DOCUMENT RESUME

ED 228 273 TM 830 122

AUTHOR	Herman, Joan L.
TITLE	Criteria for Reviewing District Competency Tests.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
REPORT NO	CSE-RP-4
PUB DATE	82
NOTE	30p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; Cutting Scores; *Evaluation Criteria; Formative Evaluation; *Minimum Competency Testing; Models; Performance Factors; *School Districts; Test Bias; *Testing Programs; Test Reliability; *Test Reviews; Test Validity
IDENTIFIERS	Standard Setting

ABSTRACT

A formative evaluation minimum competency test model is examined. The model systematically uses assessment information to support and facilitate program improvement. In terms of the model, four inter-related qualities are essential for a sound testing program. The content validity perspective looks at how well the district has defined competency goals and the extent to which the tests reflect those definitions--the match between district objectives and test items. The technical quality perspective examines the adequacy of the test as a sound measurement instrument--the goodness of the test itself. Standard setting procedures look at how the district has defined acceptable performance--the standard for determining remedial needs. Finally, curricular validity looks at how well the district's instructional program reflects its objectives and assessment efforts--the match between tests and instruction. (PN)

CRITERIA FOR REVIEWING DISTRICT COMPETENCY TESTS

Joan L. Herman

CSE Resource Paper No. 4

1982

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

INTRODUCTION

This paper reviews four inter-related qualities that are essential for a sound competency testing program:

1. Content validity: Do the tests measure meaningful and significant competencies? Do they clearly describe students' status with respect to those competencies?

2. Technical quality: Are the test items technically sound, reliable, sensitive to instruction, and free from bias?

3. Standard setting procedures: Were reasonable procedures used to establish minimum performance criteria? Are the cut-off scores defensible?

4. Curricular validity: What is the relationship between the competency tests, district curricula, and classroom instruction? To what extent are the test competencies reflected in the instructional program?

These perspectives derive from commonly advanced principles of criterion-referenced test construction, in general, and of competency testing in particular. (See, for example, Berk, 1980; Hambleton, 1979; CSE, 1979.) They also reflect a particular view of what purposes competency tests ought to serve and the nature of an optimal assessment system. We make these views explicit before moving to the test reviews.

We assume that competency tests ought to assess students' proficiency with regard to clearly specified district goals and that the results of such tests ought to be used to improve the quality of instruction for students and to facilitate student achievement. Test results can identify individual student needs, on the one hand and, in the aggregate across students, can be used to identify areas where school or district programming requires strengthening. Instructional efforts can then be targeted to areas of need, through tailored remedial efforts and through more future oriented curriculum analysis and improvement.

The idea is not that tests ought to drive district curriculum and instruction, nor that teachers, strictly speaking, ought to "teach to the test." Rather, both testing and instruction ought to reflect significant, agreed upon district competency goals. Tests should measure important objectives, and classroom or other school instruction should provide students an opportunity to attain those objectives, a view upheld by recent court discussion in minimum competency test litigation. (See Figure 1.)

