Third, you need to know what item characteristics to consider and to have a clear idea of what exactly is meant by each characteristic. That information is given in the section of this booklet entitled *Item Review Categories.*

Fourth, you need a systematic way to convey your judgments to us, bearing in mind that we must synthesize about 250,000 judgments (number of items × number of characteristics × number of reviewers). To accomplish that, we ask that you use the list at the bottom of each item record. In short, follow the procedure given below:

1. Note the information given at the top of the Test Item Review Sheet (the objective, etc.).

2. Read the item carefully.

3. Record the correct answer to the item in the space provided. (At this stage of item development, reviewers occasionally see an ambiguity in a foil and select an unexpected answer. We need to have that experience now rather than later.)

4. Consider the eight characteristics of the CONCEPTUAL section first. Go down each characteristic and determine whether the item satisfies the need defined by that characteristic. **If the item fails to satisfy one of the needs, check the box next to that characteristic.** Then turn over on the back of the item record, list the number appearing at the side of the characteristic you have checked, and explain the problem you perceive.

When you have completed the CONCEPTUAL section, check one of the blocks under "CONCEPTUAL": **yes**, if the item is OK conceptually; **marginal**, if the item needs editing before approval; and **no**, if that item should not be used in any form. NOTE: if you have checked **marginal** or **no**, be sure to reference one or more of the numbers on the back of the item record, and explain the problem.

5. After you have checked the CONCEPTUAL rating scale, go on to the LANGUAGE section and complete it; and so on through the DIAGRAM section (if there is a diagram).

6. Finally, make an overall judgment about the item by checking one of the three categories at the bottom of the page. Again, the second and third categories should be accompanied by your references on the back of the item record.

It is likely that this procedure will be slow at first, as you go through each characteristic, sometimes having to refer to the *Item Review Categories* to refresh your memory regarding the meaning of one of the briefly indexed categories. But as the process becomes more familiar, your speed should increase.

NOTE: Do not be concerned about the mix of items you are reviewing and the distribution of types of items on the final test form. Your group of items is not necessarily representative of the final selection of items, either in terms of difficulty, curricular representation, or in any of the other characteristics to be considered in the final item selection.

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

When you have completed the review of the items, please return the notated item records to us in the manner described in our letter of transmittal, together with any other material requested. If you wish, you may keep the *Instructions for Item Review*, *How to Write Multiple Choice Achievement Test Items*, and *Guidelines for Bias-Free Publishing* booklets.

ITEM REVIEW CATEGORIES

The following descriptions are given to clarify the meanings of the brief categories listed at the bottom of each Test Item Review Sheet.

CONCEPTUAL

1. **Objective match.** An item must measure some valid and significant aspect of the instructional objective listed at the top of each item record.

2. **Fair representation.** With respect to sex, race, geographical location, and other personal characteristics, test items should be neutral. When possible, references to "he" and "she" should be omitted; otherwise, they should be balanced. Fair representation should be given to ethnic groups. Phrases or attitudes that are demeaning to anyone should not be part of test item content.

3. **Cultural bias.** In writing items, an item-writer, in an attempt to make an item more interesting, may introduce some local example about which only local people have knowledge. This may (or may not) give an edge to local people and introduce an element of bias into the test. This does not mean, however, that no local references should be made if such local references are a part of the curriculum (in North Carolina history, for example). The test of bias is this: Is this reference to a cultural activity or geographic location something that is taught as part of the curriculum? If not, it should be examined carefully for potential bias.

Stereotypes of any person or group should be avoided. As an example of the sort of thing that should be avoided, we paraphrase below a portion of an item taken from a 10-year-old reading test:

> (Ice cream vendor) Boys take less time than girls to make up their minds, so I will serve you first, Jim.

This stereotyping is totally unacceptable today.

4. **Clear statement.** The question introduced by the item should be stated clearly and unambiguously.

5. **Single problem.** The item should not deal with two problems at one time. If the student chooses the incorrect answer, it may not be clear which part of the problem the student did not understand.

6. **One best answer.** Among the four foils, one answer should be clearly, unambiguously the best answer.

7. **Common context in foils.** The four foils should all relate to a common context. Sometimes, one answer will stand out as different from the others because it deals with a different context. That difference may give test takers an unintended clue to the correct answer. If foils must deal with disparate content, all answers should be different in that respect.

8. **Each foil credible.** A four-foil test item is sometimes reduced in effect to three or two foils because one or more foils would not be judged as possibly correct by even the most uninformed test taker. Each foil should be a believable answer for someone who does not really know the correct answer.

9. **Other.** This category should be used to document CONCEPTUAL characteristics that require comment but are not listed.

LANGUAGE

10. **Appropriate for age.** The language and noncurricular references in the item should be appropriate for the age-group being tested. The non-curricular language should be aimed at a reading level appropriate for the low-ability spectrum of the age group being tested, following the logic of testing for knowledge of the specific content of the curriculum, not language sophistication.

11. **Punctuation, spelling, grammar.** This is self explanatory. Punctuation, spelling, and grammar should be proper and correct.

12. **Excess words.** Flowery language is particularly out of place in testing. It serves no useful purpose, and takes up time and space.

13. **No stem/foil clues.** More than occasionally, the stem of an item will contain a key word that appears in only one foil. This "echo" provides an unwanted cue for the informed test taker. "Clangs" and "echoes" from stem to foil should be avoided.

14. No negatives in foils. Test takers who are struggling to answer a large number of questions are easily confused by negatives. "None of the above" and single foils that negate a condition should be avoided.

15. **Other.** This category should be used to document LANGUAGE categories that require comment but are not listed.

FORMAT

16. **Logical order of foils.** Other things equal, foils should be presented in logical order: order of magnitude, length of foils, location (e.g., east to west, top to bottom), etc.

17. **Familiar presentation style.** The form of the problem should be familiar to the student. For example, if the student's text presents stacked fractions, then the item should also.

18. **Print size and type.** These characteristics should be similar to the ones the student sees regularly in the classroom.

19. **Mechanics and appearance.** The general layout of the item should promote comprehension of what is being asked.

20. **Equal length foils.** Probably the best advice a poorly prepared test taker could have would be, "When in doubt, check the longest foil." Item-writers know more about the correct foil than the incorrect foils and tend to elaborate the correct foil. No more than 25% of the correct answers should be the longest foil, and it is bad policy ever to make the correct answer conspicuously longer than the wrong answers.

21. **Other.** This category should be used to document FORMAT characteristics that require comment but are not listed.

DIAGRAM

22. **Necessary.** The illustration accompanying an item should be necessary to an understanding of the question.

23. **Clean.** A diagram should be uncluttered. It should be unambiguous in what it depicts.

24. **Relevant.** A diagram just for the sake of illustration is inappropriate for test items. The diagram should contain only relevant elements.

25. **Unbiased.** It is particularly easy for a diagram to introduce bias into an item. A diagram should not deal exclusively with urban environments, with scientists who are invariably male, with dishwashers who are invariably female, and so on

26. **Other.** This category should be used to document DIAGRAM categories that require comment but are not listed.

The item reviewer should make additional general comments on the back of the item record if such comments will contribute to the development of better North Carolina achievement tests.

Notes

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

Geometry Test Item Review Sheet

1.10 Identify interiors and exteriors of geometric figures.

Difficulty Level: Easy

Art: Yes

0047

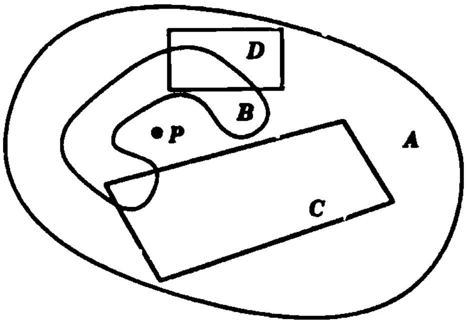
Point *P* lies in the interior of which geometric figure?

A *A*

B *B*

C *C*

D *D*



14 Correct Answer: ________

CONCEPTUAL

Yes Marginal No

15 □ □ □

16 □ Objective match

17 □ Fair representation

18 □ No cultural bias

19 □ Clear statement

20 □ Single problem

21 □ One best answer

22 □ Common context in foils

23 □ Each foil credible

24 □ Other

LANGUAGE

Yes Marginal No

25 □ □ □

26 □ Appropriate for age

27 □ Punctuation, spelling, grammar

28 □ No excess words

29 □ No stem/foil clues

30 □ No negatives in foils

31 □ Other

FORMAT

Yes Marginal No

32 □ □ □

33 □ Logical order of foils

34 □ Familiar presentation style

35 □ Print size and type

36 □ Mechanics and appearance

37 □ Equal length foils

38 □ Other

DIAGRAM

Yes Marginal No

39 □ □ □

40 □ Necessary

41 □ Clear

42 □ Relevant

43 □ Unbiased

44 □ Other

45 OVERALL RATING: □ Acceptable □ Acceptable with modifications □ Discard Item

List the number next to the box you have checked on the other side and explain the problem with the item.

Appenaix F

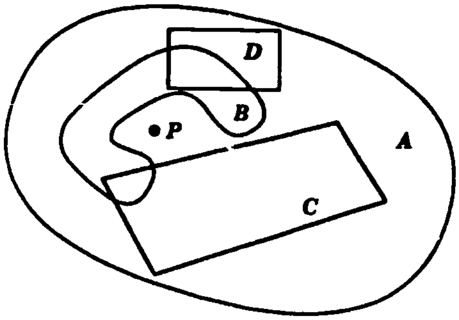
Sample Item Record with all Data Included

Geometry

Objective: 1.10 Identify interiors and exteriors of geometric figures.

GEOMETRY						PT		------CHOICE------				------BIAS-----	
ORIGNO	FORM	ITEM	OBJ	P	SD	BISER	KEY	A	B	C	D	WHITE	FEMALE
0047	F	15	1.10	0.96	0.20	0.11	1	643	22	4	1	0.0752	0.0183

		ADJ			TOTAL	WTN	--ITEM CHARACTERISTIC CURVE--					
P(A)	RASCH	RASCH	SE	DISC	T	MS	1	2	3	4	5	6
0.98	-3.32	-3.37	0.26	0.87	0.03	0.99	0.93	0.96	1.00	0.99	0.99	0.98



15. Point *P* lies in the interior of which geometric figure?

A *A*

B *B*

C *C*

D *D*

F 15 EASY WEAK PRED INVERTED ICC

Yes No	Yes No	Yes No	Yes No	Yes No
Psychometric Approval	Edit Approval	Curriculum Approval	Committee Approval	Final Approval

Appendix G

Test Review Instructions and Questionnaires

Notes to the Curriculum Supervisor:
Review of the 1989 North Carolina Test of Biology

Thank you for agreeing to provide professional leadership for the review of the 1989 North Carolina Test of Biology. These reviews will be going on simultaneously in all eight educational regions. Their success will depend upon each regional team approaching the review with some consistency and common purpose. Your leadership will be essential in bringing that about.

To help with your planning of the review, we have outlined a typical agenda. The final agenda will of course be up to you, based on your analysis of the particular needs of your review team.

The Morning Session

You may wish to begin the session by summarizing introductory comments from the pages of instructions.

- Remind your teachers that the tests were designed in accord with the goals and objectives as set forth in the *Teacher Handbook*.

- Point out that the chief purpose of the session is to validate the content of the 1989 North Carolina Test of Biology.

- Emphasize that the test items have previously undergone teacher reviews, field tests, and statistical analyses; this review is a final test review.

Next, outline your agenda for the session. (A suggested agenda is provided below.) Each teacher should examine the tests, work the items, and complete the questionnaire independently. Ask the teachers not to discuss the tests during the lunch break. Group discussion about the tests should be postponed until after the questionnaires are completed.

After an overview of the agenda, review with your teachers the nature of the task:

- The 1989 North Carolina Test of Biology consists of five forms. These items are the same on all five forms: [redacted] Each teacher should first read any one of the forms and answer the test items, circling the answer on the test form. Then, the same should be done with the other forms, but without repeating the common items listed above.

- During this time, each teacher is asked to reflect independently upon the overall content of the tests.

- After completing the tests, each teacher is asked to evaluate the content by completing the questionnaire independently.

- Teachers may feel free to mark on their tests during the review process.

You may wish to discuss the items from the questionnaire with your teachers. Answer any questions. Then pass out the tests. Please point out to the teachers that the test booklets provided for their review are in-house xerox copies. Final test booklets will be professionally printed and bound.

During the time that your teachers are working through the tests and completing the questionnaires, you are asked to examine each form of the 1989 North Carolina Test of Biology and to complete the special **Questionnaire for Curriculum Supervisor Evaluation of the 1989 North Carolina Test of Biology** (to be distinguished from the **Curriculum Supervisor's Summary of Teacher Reviews**). Your own evaluations may be at a more general level than those of the teachers. Feel free to comment on the overall layout and to make any suggestions that you think would be helpful.

The Afternoon Session

After the teachers have independently completed the tests and the questionnaires, elicit their opinions and comments within each of the questionnaire categories. If there is diversity of opinion on the questionnaire items, see if it is possible to reach some common ground. Record the outcome of this discussion on the special form entitled **Curriculum Supervisor's Summary of Teacher Reviews**. Do not, however, have the teachers change their own questionnaire responses.

At the end of the session, thank your teachers for their participation. Assure them that their opinions and comments will be carefully studied by Research and Testing Services and the Division of Science.

Test Security

In order to insure test security, please keep the tests in a locked location both before and after the period of the review. At the conclusion of the session, collect the tests and be sure that every copy is accounted for. Return these materials and the completed questionnaires in the envelope provided. Remind your teachers not to discuss the contents of the tests with anyone after the review. The items must remain secure if the tests are to serve their intended purpose.

Suggested Agenda

9:00-9:15	Provide introductory comments; state the purpose of the session; present the plan for the session.
9:15:9:30	Discuss and review the task.
9:30-12:00	Independently work through three forms of the test.
12:00-1:00	Break for lunch.
1:00-3:00	Independently work through the two remaining forms of the test and complete the questionnaire.
3:00-3:30	Discuss evaluations and attempt to reach consensus.

Procedures for Instructional Review of the 1989 North Carolina Test of Biology

Introduction

During the next two weeks key curriculum supervisors and teachers will meet across North Carolina to review the 1989 North Carolina Test of Biology. The major purpose of these reviews is to verify that the test accurately reflects the Biology curriculum. You will be asked to evaluate the extent to which the test reflects the goals and objectives of the subject matter, both with respect to the formal curriculum and as the subject is taught at your particular school.

Background

You may recall that extensive collaboration between curriculum specialists and classroom teachers led to the specifications for the North Carolina Biology curriculum. Based on a statewide survey of Biology teachers, goals and objectives were written for the curriculum. These curriculum specifications, officially adopted by the State Board of Education in the *Standard Course of Study* and the *Teacher Handbook*, form the basis for the 1989 North Carolina Test of Biology.

The items for the 1989 North Carolina Test of Biology were written by professional educators, primarily classroom teachers. These items were subjected to a rigorous series of reviews by psychometricians, technical specialists, curriculum advisors, and Biology teachers. The (revised) items were field tested with representative samples of North Carolina students. These field test data formed the basis for extensive statistical analyses. The various steps in the item development process may be enumerated as follows:

- Items are written in accord with the goals and objectives of the North Carolina Biology curriculum.
- Psychometric, technical, and editorial specialists review the items, modifying wording and format as needed.
- Curriculum specialists review the items to confirm that each measures the intended goal and objective.
- Teachers across the state review the items to confirm that they are appropriate for the target age level and that they conform to the specified goals and objectives.
- Items are field tested with representative samples of North Carolina students.

- Classical item analyses are performed on field test data.

- Item analyses are performed under the Rasch model of item response theory.

At this point in the test development process, a great deal is known about each item. Available statistics include the proportion of students answering the item correctly, the frequency with which each answer alternative is selected, the correlation between success on the item and the total test score, and the partial correlation between success on the item and group membership (ethnic, sex) with total scores partialed. Based on this information, 25-50% of the items are eliminated. From the remaining pool, items that meet certain criteria are selected for inclusion in the test.

We are presently entering the last phase in the test development process. It is now appropriate for professional educators familiar with the Biology curriculum at the classroom level to review the tests as collections of items. You and other members of this review panel are asked to provide a final check on general aspects of the 1989 North Carolina Test of Biology before its statewide administration.

Prior to the Test Review

You are undoubtedly familiar with the goals and objectives of the Biology curriculum referred to above. These curriculum guidelines are attached for your convenience. It may be helpful for you to review them briefly. If you have questions or comments concerning the relation of the guidelines to your own Biology curriculum, discuss these with your curriculum supervisor prior to the date of the test review.

The Plan for the Test Review

On the day of the review, you will be asked to read through the five forms of the 1989 North Carolina Test of Biology. These items are the same on all five forms: [redacted] First, read any one of the forms and answer the test items, circling the answer on the test form. Then, the same should be done with the other forms, but without repeating the common items listed above. As you read the items, you will be asked to reflect particularly upon the test content as it relates to the Biology curriculum at your school.

You will then be asked to complete a short questionnaire, a copy of which is included in this packet. The questionnaire is designed to elicit your evaluation and comments within five content categories. Each of these categories is explained below.

The Questionnaire

Item 1. The first two items on the questionnaire elicit your evaluation of content validity. As noted above, the 1989 North Carolina Test of Biology has been systematically designed to be in accord with the specific goals and objectives given in the *Basic Education Plan* and *Teacher Handbook*. Item-objective congruence has been confirmed both for the items and for each test as a whole by curriculum specialists. We would like your view on the degree to which this effort has succeeded. Remember, we are concerned here with the goals and objectives of the formal curriculum as detailed in the enclosed list of goals and objectives.

Item 2. The second evaluation of content validity concerns the degree to which the test reflects the goals and objectives of the Biology curriculum at your school. Differences are usually a matter of accent, with the school curriculum reflecting differing student abilities and interests. In some instances, however, there may be large differences that should be noted. While you are reading through the items, reflect upon the extent to which the tests as collections of items are in accord with the goals and objectives of the Biology curriculum at your school.

Item 3. In the third item on the questionnaire, you will be asked to indicate the extent to which, in your opinion, the vocabulary and linguistic style of the 1989 North Carolina Test of Biology is appropriate to the target age level. An important characteristic of a test designed to assess achievement within a particular subject area is that it is relatively insensitive to individual differences within other subject areas. The items on the 1989 North Carolina Test of Biology have been carefully edited for clarity and conciseness and for vocabulary usage appropriate to the target developmental group. You, as educators, however, are perhaps most thoroughly familiar with the verbal proficiency of your students. As you read through each form of the test, reflect upon the linguistic aspects in relation to your students' abilities.

Item 4. In the fourth item on the questionnaire, you will be asked to evaluate the extent to which the content of the 1989 North Carolina Test of Biology is balanced in relation to ethnicity, race, sex, socioeconomic status, and geographic district of the state. The items selected for the test have been carefully screened to eliminate bias. Panels of educators from each region of the state have reviewed the items to assure that none are objectionable to minorities or women. A series of sophisticated statistical analyses have been performed to identify items favoring one or another group. Where appropriate, such items have been eliminated. It is still possible, however, that taken as a whole a particular form of the test might exhibit bias. For example, the test content might differentially emphasize one or another geographic district or the artwork might differentially picture males or females. Please be sensitive to the possibility of overall bias as you read through the tests.

Item 5. The final test item on the questionnaire elicits your evaluation of the test at the item level. It is, of course, important that each item on a multiple choice test has a single answer that is clearly best. Equally important, although frequently overlooked, is that each item has distracter responses that are plausible. A student who has no knowledge of an objective represented by a particular item should find each of the answer choices equally inviting. As noted above, all test items have been subjected to a rigorous series of statistical analyses; among these analyses were item response analyses. Student responses to each alternative of every item have been considered in the item selection process. You are asked to provide a final check that each test item has a single answer that is best and distracter choices that are plausible.