Figure 1

Minimum Competency Test Model

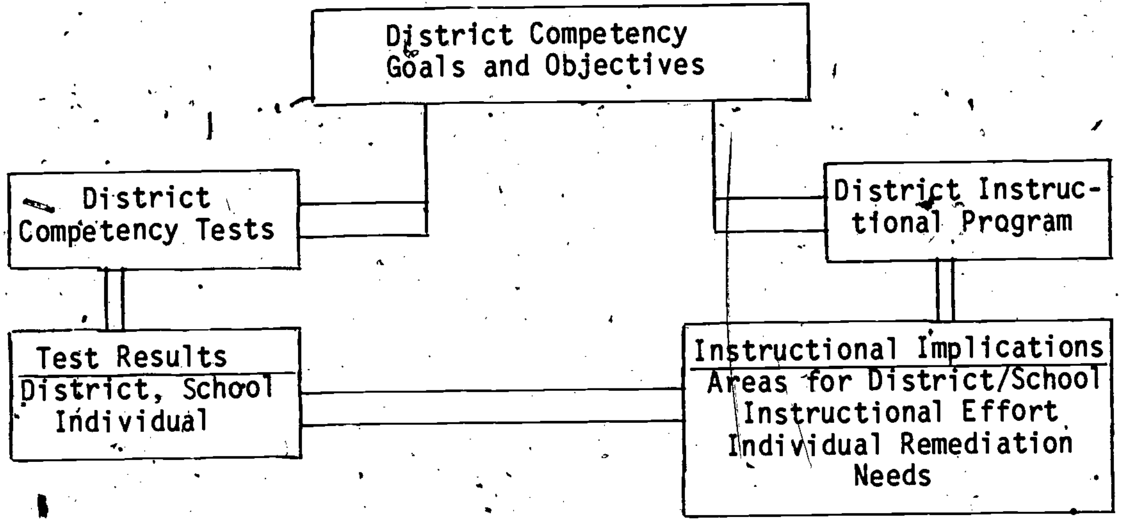
District Competency Goals and Objectives
District Competency Tests
District Instructional Program
Test Results
District, School Individual
Instructional Implications
Areas for District/School Instructional Effort
Individual Remediation Needs

Figure 1 displays a formative evaluation model that systematically uses assessment information to support and facilitate program improvement. The test reviews address the adequacy of the district's efforts for implementing such a model as well as the integrity of the tests for assessing competency. In terms of the model, the content validity perspective looks at how well the district has defined competency goals and the extent to which the tests reflect those definitions--the match between district objectives and test items. The technical quality perspective examines the adequacy of the test as a sound measurement instrument--the goodness of the test itself. Standard setting procedures look at how the district has defined acceptable performance--the standard for determining remedial needs. Finally, curricular validity looks at how well the district's instructional program reflects its objectives and assessment efforts--the match between tests and instruction. These perspectives, of course, are quite inter-dependent; for example, content validity, or any other kind of validity, is impossible without adequate technical quality.

Each of the review criteria are more fully defined in the sections which folow. We make explicit those areas where there is little agreement among experts and where there are problems in available methodologies; here, we offer our views of the best available solutions.

CONTENT VALIDITY

Scores on competency tests are not ends in themselves; they are of interest because they indicate students' proficiency with regard to particular skills and/or content areas--those deemed necessary for competence. Content validity asks whether a test score is meaningful in this sense: whether the test measures what it claims to measure. To the extent that student scores are representative of competence, the tests are content valid. Such validity is not established statistically, but rather is demonstrated by means of well documented, rational, systematic, and logical judgments. The answers to several questions are germane here:

1. To what extent do the assessed skills constitute competence? For example, do the objectives measured on the reading test fairly represent those needed for an individual to be competent?

2. Are the assessed skills adequately described? We can make judgments on how well a test measures particular skills only insofar as we are clear on what the test is intended to measure. A clear description enables item writers to write good items, teachers to teach the skills that are described, and tells consumers of test results, including parents and students, just what was tested; the test description thus gives meaning to the test score.

3. To what extent do the test items match their descriptions? The test descriptions above represent intentions in constructing tests. Here we ask "Were the intentions carried out?" If the test items are adequately described by the test description, then we have a good idea of what is tested; if not, then the items measure something else, and the test is invalid.

4. Does the item sampling scheme fairly represent the domain of interest? Does the number of items included for each skill reflect its importance relative to the total set of tested skills? Number 1 above asks whether the skills tested are a reasonable definition in a particular area, e.g., reading. Here we are interested in whether the total pool of items represents skills in proportion to their importance for competence.

We take each of these content validity issues in turn, describe procedures for optimizing content validity, and suggest actions the district may want to take.

Do the Skills Tested Constitute Competence?