As you read through the tests and reflect upon these various characteristics, keep in mind that it is at the whole-test level, rather than at the item level, that we elicit your comments. You may, of course, wish to illustrate a point with individual items. Also, if you have a problem with a particular item as you go through the test, make a note in the margins of the test. These comments will be compiled for consideration.

Test Security

Feel free to write on your copies of the tests during the review process, but please do not take the tests from the review room. The tests are secure. They will be collected at the end of the review session for return. It is imperative that you not discuss the contents of the 1989 North Carolina Test of Biology with others after the review. The items must remain secure if the test is to serve the intended purpose.

Questionnaire for Curriculum Supervisor Evaluation of the 1989 North Carolina Test of Biology

Directions: Indicate the extent to which, in your opinion, the 1989 North Carolina Test of Biology achieves the characteristics described below. Your comments are welcomed.

1. The test content reflects the goals and objectives of the Biology curriculum as <u>outlined on the enclosed list</u> of Biology Goals and Objectives.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

____

____

____

____

____

2. The test content reflects the goals and objectives of the Biology curriculum as Biology is <u>taught in my school system</u>.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

____

____

____

____

____

3. The items are clearly and concisely written, and the vocabulary is appropriate to the target age level.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

____

____

____

____

____

4. The content is balanced in relation to ethnicity, race, sex, socioeconomic status, and geographic district of the state.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

5. Each of the items has one and only one answer that is best; however, the distracters appear plausible for someone who has not achieved mastery of the represented objective.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

Other Comments

Curriculum Supervisor

LEA

Curriculum Supervisor's Summary of Teacher Reviews of the 1989 North Carolina Test of Biology

Directions: For each questionnaire item on which there were substantial differences of opinion, please give the nature of the disagreement and tell whether it was resolved during discussion.

1. The test content reflects the goals and objectives of the Biology curriculum as <u>outlined on the enclosed list</u> of Biology Goals and Objectives.

   Were there substantial differences of opinion on this item?

   ( ) No ( ) Yes (please explain)

   ____

   ____

   ____

   ____

   ____

2. The test content reflects the goals and objectives of the Biology curriculum as Biology is <u>taught in my school</u>.

   Were there substantial differences of opinion on this item?

   ( ) No ( ) Yes (please explain)

   ____

   ____

   ____

   ____

   ____

3. The items are clearly and concisely written, and the vocabulary is appropriate to the target age level

   Were there substantial differences of opinion on this item?

   ( ) No ( ) Yes (please explain)

   ____

   ____

   ____

   ____

   ____

4. The content is balanced in relation to ethnicity, race, sex, socioeconomic status, and geographic district of the state.

Were there substantial differences of opinion on this item?

( ) No ( ) Yes (please explain)

5. Each of the items has one and only one answer that is best; however, the distracters appear plausible for someone who has not achieved mastery of the represented objective.

Were there substantial differences of opinion on this item?

( ) No ( ) Yes (please explain)

Other Comments

Curriculum Supervisor

LEA

Questionnaire for Teacher Evaluation of the 1989 North Carolina Test of Biology

Directions: Indicate the extent to which, in your opinion, the 1989 North Carolina Test of Biology achieves the characteristics described below. Your comments are welcomed.

1. The test content reflects the goals and objectives of the Biology curriculum as outlined on the enclosed list of Biology Goals and Objectives.

   ( ) To a superior degree
   ( ) To a high degree
   ( ) To an average degree
   ( ) To a low degree
   ( ) Not at all

2. The test content reflects the goals and objectives of the Biology curriculum as Biology is taught in my school.

   ( ) To a superior degree
   ( ) To a high degree
   ( ) To an average degree
   ( ) To a low degree
   ( ) Not at all

3. The items are clearly and concisely written, and the vocabulary is appropriate to the target age level.

   ( ) To a superior degree
   ( ) To a high degree
   ( ) To an average degree
   ( ) To a low degree
   ( ) Not at all

4. The content is balanced in relation to ethnicity, race, sex, socioeconomic status, and geographic district of the state.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

5. Each of the items has one and only one answer that is best; however, the distracters appear plausible for someone who has not achieved mastery of the represented objective.

( ) To a superior degree
( ) To a high degree
( ) To an average degree
( ) To a low degree
( ) Not at all

Other Comments

Reviewer

LEA

Appendix H

Technical Manual

NC TESTS

Technical Characteristics of the

North Carolina Test of Algebra I

Forms A-O

North Carolina Department of Public Instruction
Division of Research\Raleigh, NC 27603-1332

Published 1988

Foreword

NCDPI Division of Research, in cooperation with the NCDPI Instructional Services, has developed diagnostic achievement tests of basic skills for public school students in Grudes 3, 6, and 8; survey achievement tests of Science and Social Studies for students in Grades 3, 6, and 8; and end-of-course achievement tests for students taking Algebra I, Algebra II, Biology, and US History. Chemistry and Geometry achievement test will be added in 1989, and other tests are being planned. [a]

To facilitate the proper technical use of the test scores obtained from the administrations of the tests, the curricular and psychometric characteristics of the tests will be described in a series of technical manuals. This manual, the first in the series, contains a description of the characteristics of the North Carolina Test of Algebra I.

[a] Readers who have an interest in the origins of the test development program are referred to the North Carolina Elementary and Secondary School Reform Act of 1984, the North Carolina Basic Education Program, the North Carolina Standard Course of Study, and the Teacher Handbook.

CONTENTS

Foreword ii

List of Tables iv

List of Figures v

DESCRIPTION 1

VALIDITY 2

METHOD FOR DERIVING TEST SCORES 5

RELIABILITY AND OTHER STATISTICS 7

CURRICULAR ASSESSMENT 15

CONTENT OF THE TEST 16

TEST NORMS 26

APPENDIX 28

LIST OF TABLES

1. Organization of the North Carolina Test of Algebra I 1
2. Descriptive Statistics for the NC Tests of Algebra I Cores 8
3. Test Content - Item Representation by Goal and Objective (1987)...................17
4. Item Difficulty by Item Number for 1986 and 1987 NC Test of Algebra I........22
5. Norms Table for Student Scores on the NC Test of Algebra I 27

LIST OF FIGURES

1. Comparison of letter grades teachers expected students to receive and scores subsequently earned on the 60-item core of the NC Test of Algebra I (N = 52,648 students) 6
2. Frequency distribution of the 1986 NC Test of Algebra I 60-item core 9
3. Equipercentile comparison of the 1986 NC Test of Algebra I core A (administered statewide) with core test 6 10
4. Equipercentile comparison of the 1986 NC Test of Algebra I core A (administered statewide) with core test 7 11
5. Equipercentile comparison of the 1986 NC Test of Algebra I core A (administered statewide) with core test 8 12
6. Equipercentile comparison of the 1986 NC Test of Algebra I core A (administered statewide) with core test 9 13
7. Equipercentile comparison of the 1986 NC Test of Algebra I core A (administered statewide) with revised core test 8 14

DESCRIPTION

The North Carolina Test of Algebra I was developed for use as an end-of-course test following completion of the Algebra I course of study. Its design serves a dual purpose: that of a normative measurement of student achievement, and of an objective-based measurement of curriculum coverage.

The measurement of student achievement is attained by administering a basic core of 60 items to all students. The measurement of curriculum is met by the addition of 35 items that vary across five forms of the test. All five forms of the test, each form containing the same 60 core items and 35 different items, are administered in each classroom, one form per student. Under this system, 20% of the students in a classroom will take Form A of the test, 20% Form B, and so on (see Table 1). The normative student scores are based on the 60 items all forms have in common. Curriculum assessment is achieved by combining the results from all five forms, which provide an assessment based on the 60 core items + 5(35 items), or 235 items in all.

Table 1
Organization of the North Carolina Test for Algebra I

<table>
<tr><td colspan="5">60 core items</td></tr>
<tr><td>35
variable
items</td><td>35
variable
items</td><td>35
variable
items</td><td>35
variable
items</td><td>35
variable
items</td></tr>
<tr><td>95 items
Form A</td><td>95 items
Form B</td><td>95 items
Form C</td><td>95 items
Form D</td><td>95 items
Form E</td></tr>
</table>

VALIDITY

The purpose in developing an Algebra I achievement test is to obtain test scores from which inferences may be drawn concerning the degree of success a particular student, classroom, school, or school district has had in mastering the Algebra I curriculum, and the degree to which the curriculum has been mastered by students in the aggregate. To the extent this can be done meaningfully, test scores may be said to be valid. Thus one inference drawn from a test score may be valid, while another inference may not be valid.

Theoreticians insist correctly that only inferences concerning test scores can be said to have validity. Generally, readers understand this, and here the convenient shorthand will be employed of speaking about "test validity" rather than "inferences about achievement drawn from scores obtained from tests."

Test validity is a predominant theme in test development, from the time the idea for a test is conceived until the final test scores have been analyzed and interpreted. For convenience, the various components of test validity will be described as if they were unique, independent components rather than interrelated parts. The first component of test validity to be described will be curricular validity.

Curricular validity. If a test is to be used to measure the degree to which a course of study has been mastered, the first step is to define the curriculum. In the case of Algebra I, that was done through a cooperative effort, led by the NCDPI Instructional Services, involving curriculum specialists, teachers, administrators, university professors, and others. The result was a list of 12 goals encompassing 111 objectives. Supported by expert opinion and a statewide consensus, these goals and objectives were approved by the State Board of Education as the basis for instruction in Algebra I. Curricular validity, the first step in establishing construct validity, was established by this method.

Instructional Validity. A basic course of study may not include all of the objectives taught under various circumstances in Algebra I. For example, some advanced classes may cover material that would be beyond the reach of 95% of all Algebra I students. For that reason and several others, it becomes important to know just what is being taught in the majority of Algebra I courses in the state. To determine this, a survey was made of all Algebra I teachers (N = 1,056 teachers). The analysis of results was based on 981 responses, or 93% of all possible responses.

The Algebra I teachers examined all 111 objectives and noted whether they taught each objective every year, and whether each objective was basic to Algebra I instruction. The two answers turned out to be equivalent: if an objective was taught every year, it was considered to be basic. The majority of teachers judged somewhat more than 80 of the 111 objectives as basic. After deliberation by curriculum specialists and the North Carolina Competency Test Commission, whose duty it was to advise the NC Board of Education on statewide testing, it was decided that 87 of the objectives formed the basic curriculum for Algebra I. The objectives that were accepted are given in Table 3, together with the proportion of teachers that judged each one as basic. Similar information is given in the Appendix for the objectives that were rejected.

Instructional validity, the second step in defining construct validity, was established by these procedures. It limits inferences drawn from Algebra I test scores to the basic instructional program comprising the 87 objectives.

In summary, it was concluded that curricular and instructional validity depended jointly on the 87 objectives and the 12 goals under which they were collected, and that the Algebra I test should be built on that foundation. (Items created for the additional 24 objectives were not field tested, but were retained as a separate item pool for possible use at a later time.)

Content validity of the item pool. Content validity—the degree to which test items reflect the basic instructional program—was defined through a number of operations:

First, twelve items were created for each objective by classroom teachers who were trained for two days in the technical aspects of item-writing. The use of classroom teachers helped to insure that instructional validity was maintained, since their items would be drawn from their classroom experiences.

Second, the item pool was edited for grammar, syntax, psychometric form, and linguistic bias.

Third, the item pool was analyzed by curriculum specialists to assure that the items were valid representations of the objectives for which they were written.

Fourth, the items were collected into test forms for tryouts. Although the forms were not the final forms for the Algebra I tests, they were organized in such a way that the objectives were represented equitably across all forms. Each test contained an average of 80 items, 15 of which were common across all test forms for purposes of ability equating should that become necessary.

Fifth, test administration instructions were written, test distribution procedures were organized, and administrators were trained to conduct the test administration. The experienced test administration organization used to administer statewide tests in North Carolina was employed to accomplish the testing. The procedure followed the routine eventually expected to be used to administer the test forms when the test of record was given.

Sixth, a sample of 8,500 students was selected to take the 16 tryout test forms containing a total of 1,284 items. To insure broad representation, four schools were selected from each of the eight North Carolina educational regions. All Algebra I students in each school took one form of the tryout test. Each test form was administered in two regions. Each item was answered by approximately 530 students.

Seventh, the field test data were analyzed using both the classical psychometric model and the one-parameter Rasch model (Bical program). Sixteen statistics were assembled for each item, i.e., p-value, Rasch difficulty index, Adjusted Rasch difficulty index, standard error of the mean, fit mean-square, item validity (point-bisearl correlation), and the item characteristic curve groupings.

The item statistics were submitted to computer analysis using a program designed to scan a range of statistics and print out an appropriate decision based on the criteria that had been built into the program. For example, if an item was answered on a chance level, the computer would print the comment, "Too hard." If the item characteristic curve exhibited irregularities, that fact was noted. And so on.

These notations were reviewed and decisions made about the psychometric adequacy of the items. The decisions were then conveyed to curriculum specialists, who also reviewed the items and reached a decision about their curricular adequacy.

The content of Algebra I cannot be represented by a single factor. Therefore, the maximization of item-total correlations was not a goal of item development. Once an item was shown to have at least a modest correlation with a corrected total score and was judged to measure an objective, it was included in the item pool. While this may have reduced the potential internal reliability as measured by coefficient alpha, it increased the validity of the test by allowing for an objective factor structure that was not expected to be unitary.

This information was then placed on an item record, which became the basic document to which all other records were referred. The item record contains the goal, objective, historical information, a copy of the item itself, the test tryout statistics, and the psychometric and curricular decisions concerning the item's suitability for use in a test. Each item has a separate item record.

Of the field-tested items, 29% were judged to be too difficult or too easy for Algebra I students, or had other obvious flaws. In content areas, these items are usually discarded. In Algebra I, it was felt that an analysis of defective items was possible, and that the curriculum specialist could revise the items to bring them within a usable range of values. Thus in Algebra I a revised item pool, as well as the statistically defined item pool, was available for use in test construction. The revised items were employed randomly throughout the selection of items, in order to assure that their effects would fall equally among all conditions. This, however, was the general method used to select all items (to be discussed next).

Content validity of the test. After a consideration of the logistics involved, it was decided to prepare one complete test for administration in May 1986, and to test four additional core tests of 60 items each for use in succeeding years. Core tests were based on a random selection of 60 objectives, each objective represented by one test item randomly chosen from the approved item pool. Thus the content of the test cores directly reflected all of the decisions that had been made earlier.

This method of item selection is a modified domain sampling model, with the various forms and cores randomly equivalent. The domain sampling model in its pure form is highly inefficient because it allows the entry of items that are grossly inappropriate for normative measurement—items that no one can answer or that everyone can answer, or that have psychometric deficiencies of a more complex form. In the modification used here, the domain of items was limited to those items that had satisfactory psychometric characteristics. This was determined by the analysis of the item tryout data, which was used only to verify the psychometric adequacy of the item pool and to direct where item revisions should be made.

Although the initial equating of the core tests depended upon random selection of items from the pool, the final equating was based on statistics obtained at the time the first test of record was administered. This second psychometric analysis, described next, was used to eliminate random differences among cores and thus to facilitate the precision of measurement from one year to another.

Standardization sample. The first NC Test of Algebra I of record consisted of five forms, each form containing the same 60 core items and a unique set of variables items. This test was administered to 63,330 North Carolina Algebra I students in May, 1986. The state norm population comprises these 63,330 students.

At the same time, four additional core tests of 60 items each were administered as separate forms to 2,400 students (600 students per form). These students attended schools selected to be representative of the state on the basis of criteria that were judged to be related at least partially to Algebra I ability levels—school performance on the California Achievement Test, for example. (Prior to the first NC Test of Algebra I of record, no comparative information existed on Algebra I achievement or class ability composition.)

The four future core tests were interleaved in all student samples. This produced an even spread of ability across all four tests. The agreement of mean test scores on all four future core tests and the agreement of these mean scores with the state norm mean for the first test of record (Table 2) support the view that the samples were representative of all North Carolina students.

Concurrent validity. When the 1986 NC Test of Algebra I of record was administered, Algebra I teachers were asked to indicate the expected final letter grade for each student in their classes. Figure 1 displays a comparison of letter grades and the mean Algebra I test score corresponding to each letter. The results of this test for concurrent validity conform closely to expectations and contribute to the validity of inferences concerning student achievement as measured by the Algebra I test.

METHOD FOR DERIVING TEST SCORES

Item information was available to support both the classical scoring model and the Rasch scoring model. The classical scoring model gives a unitary weight to each item; a correct choice adds 1 to the total score, an incorrect choice adds 0. The one parameter Rasch model also uses unitary weighting. (Two- and three-parameter item response models give more credit for answering some items correctly, and less credit for answering other items correctly. These models assume that each item has a fundamental, unchanging difficulty level.) The classical model was utilized to score the NC Test of Algebra I test because it was fundamentally sound, simple to use, and easy to interpret.

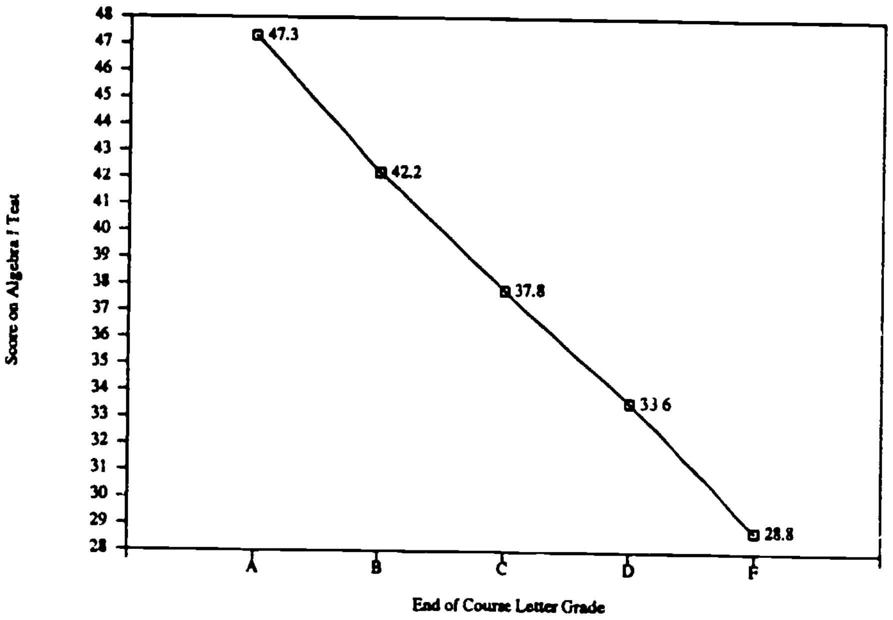


Figure 1. Comparison of letter grades teachers expected students to receive and scores subsequently earned on the 60-item core of the North Carolina Test of Algebra I (52,648 students).

RELIABILITY AND OTHER STATISTICS

The descriptive statistics, the standard errors of measurement, alternate form reliability, and the alpha coefficients for all five core tests are given in Table 2. The alternate form reliabilities are approximately .83, the alpha reliabilities .39. The scores are symmetrical about a mean score of 63% correct (Figure 2).

Of especial significance to comparability of student scores across the years is the equivalence of the four future core tests to the first core test of record. An equipercentile analysis was made of the relationship of the four future core tests to the 1986 core test of record. To make the equipercentile comparison, the mean of a block of scores within successive five percentile points on the first core test of record was taken to compare with the mean of the block of scores within the same successive five percentile points on the second core test. This yielded 20 reasonably reliable points of comparison. The results are displayed in Figures 3 through 6.