There certainly is no sure and fast definition of competence. The issue is fraught with philosophical and value judgments, methodological problems barring empirical validation and, not surprisingly, subject to wide interpretation. Despite these problems, however, a commonly accepted approach to defining competencies has evolved. This approach features the consensus of community and school personnel and subject area specialists. Typically, a committee is charged with the responsibility for generating a list of skills to be assessed based on input from teachers, administrators, parents, students, other interested citizenry, and extant curricula and textbooks. The most important skill areas are identified, again based on the input of all interested parties. Lastly, the final selection of test content is validated by surveying the opinions of teachers, parents, administrators, students, subject area scholars, and community members about the comprehensiveness, representativeness, and relevance of the selected test content.

Are the Assessed Skills Adequately Described?

Clear descriptions of tested skills are essential for competency testing and accountability. Optimally, content, format, response mode, and sampling for each skill are described thoroughly enough so that

on the testing side:

a. different test writers should produce equivalent test items by following the instructions inherent in the description;

b. it is clear whether any test item, or set of items, falls within or outside the test domain (CSE, 1979);

on the instruction side:

c. teachers can provide instruction and equivalent practice for the skill domain;

d. it is clear whether or not instructional activities directly address the domains of interest;

and on the fairness side:

e. the test expectations are clear to all.

Several approaches to test descriptions have been advanced. Among them are item forms, amplified objectives, domain specifications, and mapping sentences. Domain specifications, (following the work of Popham, 1978, 1980; and Baker, 1974, among others) seem to provide an optimal compromise between practicality and technical sophistication.

Following this approach, domain specifications are created for each objective or skill tested. The domain specification includes:

1. A general description of the skill to be assessed, e.g., the instructional objective.
2. A sample item including directions for administration.
3. Content limits, i.e., a description of the content presented to the students, and the critical features of the task to which students respond, including, e.g., eligible information, concepts and/or principles, structural constraints (length, organization), and language constraints (semantic and linguistic complexity).
4. Response limits, i.e., the nature of the required response. For selected response items rules for generating correct and incorrect alternatives are given; for constructed response items, criteria are included for judging the adequacy of students' responses.

Several sample domain specifications are provided in the appendix as illustrations.

Do the Test Items Match Their Descriptions?

As noted above, the test descriptions portray a test maker's intentions. It is still necessary to verify that the intentions were carried out, and that the test items really measure the skills they purport to measure. While evidence that the test items were generated from the detailed test specifications is one step toward such verification, independent confirmation from qualified judges is also highly desirable. This confirmation process might also examine whether there are extraneous factors in the test items that detract from measurement validity--e.g., linguistic complexity, cultural bias, vocabulary level, unclear directions, confusing format, etc.--and that may confound a student's ability to demonstrate a particular skill.

Does the Item Sampling Scheme Proportionately Represent the Domain of Interest?

This last aspect of content validity is a simple one, but it is frequently overlooked. When judgments about a student's competency are made on the basis of a single test score, then decisions about the number of items to include per skill objective should be based on the relative importance of each objective; or alternatively student scores should be weighted accordingly. For example, if you were to construct a twenty-item test measuring four objectives of equal importance, you would want to include five items for each objective. Alternatively, if two of the objectives were judged twice as important as the remaining ones, you might want to allocate your twenty items differently, e.g., seven for each of the more important objectives and three each for the remainder. Parenthetically, it should also be noted that the reliability of an individual's score on a particular skill is a direct function of the number of items in the skill objective. While no absolute minimum number of items can be specified without knowing the difficulty and variation of test item performance, it has been recommended that reasonably accurate estimates of individual abilities are obtained with a dozen or so items per skill objective.

TECHNICAL QUALITY

Authorities in the field agree that content validity and descriptive rigor are essential for a good minimum competency test. However, there is less agreement on the need for empirical (data based) validation of tests. CSE maintains the position that both types of validity are inter-dependent and that both are necessary to assure test integrity. Without empirical validation, a test that appears to be conceptually sound may give measures that are not consistent, that are insensitive to students' competence levels, that are biased, and/or that measure unintended skills and/or abilities. Indices of technical quality help prevent such occurrences by signalling potential problems in a number of areas:

1. Does the test provide consistent estimates of students' performance?
2. Is the test sensitive to school learning? Do the test items differentiate between students who are competent and those who are not?
3. Does the test measure a coherent skill?
4. Are the tests free from bias, or do they seem to discriminate against particular subgroups?

Test statistics alone cannot either discredit or guarantee test validity, but they are useful for identifying items or subscales that need further scrutiny.

Does the Test Provide Consistent Estimates of Student Performance?

A test is consistent if the difference in a student's score on two occasions is due to a real change in achievement. If a student's score changes as a result of poor directions, variations in testing conditions, or other irrelevant factors, then the test's scores are not consistent--and the test scores do not reflect real learning or achievement.

Test-retest reliability and alternate forms reliability are two indices of test consistency that are particularly important for minimum competency tests. Test-retest reliability indicates whether, in the absence of new learning, test scores are consistent over time. That is, if a test is given on two occasions, and no relevant instruction or learning occurs in the interim, then students' skill levels and their test scores should be the same. If, under these conditions, test scores vary substantially, then the test does not provide a good estimate of student skill proficiency.

Alternate forms reliability indicates the extent to which two or more forms of a test give parallel information. If both provide similar estimates of students' performance, this constitutes evidence that both tests are consistent and measure the same skill--the skill they are supposed to measure.

Test-retest reliability and alternate forms reliability are indices that developed out of classical, norm-referenced test methodologies. In the context of minimum competency testing and pass-fail decisions, consistency needs to be demonstrated for pass-fail judgments: i.e., are pass or fail judgments consistent from one test occasion to the next, and/or do two supposedly equivalent test forms yield consistent pass or fail decisions? Although several methods have been advanced for calculating these two reliability indices, proportion of agreement seems the simplest and most straightforward computation.

A district may wish to administer its tests on two occasions, e.g., in two or so weeks interval, and then compute the proportion of agreement in pass-fail judgments, i.e., the proportion of students who were similarly classified on both testing occasions (either pass-pass or fail-fail). Similarly, it would be useful to administer the "pre" (fall) and "post" (spring) versions of the test simultaneously to a sample of students and determine the extent to which the two forms yield consistent pass or fail decisions. The district may also wish to examine the consistency of pass-fail judgments for the specific competencies measured by the test, particularly if the tests are intended to function as diagnostic tools.

While the reported measurement quality indices presumably give some indication of alternate forms reliability a district also may wish to investigate test-retest reliability for the high school proficiency tests.

Is the Test Sensitive to Learning and Competency?

Students' scores on a test may or may not reflect actual student learning and may or may not accurately portray their competency with respect to skills the test is intended to measure. A test which does provide such an accurate portrayal is described as sensitive to the phenomenon of interest--sensitive to school-learned basic competency--and scores from such tests provide a reasonable basis for competency or non-competency judgments. Naturally, evidence of such sensitivity is important for establishing test validity.

Two aspects of sensitivity are implied above. First, are test scores sensitive to what students learn in school and do they reflect the positive effects of instruction? For example, do students who have been instructed in the assessed skills outperform those who have not been so instructed? While quality of instruction may affect the answer, it is important to demonstrate that test content is teachable, and that test scores indeed reflect school learning. Otherwise, the utility of test scores for school or individual accountability is negligible.

A second aspect of sensitivity focuses on test accuracy in differentiating between masters and non-masters, or between those who are "competent" and those who need additional remediation. For example, do those who are independently judged competent pass the test items and those who are not so judged fail the items?

Similar item analyses can be conducted to investigate both aspects, although there is not yet consensus on how to prove a test's sensitivity. While acknowledging that all available techniques suffer from some technical problems, several easy to compute alternatives are suggested below. These methods identify items that appear to be insensitive and that therefore need additional review. This additional review may uncover problems, e.g., ambiguous wording, poor distractors, unfamiliar vocabulary, poor construction. If such defects are discovered, items should be revised or discarded. Alternatively, closer inspection may not reveal any defect, in which case no revision is necessary. In other words, an item should not be rejected solely on the basis of an item statistic, but only when both the empirical analyses and substantive review indicate a problem.