In Figures 3 through 6, the differences of the data points from a slope of 1.00 are small. These differences could be adjusted statistically by providing a separate set of norms for each form. Just as easily, and far more efficiently, the core tests could be redeveloped slightly so that even the small differences disappeared. With this technique, a single norms table could be used for all five core tests. To accomplish this transformation, the test developer had available the 60 items from the first core test of record plus 200 items from the variable set of items from the first test of record—a total of 260 items for which comparable psychometric data were available across all four future core tests. This made the task of adjusting the future cores tests to the profile of the first core test a relatively simple matter.

The results of the adjustment for core test 8, employed in 1987, are given in Figure 7. The required changes were so small that the alternate form and alpha reliabilities did not change from their respective values of .84 and .90. The overall means and standard deviations of the two tests are identical to three digits.

Similar adjustments based on the 1986 administration will be made to the third, fourth, and fifth core tests as needed in the future. This will assure the continuity of the norms table for the next three years while providing new test items each year The new test items will prevent the loss of test confidentiality, and therefore validity, that occurs with the continued use of the same items. Student scores will have a common reference point from 1986 to 1990, barring changes in the definition of the basic instructional program.

TABLE 2

Descriptive Statistics for the NC Test of Algebra I Core

Test		Mean	Median	SD	SE_{meas}	Alternate Form Reliability	Coefficient Alpha Reliability
A	Raw Score	37.7	38	9.3	3.04	0.84,0.84,0.82,0.82	0.89
	Proportion	0.63	0.63	0.15	0.05		
6	Raw Score	38.1	38	9.7	3.13	0.84	0.90
	Proportion	0.63	0.63	0.16	0.05		
7	Raw Score	37.1	39	9.8	3.21	0.82	0.89
	Proportion	0.62	0.65	0.16	0.05		
8	Raw Score	38.0	38	9.4	3.24	0.84	0.88
	Proportion	0.63	0.63	0.15	0.05		
9	Raw Score	36.6	39	10.0	3.26	0.82	0.89
	Proportion	0.61	0.65	0.17	0.05		

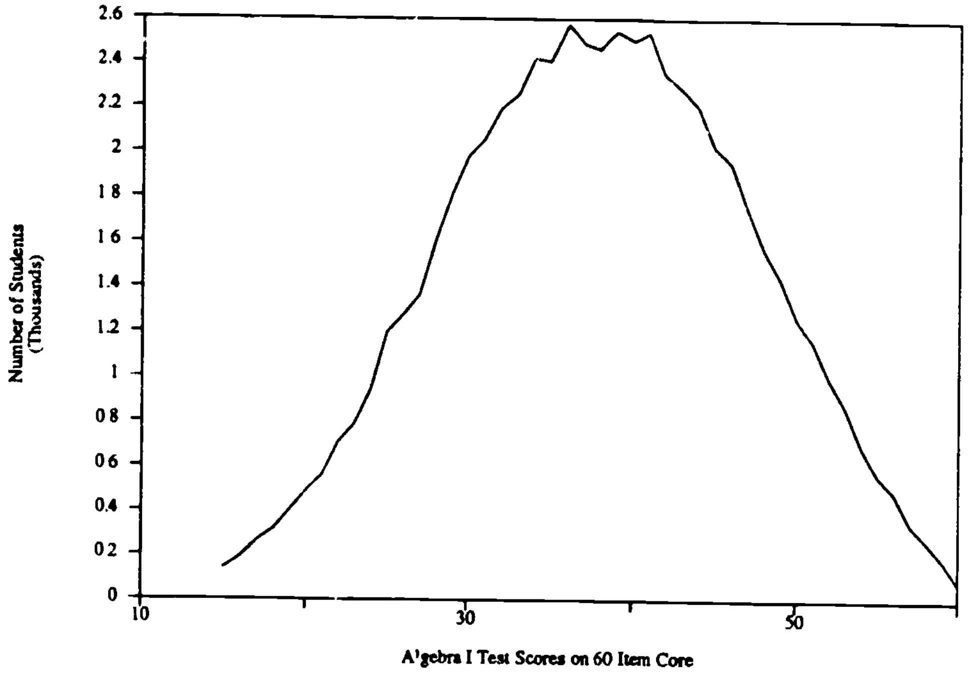


Figure 2. Frequency distribution of the 1986 Algebra I 60-item Core Test.

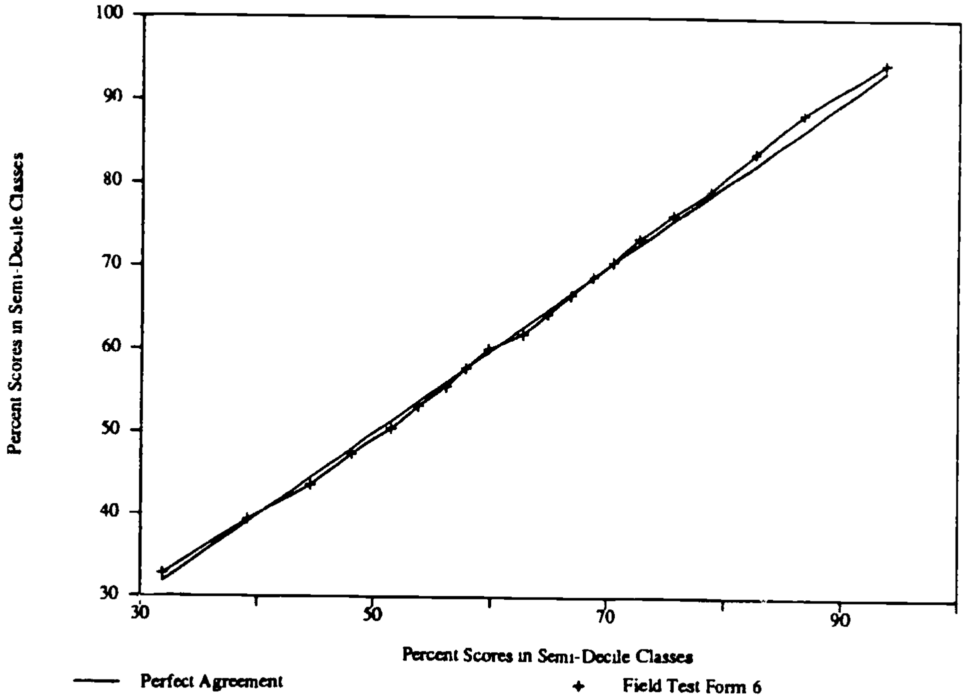


Figure 3. Equipercentile comparison of the 1986 Algebra I Core Test A (administered statewide) with Core Test 6.

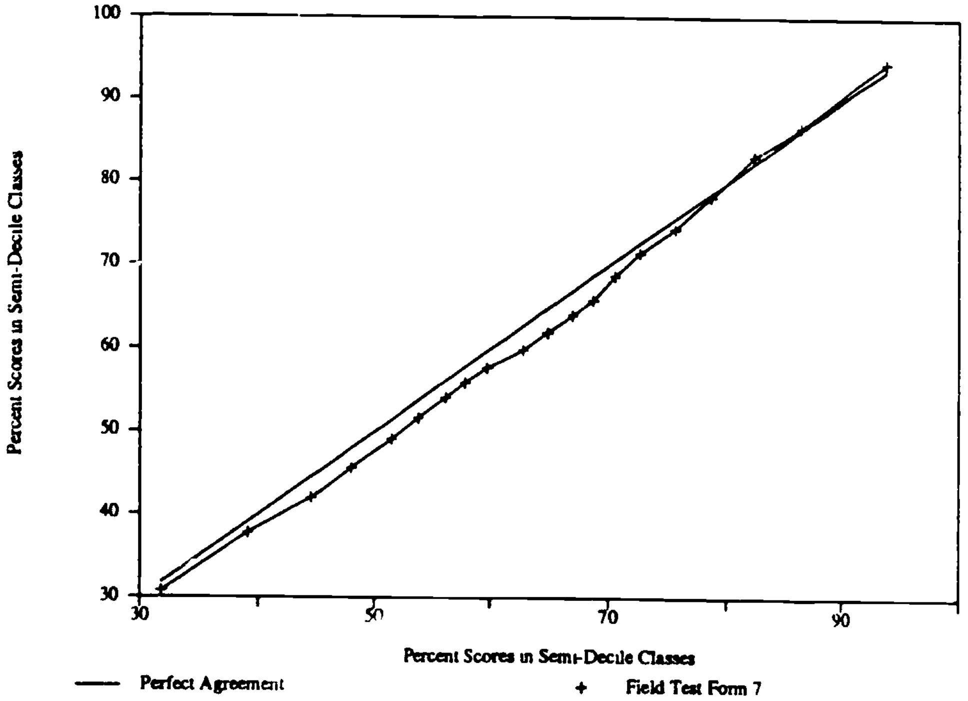


Figure 4. Eqı ˌercentile comparison of the 1986 Algebra I Core Test A (administered statewide) with Core Test 7.