To determine whether the tests are sensitive to school learning, the district may wish to administer the same test at several grade levels, and examine the extent to which students' scores improve with instructional exposure. For example, one would expect students' achievement to increase over time, especially from prior to post instructional exposure. Pre-to-post instruction growth is evidence that a test item is sensitive to school learning. Most simply, the question is, do instructed students in high grades outperform those in lower grades or do students' test scores increase from the beginning to the end of the school year?

A parallel question addresses the issue of sensitivity to minimum competency: do students who are clearly competent outperform students who are clearly not competent? There are problems in independently identifying students who fall into each category, and many schemes have been used, e.g., teacher judgments, school grades, other test scores. However, demonstrating that test items and passing scores differentiate between the competent and the non-competent is an important validity issue.

Sensitivity indices indicate the extent to which test items differentiate between criterion groups, between instructed and uninstructed students, (either at different grade levels, or students at the beginning of the school year versus the same students at the end of the year) and/or between masters (competents) and non-masters (not competents).

Several easy to calculate statistics are based on item difficulty--the proportion of students who answer an item correctly:

> Item difficulties are computed separately for the two criterion groups, and then compared to see whether there are differences in the expected direction--e.g., the proportion of students who answered an item correctly on the postest minus the proportion who answered it correctly on the pretest. One would expect considerably higher item difficulty values for the instructed or for the "competent" groups.

Other, more sophisticated indices also have been developed, but these cannot be calculated efficiently without a computer program.

Do the Tests Measure Coherent Skills?

Items on the competency tests are developed to assess specific competencies, and ideally there should be evidence that each of these competencies represents a coherent skill. Some believe that such coherence is demonstrated by measures of the extent to which all items for a given skill function alike. For example, the more students' scores on one test item are similar to their scores on the other items measuring the same skill, the greater the coherence of the measure. Such information can supplement and help confirm content validity judgments.

In practice, however, item coherence (or homogeneity) is often unrealistic, because a variety of skills may define the content of an instructional objective or competency. For example, a phonics competency might deal with a variety of categories of consonants (e.g., stops, liquids, nasals). While student performance might be uniform within each category, it would not necessarily be consistent across categories.

A district may wish to examine item homogeneity and consistency within subscales to signal possible aberrant items and/or to help verify item-test description match. Factor analyses of competencies including at least ten items and appraisal of item difficulties within subscales would also be useful so long as a full range of abilities exists in the data being analyzed.

Are the Tests Free from Bias?

A test is biased for a given group (e.g. a particular ethnic or language group) of students if it does not permit them to demonstrate their skills as completely as it permits other groups to do so; and/or taps different skills and/or abilities in different groups. For obvious reasons, bias has been a controversial issue in achievement testing. It is particularly significant in minimum competency testing because of the potential consequences of such tests.

Bias can be apparent in a test in a number of ways, including obvious presentation defects (e.g., items that disparage some groups, that depict solely majority customs or activities that are stereotypic, etc.), linguistic and semantic problems, and socio-cultural and contextual bias. A careful item review can minimize the more obvious problems, but such analyses should be supplemented with statistical procedures for detecting bias.

These statistical analyses are derived conceptually from the nature of an unbiased test: one that measures the same skill or ability, and is equally reliable and sensitive for all groups. Evidence that the patterns of performance are similar for all groups is one way to document that a test is not biased. Demonstrating that technical quality indices are similar for all groups--e.g., consistency, coherence, sensitivity--is additional evidence that a test is relatively free of bias.

STANDARD SETTING PROCEDURES

There is no simple answer to the problem of setting reasonable standards for competency tests. A variety of methods for setting passing scores have been advanced, but all have been criticized as at least somewhat arbitrary, because all require human judgment. But imperfection does not obviate the need for decisions, and more reasoned judgments tend to produce more reasoned and defensible decisions.

Most recent approaches to setting passing scores acknowledge the need for multiple sources of information, and combine judgmental and empirical data. Many advocate input in the judgment process from a broad cross-section of constituents. Several methods are described below to illustrate the range of available alternatives.

Several principally judgmental methods require judges to examine each item on a test and decide whether or not a minimally competent student should be able to answer the item correctly--or some variation of such an individual item rating. Passing scores are then computed by averaging over judges the total number of correct responses that a minimally competent student should be able to provide. Most recent variants of this approach require the use of pilot-test data to help assure that judges' ratings are realistic. For example, judges are provided with item analyses from a district pilot test to help them ascertain the difficulty of the item, and whether or not a minimally competent person should be able to correctly answer the item. An iterative process of rater judgments, resultant passing standards, and the normative implications of those passing standards (e.g., the percentage of high school students likely to fail) is then used to arrive at a final decision. (See, for example, Jaeger, 1978.)

Another approach to setting standards asks raters to make judgments about mastery levels of students rather than about test items. Judges (most likely teachers) identify students as "competent," "incompetent," or "borderline" with respect to the subject domain being tested. In the "borderline group" method, the students so identified are administered the test, and the median test score for this group becomes the standard. Alternatively, in the contrasting groups methods, the test is administered to students who are identified as clearly competent and to those who are identified as clearly incompetent. Score distributions are plotted for the two groups, and the point at which the two distributions intersect becomes the first estimate of the standard. This estimate can then be adjusted up or down to minimize different types of decision errors, i.e., misclassifying a competent student as incompetent and vice versa.

This latter consideration is an important one, regardless of the method used. Students' test scores, at best, provide only estimates of their competence. Indices such as the standard error of measurement provide some indication of the quality of the estimate, and/or the potential error incorporated into the test scores. Passing scores should not be set without some consideration of measurement errors and likely classification errors.

CURRICULAR VALIDITY

When a local district sets competency standards, it is defining the components of an adequate education, and is enjoining the responsibility for providing such an education. That competency tests assess skills and objectives that are actually taught in school is essential to the logic and legality of any such program. If students are not provided with the opportunity to learn the test content, and if test content does not match what students are taught in classrooms, then the system is senseless and unfair, a view affirmed in recent U.S. court rulings in the Florida minimum competency litigation.

Curricular validity focuses attention on this very important requirement of minimum competency testing programs: does the test measure skills and objectives that are fully covered in the district curriculum? Does classroom instruction afford students relevant practice in the assessed skills? While these questions appear simple and straightforward, a methodology for providing answers is only now emerging, and a number of issues are yet to be resolved. For example, how do you document classroom instruction to demonstrate that students are actually exposed to the minimum competency objectives? How similar must instructional activities and test content be to count as a reasonable match? How much instruction and practice in the assessed skills is sufficient to fulfill district responsibilities?

Formal attempts to deal directly with the problem of match have developed along two different lines: detailed curriculum analyses and teacher-based estimations. Approaches to curriculum analyses have generally involved comparisons between curriculum scope and sequence charts and test descriptions of content covered. Typically these analyses have not included information on how much of the scope and sequence was actually covered. They have also assumed that similar content or topic labels mean the same thing, e.g., inferential comprehension means the same thing to both the test developer and the curriculum developer.

More recent work has started with a detailed taxonomy of objectives in a subject domain. Curricular and test coverage are then mapped on this taxonomy and the extent of overlap is ascertained. Following this approach one would start with domain specifications, as described earlier in this paper, and then examine test items and curriculum materials to verify that the specified skills were indeed assessed by the test and included in the curriculum. Such an approach yields more precise estimates of curriculum coverage but is limited in that it considers only the formal curriculum, not teacher presentations, nor teacher generated instructional activities nor differing rates of actual curriculum use. Having teachers indicate whether students in their class have been exposed to the minimum material necessary to pass each item has been one response to the problem, but the credibility of estimation in the context of minimum competency testing is probably suspect.

Providing supplementary instruction and appropriate practice materials for each objective covered on the competency tests also insures instructional opportunity for all, and is clearly a necessity for remediation. Ideally these practice materials would be developed from the same specifications that guided test development, and/or could be selected from relevant portions of available curriculum materials.