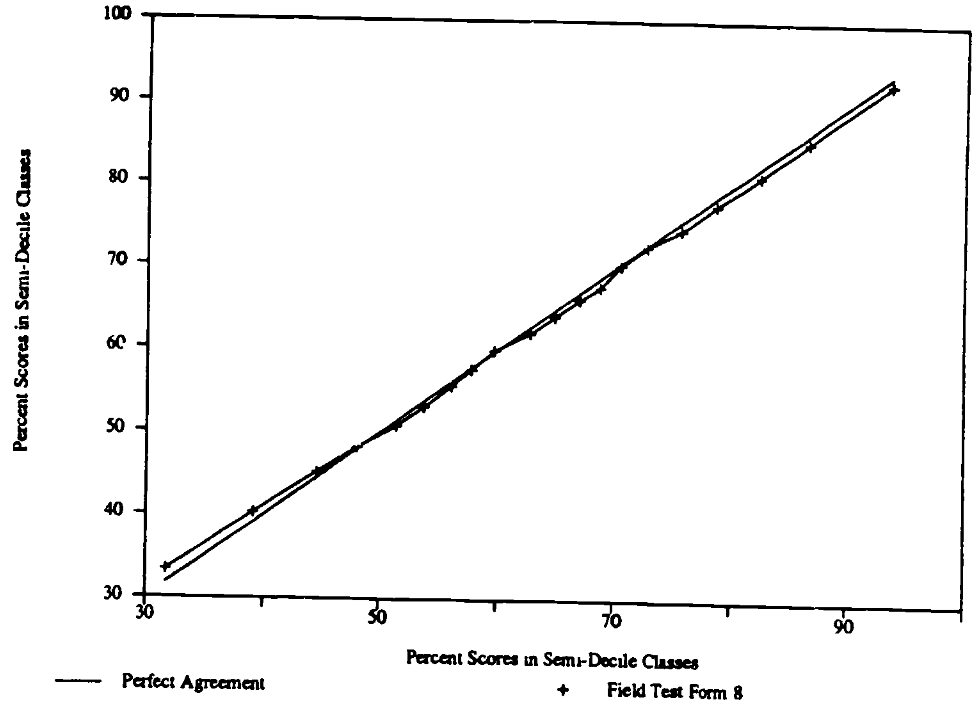


Figure 5. Equipercentile comparison of the 1986 Algebra I Core Test A (administered statewide) with Core Test 8.

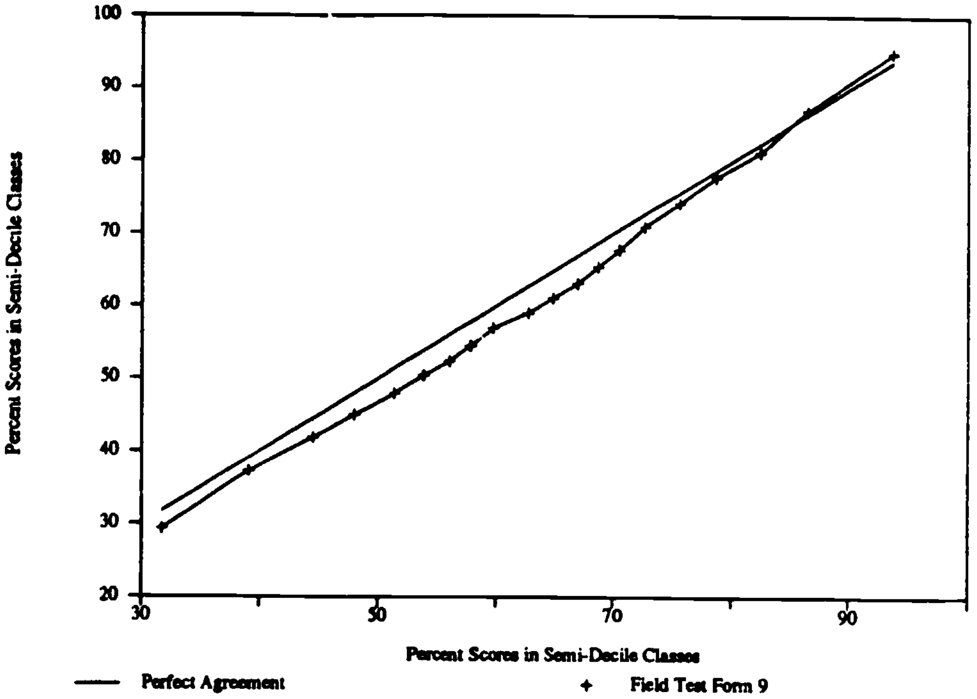


Figure 6. Equipercentile comparison of the 1986 Algebra I Core Test A (administered statewide) with Core Test 9.

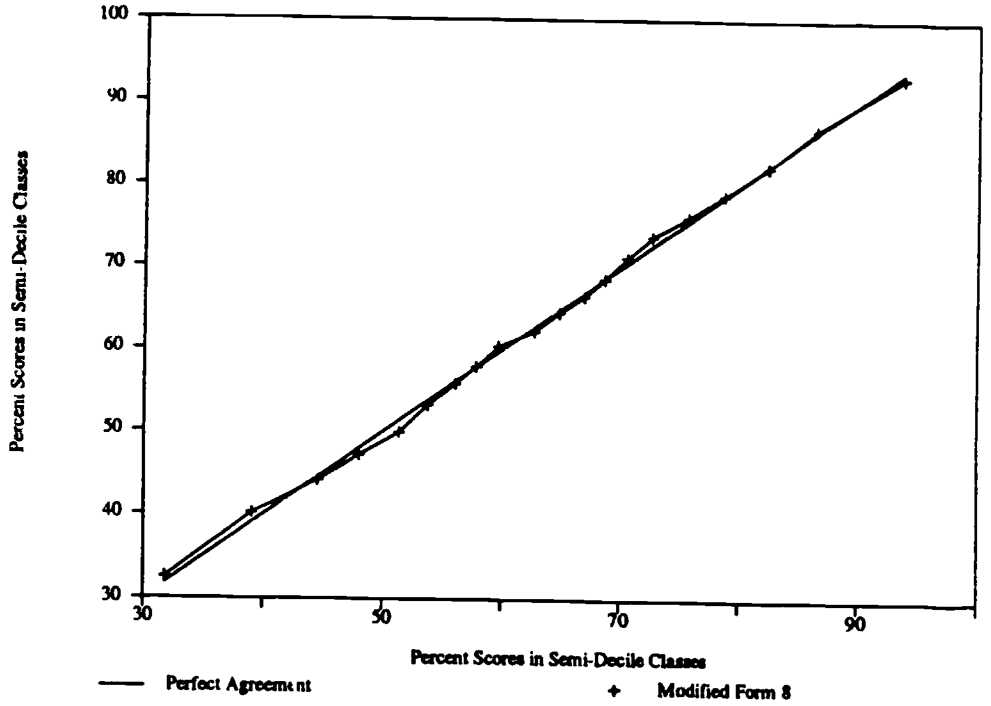


Figure 7. Equipercentile comparison of the 1986 Algebra I Core Test A (administered statewide) with the revised Core Test 8.

CURRICULAR ASSESSMENT

For purposes of exposition, the 60-item NC Test of Algebra I core tests were discussed above as if they made up the entire test. In the 1986 and 1987 core tests of record, the core tests of 60 items were accompanied by additional items that varied across five forms (as discussed earlier—see Table 1). These items were not intended to contribute to individual student scores, but to curriculum assessment. Each item was to be answered by one-fifth of the students.

At the classroom level, 235 items will be answered in 1987 by an average of 5 students each, which will provide a data base of three items per objective across five students. This will permit an estimate of how various portions of the curriculum are being mastered in the classroom. At the school, school district, or state level, the 235 items will be answered by larger numbers of students: up to 13,000 students per item. This assures a more stable measurement, but does not of course include a larger number of objectives or items. That accumulation depends upon measurement across successive years.

The measurement afforded by the 175 variable, non-core items, while they cannot be referenced to individual students, are critical in assessing curriculum mastery at the classroom, school, school district, and state levels. Each year adds to the data and gives a more and more detailed picture of curricular success. During one year, each objective is measured by three items; the second year, by three more items; and so on, year after year.

In summary, the utility of the test is in its initial norms table, its statistical equivalence of core tests from year to year, and its broad sampling of the curriculum across time.

CONTENT OF THE TEST

The North Carolina Test of Algebra I is objective-referenced; that is, its reference is to a domain of objectives. This domain is mapped over by a domain of items. The items reflect the objectives, equal in kind and number except for random fluctuations.

The first year, a different pattern was followed for selecting the variable items, which would have permitted the variable item scores to be added to the student test scores to increase the reliability of the student test scores. This would have required statewide norms for the variable items. These could not be calculated in time to be of practical use, so the more tractable strategy of basing the student scores exclusively on the 60 core items was accepted. This freed up the variable items for their most efficient use in curriculum assessment. Subsequent programming of the variable items aimed toward the goal of even assessment across all objectives; in short, each objective was to be represented by the same number of items. This is consistent with the concept of a domain of objectives mapped over by a domain of items.

Although the objectives have unit weighting, the goals are weighted by the number of objectives assigned to them. As born out by empirical analyses, this is part of the natural history of curriculum development: the more important a goal is believed to be, the greater the number of objectives that will be developed for it. Thus an intrinsic system of weights exists for curricular goals.

Table 3 lists each goal and objective and the numerical item representation on the NC Test of Algebra I for the second year (which sets the pattern for future years). For the historical record, the items per objective for the first two years are also given.

Table 4 gives the difficulty level for all items on the NC Test of Algebra I tested in 1986 in terms of the proportion of all students answering the item correctly.