Clear articulation of competence across grade levels and a logical progression of skill development further supports students' opportunity to learn the assessed competencies. For example, do the reading competencies at grade five and grade seven include the necessary pre-requisites for the required high school reading proficiencies? The judgment of subject area experts might provide evidence of a reasonable sequence of skills, and thus reasonable notice and opportunity to learn.

A district ought to consider an analysis of the formal curriculum --e.g., basic texts--to determine whether and where each assessed competency is covered. Supplementary exercises could be developed or located in other available materials to compensate for any gaps--and to support remedial needs.

REFERENCES

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. In W. Hively (Ed.), Domain referenced testing. Englewood Cliffs, New Jersey: Educational Technology Publications, 1974.

Berk, R. A. (Ed.). Criterion-referenced measurement: The state of the art. Baltimore, Maryland: The Johns Hopkins University Press, 1980.

Center for the Study of Evaluation. CSE criterion-referenced test handbook. Los Angeles, California: University of California, Center for the Study of Evaluation, 1979.

Hambleton, R. K. Competency test development, validation and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), Minimum competency achievement testing. Berkeley, California: McCutchan, 1979.

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill, N.C., 1978.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, New Jersey: Prentice Hall, 1978.

Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, Maryland: The Johns Hopkins University Press, 1980.

APPENDIX

Sample Domain Specifications

Grade Level: Grade 3

Subject: Reading Comprehension

Domain Description: Students will select from among written alternatives the stated main idea of a given short paragraph.

Content

1. For each item, student will be presented with a 4-5 sentence expository paragraph. Each paragraph will have a stated main idea and 3-4 supporting statements.

2. The main idea will be stated in either the first or the last sentence of the paragraph. The main idea will associate the subject of the paragraph (person, object, action) with a general statement of action, or general descriptive statement. E.g., "Smoking is dangerous to your health," "Kenny worked hard to become a doctor," "There are many kinds of seals."

3. Supporting statements will give details, examples, or evidence supporting the main idea.

4. Paragraphs will be written at no higher than a third grade reading level.

Distractor Domain:

1. Students will select an answer from among four written alternatives. Each alternative will be a complete sentence.

2. The correct answer will consist of a paraphrase of the stated main idea. Paraphrased sentences may be accomplished by employing synonyms and/or by changing the word order.

3. Distractors will be constructed from the following:

   a. One distractor will be a paraphrase of one supporting statement given in the paragraph (e.g., alternative "a" in the sample item).
   b. One - two distractors will be generalizations that can be drawn from two of the supporting statements, but do not include the entire main idea (e.g., alternative "d" in the sample item).
   c. One distractor may be a statement about the subject of the paragraph that is maore general than the main idea (e.g., alternative "b" in the sample item).

Format: Each question will be multiple choice with four possible responses.

Directions: Read each paragraph. Circle the letter that tells the main idea.

Sample Item: Indians had many kinds of homes. Plains Indians lived in tepees which were made from skins. The Hopi Indians used bushes to make round houses, called hogans. The Mohawks made longhouses out of wood. Some Northeast Indians built smaller wooden cabins.

What is the main idea of this story?

a. Some Indians used skins to make houses.

b. There wefe different Indian tribes.

*c. Indians built different types of houses.

d. Indian houses were made of wood.

Grade Level: Grade 8

Subject: Introduction to Algebra

Domain Description: Using basic operations and laws governing open sentences, solve equations with one unknown quantity.

Content Limits:

1. Stimuli include a number sentence with one unknown quantity, represented by a lower case letter in italics, and array of four solution sets or single answers, only one of which is correct.

2. Number sentences may be statements of equalties or inequalities.

3. The number sentences may require simplifying before solving by combining like terms or carrying out operations indicated (e.g., by parentheses).

4. Number sentences will have no more than five terms. Fractions may be used but not decimal fractions and non-decimal fractions in the same expression. Exponents (powers) may appear in the expression only if they cancel out and need not be solved or modified.

5. Solution sets for equations and inequalities will be drawn from the set of rational numbers (±). The null set (/) may be used as a correct solution set.