Table 3

Test Content - Item Representation by Goal and Objective (1987)

Goal	Objective - Description	No. Items 1987	No. Item 1986+87	% Teachers Rating as Basic[a]
Goal 1:	Use the language of Algebra.	17	35	
1.1	Simplify numerical expressions.	2	7	98
1.2	Evaluate variable expressions.	3	4	99
1.3	Evaluate exponential expressions	3	5	96
1.4	Use 'order of operations' to evaluate expressions.	3	5	98
1.5	Evaluate formulas when the replacement values are given.	3	5	91
1.6	Convert word phrases into symbols.	3	9	94
Goal 2:	Use the structural properties of number systems.	18	38	
2.1	Use the commutative property of addition to simplify expressions or computational processes with real numbers.	3	5	93
2.2	Use the associative property of addition to simplify expressions or computational processes with real numbers.	3	4	93
2.3	Use the distributive property of multiplication over addition to simplify expressions or computational processes withreal numbers.	2	7	97
2.4	Use the reciprocal, or multiplicative inverse, of a number to simplify expressions or computational processes with real numbers.	3	5	94
2.5	Use the commutative property of multiplication to simplify expressions or computational processes with real numbers.	2	8	93
2.6	Use the associative property of multiplication to simplify expressions or computational processes with real numbers.	2	4	93
2.7	Use the distributibve property to simplify expressions.	3	5	93
Goal 3:	Perform operations with rational numbers.	5	11	
3.1	Use < or > to compare two rational numbers	2	7	92
3.2	Express rational numbers in fraction or decimal form.	3	4	83

Table 3 (continued)

Test Content - Item Representation by Goal and Objective (1987)

Goal	Objective - Description	No. Items 1987	No. Items 1986+87	% Teachers Rating as Basic[a]
Goal 4:	Locate numbers on the number line or rectangular coordinate plane.	16	34	
4.1	Graph sets of real numbers on the number line.	3	4	96
4.2	Use the number line to add real numbers.	2	7	71
4.3	Graph ordered pairs of numbers on the coordinate plane.	2	5	92
4.4	Graph a relation of the coordinate plane.	2	7	80
4.6	Graph a linear equation in two variables.	3	6	80
4.7	Graph a line given its slope and y-intercept.	3	5	79
Goal 5:	Perform operations with real numbers.	30	62	
5.1	Determine the opposite, or additive inverse, of a number.	3	4	98
5.2	Find the absolute value of a number.	3	4	96
5.3	Use < or > to compare two numbers.	2	7	96
5.4	Add real numbers.	2	8	96
5.5	Subtract real numbers.	3	6	96
5.6	Multiply real numbers.	2	8	96
5.7	Divide real numbers.	3	5	96
5.8	Distinguish between rational and irrational numbers.	3	5	68
5.9	Find the square root of a number which is perfect square.	3	5	87
5.10	Use a calculator, table of square roots, or an algorithm to find a decimal approximation for the square root of a real number.	3	6	53
5.11	Find the union and intersection of two sets of real numbers.	3	4	73
Goal 6:	Solve linear equations.	32	67	
6.1	Find the solution set of an open sentence when replacement values are given for the variable.	3	5	94
6.2	Solve a simple equation by using the additive property of equality.	3	5	97
6.3	Solve a simple equation by using the subtraction property of equality.	2	7	96
6.4	Solve a simple equation by using the multiplicative property of equality.	3	5	97
6.5	Solve a simple equation by using the division property of equality.	2	7	96

Table 3 (continued)

Test Content - Item Representation by Goal and Objective (1987)

Goal	Objective - Description	No. Items 1987	No. Items 1986+87	% Teachers Rating as Basic[a]
6.6	Solve an equation by using more than one property of equality.	3	4	98
6.7	Solve an equation which contains similar terms.	2	7	98
6.8	Solve an equation which has the variable in both members.	3	5	97
6.9	Solve 'age', 'coin', and 'integer' problems.	3	5	72
6.10	Solve an equation in which the numerical coefficient is a fraction.	3	5	88
6.11	Solve problems involving percents.	3	5	67
6.12	Solve 'percent-mixture', 'investment', 'uniform motion', and 'rate-of-work' problems.	2	7	40
Goal 7:	Solve linear inequalities.	5	12	
7.1	Find the solution set for a linear inequality when the replacement values are given for the variables.	2	8	86
7.2	Solve a linear inequality by using transformations.	3	4	84
Goal 8:	Understand and solve systems of linear equations.	22	46	
8.1	Find the slope of a non-vertical line given the graph of the line, or an equation of the line, or two points on the line.	3	5	81
8.2	Write the slope-intercept form of an equation of a line.	3	5	81
8.3	Write the equation of a line given the slope and one point on the line, or two points on the line.	3	5	72
8.4	Find the solution set of open sentences in two variables when given replacement values for the variables.	3	5	73
8.5	Use a graph to find the solution of a pair of linear equations in two variables.	2	8	71
8.6	Use the substitution method to find the solution of a pair of linear equations in two variables	3	5	72
8.7	Use the addition or subtraction method to find the solution of a pair of linear equations in two variables.	2	8	75
8.8	Use multiplication with the addition or subtraction method to solve systems of linear equations.	3	5	7

Table 3 (continued)

Test Content - Item Representation by Goal and Objective (1987)

Goal	Objective - Description	No. Items 1987	No. Items 1986+87	% Teachers Rating as Basic[a]
Goal 9: Perform operations with polynomials.		49	102	
9.1	Add polynomials.	3	5	99
9.2	Subtract polynomials.	3	5	99
9.3	Multiply monomials.	3	5	99
9.4	Find an indicated power of a monomial.	3	5	97
9.5	Multiply a polynomial by a monomial.	3	5	99
9.6	Multiply two polynomials.	2	8	98
9.7	Factor a monomial	3	5	97
9.8	Divide two monoi	3	5	97
9.9	Divide a polynomial by a monomial.	3	5	96
9.10	Divide a polynomial by a binomial.	3	5	84
9.11	Find a common monomial factor in a polynomial.	3	5	97
9.12	Find the product of the sum and difference of two binomials.	3	5	95
9.13	Factor the difference of two squares.	2	8	96
9.14	Square a binomial without using long multiplication.	2	7	86
9.15	Factor a perfect square trinomial.	2	7	89
9.16	Find the product of two binomials.	2	7	97
9.17	Factor a quadratic trinomial when the coefficient of the quadratic term is one.	3	5	92
9.18	Factor a quadratic trinomial when the coefficient of the quadratic term is not one.	3	5	85
Goal 10: Solve quadratic equations.		11	23	
10.1	Solve a second degree equation when one member is in factored form and the other member is zero	3	5	90
10.2	Solve a second degree equation by factoring.	3	5	90
10.3	Use factoring to solve a verbal problem.	2	8	57
10.4	Solve a quadratic equation that is in the form: perfect square = constant.	3	5	58
Goal 11: Perform operations with algebraic fractions.		25	52	
11.1	Write an algebraic fraction in its simplest form.	3	5	93
11.2	Solve proportions.	3	4	85
11.3	Use ratios and proportions to solve problems.	2	7	71
11.4	Multiply algebraic fractions.	2	7	93
11.5	Divide algebraic fractions.	3	5	93
11.6	Simplify algebraic expressions involving multiplication and division of algebraic fractions.	3	5	88

Table 3 (continued)

Test Content - Item Representation by Goal and Objective (1987)

Goal	Objective - Description	No. Items 1987	No. Items 1986+87	% Teachers Rating as Basic[a]
	11.7 Add and subtract algebraic fractions.	3	5	89
	11.8 Change a mixed expression to an algebraic fraction and a fı action to a mixed expression.	3	5	70
	11.9 Solve fractional equations.	2	8	81
Goal 12: Simplify expressions which contain radicals.		5	13	
	12.1 Simplify products and quotients of radical expressions.	2	8	62
	12.2 Simplify sums and differences of radical expressions.	3	5	59

Table 4

Item Difficulty by Item Number for 1986 and 1987 NC Test of Algebra I

1986 Statewide Core - Form A

Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Diffiulty
1	0.46	16	0.83	31	0.32	46	0.74
2	0.66	17	0.90	32	0.43	47	0.74
3	0.80	18	0.71	33	0.38	48	0.74
4	0.89	19	0.76	34	0.28	49	0.73
5	0.91	20	0.91	35	0.65	50	0.59
6	0.76	21	0.97	36	0.68	51	0.57
7	0.54	22	0.54	37	0.50	52	0.27
8	0.93	23	0.89	38	0.56	53	0.83
9	0.84	24	0.42	39	0.56	54	0.50
10	0.75	25	0.56	40	0.53	55	0.53
11	0.97	26	0.84	41	0.80	56	0.48
12	0.86	27	0.44	42	0.87	57	0.41
13	0.43	28	0.55	43	0.46	58	0.33
14	0.56	29	0.47	44	0.59	59	0.33
15	0.90	30	0.61	45	0.49	60	0.17

1986 Variable Items

Item	Diff.	Item	Diff.	Item	Diff.	Item	Diff.	Item	Diff.
1	0.93	41	0.94	81	0.90	121	0.87	161	0.67
2	0.94	42	0.78	82	0.34	122	0.89	162	0.92
3	0.89	43	0.79	83	0.83	123	0.84	163	0.91
4	0.89	44	0.79	84	0 63	124	0.70	164	0.92
5	0.84	45	0.95	85	0.95	125	0.94	165	0.89
6	0.95	46	0.75	86	0.82	126	0.60	166	0.78
7	0.85	47	0.66	87	0.64	127	0.62	167	0.84
8	0.75	48	0.22	88	0.77	128	0.61	168	0.88
9	0.36	49	0.56	89	0.76	129	0.36	169	0.65
10	0.63	50	0.79	90	0.74	130	0.57	170	0.37
11	0.95	51	0.67	91	0.80	131	0.39	171	0.75
12	0.85	52	0.76	92	0.70	132	0.70	172	0.73
13	0.73	53	0.97	93	0.92	133	0.90	173	0.98
14	0.81	54	0.71	94	0.97	134	0.72	174	0.90
15	0.90	55	0.85	95	0.79	135	0.83	175	0.86
16	0.89	56	0.76	96	0.88	136	0.62	176	0.63
17	0.69	57	0.72	97	0.68	137	0.71	177	0.83
18	0.62	58	0.59	98	0.70	138	0.24	178	0.83
19	0.76	59	0.46	99	0.47	139	0.17	179	0.66
20	0.53	60	0.26	100	0.31	140	0.46	180	0.59

Table 4 (continued)