6. Factoring may be a requisite operation for solving the equation.

7. Application of the distributive property of multiplication and the use of reciprocal values may be requisite operations for solving the equation.

Distractor Domain:

1. Distractors may be drawn from the set of wrong answers resulting from errors involving any one of the following operations:

   a. combining terms

   b. transformations that produce equivalent equations (e.g.,transferring terms using the principle of reciprocal values)

   c. distributing multiplication, with positive or negative numbers (e.g., across parentheses)

   d. carrying out basic operations using brackets or parentheses

2. Distractors may also be drawn from the set of wrong answers due to incomplete solution sets.

3. Distractors may not reflect errors due to wild guessing, calculations involving negative numbers, errors in basic operations.

4. "None of the above" is <u>not</u> an accepatable alternative.

<u>Format:</u> Multiple choice; five alternatives.

<u>Directions:</u> Solve the equation. Then select the correct answer or solution set from the choices given.

<u>Sample Item</u> (see directions)

1. $8n + 2 = 2n + 38$; $n = \underline{?}$

   a) $n = 3$
   *b) $n = 6$
   c) $n = 4$
   d) $n = 5$
   e) $n = 7.6$

2. $16x \leq 32$; $x = \underline{?}$

   a) $x = 48$
   *b) $x = [0,1,2]$
   c) $x = 2$
   d) $x = /$
   e) $x = [3,4,5,...]$

Grade Level: Grade 9

Subject: English Punctuation.

Domain Description: Correctly punctuating given paragraphs adapted from a standard eighth grade text of a practical/informative nature.

Content Limits: The student will be presented with one paragraph in which all the correct punctuation marks have been omitted, except for apostrophes in contractions (I'll), and possessives (Jane's), dashes, and, semi-colons.

For each question, students will be asked to choose all the correct punctuation marks which must be added in a given sentence to make it correct. Punctuation marks to be indentified and added may include:

a. periods at the end of a declarative or imperative sentence, after an abbreviation, or an initial
b. question marks following an interrogative sentence
c. exclamation point after exclamatory sentences or interjections
d. colon after the salutation in a business letter, or to separate minutes and hours in expressions of time, and to show that a series of things or events follows
e. quotation marks enclosing a quotation or a fragment of it, enclosing the title of a story or poem which is part of a larger book.
f. comma in a date or address; to set off such words as "yes" at the beginning of a sentence; to set off names of persons or words (phrases) in apposition; to separate words in a series, direct quotations, parallel adjectives, parenthetical phrases; after the salutation and closing in a friendly letter; to separate a dependent clause and independent clause in a complex sentence.

Distractor Domain: The alternate responses to the questions may include:

a. omission of punctuation mark(s) within a given sentence which should be included, or
b. inclusion of a punctuation mark or marks not necessary or correct in the given sentence

Directions: The directions will be given: "Choose the letter which contains all the necessary punctuation marks in the given sentence which will make the sentence correct."

Format: Each question will be multiple choice, with four possible responses.

Sample Item: 1. If she starts to sing I'll crack up 2. It is funny how it hurts to hold back a laugh 3. I was sitting in the auditorium at 10:00 am and we were having a singing rehearsal for graduation 4. Sit up Get off those shoulders Think tall Sing tall Sing like this said Ms Small 5. I knew that if she was going to tweet like a bird again I would laugh 6. But I could not laugh because Ms Small would kick me out of the auditorium and that meant Felson's office--and no graduation 7. La la la--sing children Sing with your hearts said Ms Small 8. I couldn't hold it 9. She was so funny I almost rolled.off the auditorium seat 10. The other students didn't laugh but me I sounded like Santa Claus 11. It became quiet for a second 12. What are you doing Joe I know it is you Present yourself to Mr Felson at once that voice said 13. Ms Small is a foot shorter than a tall Coke but she has the bark of a hungry hound dog

1. The first sentence should be written:

a. If she starts to sing again I'll crack up.
b. If she; starts to sing again, I'll crack up
* c. If she starts to sing again, I'll crack up.
d. If she starts, to sing again, I'll crack up.