1986 Variable Items

Item	Diff.	Item	Diff.	Item	Diff.	Item	Diff.	Item	Diff.
21	0.47	61	0.35	101	0.45	141	0.29	181	0.68
22	0.58	62	0.41	102	0.46	142	0.48	182	0.53
23	0.54	63	0.41	103	0.60	143	0.51	183	0.35
24	0.37	64	0.52	104	0.61	144	0.62	184	0.36
25	0.60	65	0.53	105	0.45	145	0.64	185	0.57
26	0.85	66	0.60	106	0.71	146	0.47	186	0.47
27	0.46	67	0.48	107	0.73	147	0.80	187	0.67
28	0.50	68	0.70	108	0.44	148	0.55	188	0.65
29	0.69	69	0.57	109	0.39	149	0.44	189	0.36
30	0.38	70	0.82	110	0.86	150	0.72	190	0.57
31	0.32	71	0.72	111	0.81	151	0.65	191	0.70
32	0.57	72	0.46	112	0.75	152	0.60	192	0.59
33	0.55	73	0.63	113	0.70	153	0.71	193	0.37
34	0.58	74	0.60	114	0.68	154	0.31	194	0.23
35	0.61	75	0.39	115	0.26	155	0.64	195	0.56
36	0.57	76	0.57	116	0.56	156	0.49	196	0.81
37	0.62	77	0.74	117	0.40	157	0.65	197	0.56
38	0.28	78	0.30	118	0.38	158	0.46	198	0.29
39	0.25	79	0.45	119	0.45	159	0.29	199	0.40
40	0.31	80	0.43	120	0.40	160	0.35	200	0.35

Table 4 (continued)

Item Difficulty by Item Number for 1986 and 1987 Algebra I Tests

1987 Statewide Core - Form 8

Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
1	0.93	16	0.70	31	0.48	46	0.44
2	0.76	17	0.85	32	0.25	47	0.61
3	0.66	18	0.70	33	0.55	48	0.73
4	0.40	19	0.91	34	0.56	49	0.69
5	0.88	20	0.61	35	0.57	50	0.59
6	0.73	21	0.51	36	0.64	51	0.50
7	0.57	22	0.93	37	0.37	52	0.42
8	0.61	23	0.50	38	0.82	53	0.35
9	0.93	24	0.68	39	0.93	54	0.79
10	0.86	25	0.71	40	0.73	55	0.59
11	0.68	26	0.65	41	0.59	56	0.78
12	0.75	27	0.87	42	0.69	57	0.46
13	0.78	28	0.38	43	0.59	58	0.37
14	0.83	29	0.38	44	0.80	59	0.42
15	0.71	30	0.49	45	0.84	60	0.33

1986 Pilot Test - Form 6

Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
1	0.78	16	0.75	31	0.58	46	0.84
2	0.59	17	0.84	32	0.46	47	0.63
3	0.75	18	0.88	33	0.26	48	0.33
4	0.90	19	0.98	34	0.58	49	0.78
5	0.81	20	0.96	35	0.59	50	0.76
6	0.82	21	0.97	36	0.45	51	0.36
7	0.81	22	0.64	37	0.52	52	0.41
8	0.84	23	0.61	38	0.63	53	0.83
9	0.90	24	0.39	39	0.38	54	0 84
10	0.72	25	0.79	40	0.61	55	0.62
11	0.70	26	0.72	41	0.67	56	0.59
12	0.93	27	0.64	42	0.56	57	0.29
13	0.92	28	0.35	43	0.82	58	0.33
14	0.20	29	0.50	44	0.49	59	0.49
15	0.92	30	0.31	45	0.53	60	0.23

Table 4 (continued)

Item Difficulty by Item Number for 1986 and 1987 Algebra I Tests

1986 Pilot Test - Form 7

Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
1	0.68	16	0.54	31	0.59	46	0.77
2	0.48	17	0.61	32	0.37	47	0.72
3	0.68	18	0.87	33	0.43	48	0.40
4	0.85	19	0.79	34	0.35	49	0.64
5	0.83	20	0.56	35	0.40	50	0.71
6	0.69	21	0.98	36	0.47	51	0.32
7	0.85	22	0.72	37	0.53	52	0.41
8	0.93	23	0.92	38	0.80	53	0.50
9	0.87	24	0.88	39	0.49	54	0.40
10	0.73	25	0.80	40	0.51	55	0.66
11	0.90	26	0.86	41	0.45	56	0.45
12	0.66	27	0.79	42	0.60	57	0.31
13	0.90	28	0.57	43	0.75	58	0.50
14	0.53	29	0.63	44	0.47	59	0.32
15	0.92	30	0.17	45	0.75	60	0.37

1986 Pilot Test - Form 9

Item	Difficulty	Item	Difficulty	Item	Difficulty	Item	Difficulty
1	0.92	16	0.37	31	0.32	46	0.35
2	0.80	17	0.97	32	0.31	47	0.55
3	0.62	18	0.77	33	0.39	48	0.82
4	0.91	19	0.69	34	0.66	49	0.77
5	0.88	20	0.60	35	0.56	50	0.55
6	0.71	21	0.68	36	0.53	51	0.40
7	0.82	22	0.70	37	0.45	52	0.40
8	0.93	23	0.70	38	0.75	53	0.41
9	0.89	24	0.87	39	0.35	54	0.39
10	0.61	25	0.60	40	0.65	55	0.75
11	0.97	26	0.59	41	0.54	56	0.36
12	0.78	27	0.48	42	0.52	57	0.43
13	0.91	28	0.49	43	0.53	58	0.34
14	0.67	29	0.42	44	0.54	59	0.44
15	0.62	30	0.52	45	0.82	60	0.35

TEST NORMS

Students who answer all 60 core items on the NC Test of Algebra I correctly could be assumed to be excellent Algebra I students. If everyone answered all items correctly, however, a different interpretation would have to be placed on the scores. At some point, scores must have a reference grounded in the experience of all students. In some respects, at least, everything is good or bad by comparison. Norms tables provide that reference. Given a norms table, a student's score can be compared with other students' scores.

Norms tables commonly have two points of reference: a scale of percentiles and a scale of standard scores. The former permits the location of a score within percentile ranks; thus a student is said to have exceeded the performance of 80% of the students in the norm group (in this case, all Algebra I students taking the NC Test of Algebra I in North Carolina in May 1986). The latter, standard scores, permits the location of a score within normally-distributed standard scores. This reference is appropriate if the student abilities are believed to be normally distributed. In a normal distribution, raw scores are given greater and greater weight as they diverge from the mean in either direction.

The choice of a metric for the standard scores is arbitrary. To avoid inappropriate and confusing comparisons with some of the more common metrics, such as those employed in IQ scores or NCE scores, a metric having a mean of 50 and a standard deviation of 10 was chosen. Most curriculum research studies involving the summation of scores will find the standard score to be the statistic of choice.

The norms table for student scores on the NC Test of Algebra I is given in Table 5. These scores set a baseline of comparison for present and future achievement in Algebra I. Thus a student score in 1986, 1987, and future years can be referenced to the scores of all 1986 Algebra I students in North Carolina.

Table 5

Norms Table for Student Scores on the NC Test of Algebra I

Number Correct	State Percentile	Standard Score[a]
60	99	74.0
59	99	72.9
58	99	71.8
57	98	70.8
56	98	69.7
55	97	68.6
54	96	67.5
53	95	66.5
52	93	65.4
51	91	64.3
50	90	63.2
49	87	62.2
48	85	61.1
47	82	60.0
46	79	58.9
45	76	57.8
44	73	56.8
43	69	55.7
42	66	54.6
41	62	53.5
40	58	52.5
39	54	51.4
38	50	50.3
37	46	49.2
36	42	48.2
35	38	47.1
34	34	46.0
33	31	44.9
32	27	43.9
31	24	42.8
30	21	41.7
29	18	40.6
28	15	39.6
27	13	38.5
26	10	37.4
25	8	36.3
24	7	35.3
23	5	34.2
22	4	33.1
21	3	32.0
20	2	31.0
19	2	29.9
Less Than 18	1	28.8

[a] Adjusted to a mean of 50.0 and a standard deviation of 10.0

APPENDIX

Goals and Objectives Rejected for use in the North Carolina Test of Algebra I

Goals and Objectives Rejected for Use in the NC Test of Algebra I

Goal	Objective - Description	Number of Items	% Teachers Rating as Basic[a]
Goal 4:	Locate numbers on the number line or rectangular coordinate plane.		
4.5	Use the vertical line test to determine if a relation is a function.	12	54
4.8	Graph a linear inequality in two variables.	12	48
4.9	Graph the solution sets of linear inequalities in two variables.	12	41
4.10	Graph a quadratic equation.	12	25
4.11	Use the discriminant to determine the numbers of roots of an equation of the form $Y = AX^2 + BX + C = 0$.	12	14
Goal 6:	Solve linear equations.		
6.13	Solve a simple linear equation involving absolute value.	12	66
Goal 7:	Solve linear inequalities.		
7.3	Use inequalities to solve verbal problems.	12	41
7.4	Find the solution set of combined linear inequalities.	12	52
Goal 8:	Understand and solve systems of linear equations.		
8.9	Use systems of pairs of linear equations to solve certain puzzle problems (digit, age, fraction, uniform-motion, coin, mixture).	12	27
Goal 10:	Solve quadratic equations.		
10.5	Solve a quadratic equation by completing the square.	12	31
10.6	Use the Quadratic Formula to solve quadratic equations.	12	40
10.7	Use quadratic equations to solve problems.	12	29
Goal 12:	Simplify expressions which contain radicals.		
12.3	Multiply two binomials which contain square roots.	12	42
12.4	Solve simple equations which contain radicals.	12	43
Goal 13:	Use logarithms to solve problems.		
13.1	Classify angles as acute, right, or obtuse.	12	28
13.2	Identify vertical, adjacent, complementary and supplementary angles.	12	24
13.2	Solve problems related to the measurement of the angles of triangles.	12	31

Goals and Objectives Rejected for Use in the NC Test of Algebra I

Goal	Objective - Description	Number of Items	% Teachers Rating as Basic[a]
13.4	Solve problems about perimeters and areas.	12	67
13.5	Use the similar triangle relationship to solve problems.	12	15
13.6	Use the Pythagorean Theorem and its converse to solve geometric problems.	12	29
13.7	Find the sine, cosine, and tangent of the acute angles in a right triangle.	12	7
13.8	Find the values for trigonometric functions for given angles.	12	7
13.9	Find the measures of angles for given values of trigonometric functions.	12	6
13.10	Use trigonometric ratios to solve problems.	12	7