DOCUMENT RESUME

ED 269 477 TM 860 336

AUTHOR Siskind, Teri G.; Rose, Janet S.
TITLE CRT Development: An Overview of Why and How--One District's Perspective.
PUB DATE Apr 86
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986). For related document, see TM 860 339.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Behavioral Objectives; *Criterion Referenced Tests; Educational Objectives; *Educational Testing; Elementary Secondary Education; Feedback; Language Tests; Mathematics Tests; Measurement Objectives; Pilot Projects; Pretesting; *School Districts; Task Analysis; *Test Construction; Test Format; Test Items; *Test Theory
IDENTIFIERS *Charleston County School District SC; *Test Specifications

ABSTRACT

The Charleston County School District (CCSD) has recently begun development of criterion-referenced tests (CRT) in different subject areas and for different grade levels. This paper outlines the process that CCSD followed in the development of math and language arts tests for grades one through eight and area exams for required high school courses. The test development process begins with a statement of instruction or an objective. Methods for making objectives testable include revision, creation of new objectives, or grouping objectives into domains. The test purpose and feedback desired should be a guide in making objectives testable. Blueprinting refers to outlining the content to be tested and requires the developer to decide which objectives to test and the number of items needed to test each objective. Test and item specifications provide guidelines for writing items for a given objective. Objectives identification and specifications development provide the foundation for item construction. The pilot test tries out items and administration procedures, obtains empirical data for item evaluation and test form composition, and obtains performance data from students. Decisions to eliminate or retain items for future test forms depend upon the analysis of test data and the preference of the instructional staff. (PN)

CRT DEVELOPMENT: AN OVERVIEW OF WHY AND HOW--
ONE DISTRICT'S PERSPECTIVE

Teri G. Siskind
Charleston County (S.C.) School District

Janet S. Rose
Charleston County (S.C.) School District

AREA Symposium: Problems and Practical Solutions in the Development of Districtwide Curriculum-Referenced Tests

San Francisco
April, 1986

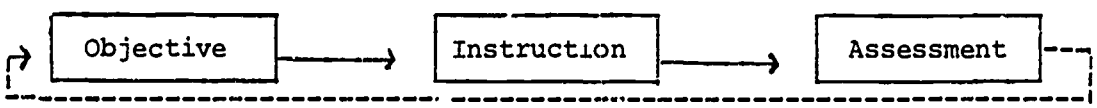
The Charleston County School District (CCSD) has recently begun development of criterion-referenced tests (CRTs) in different subject areas and for different grade levels. CRT development in CCSD has been termed "curriculum-referenced testing," thus highlighting the rationale for test development efforts. Although the district participates in statewide norm-referenced and criterion-referenced programs, neither of these programs totally fulfills curricular/instructional validation needs of the district. In the case of the norm-referenced program, CCSD's situation is hardly unique. NRT's are useful in that they allow comparisons with national norms, but their utility is limited by content derivation which does not completely match CCSD's. Although the district certainly includes state objectives in its curricula, CCSD's curricula go beyond state-required objectives.

THE TEST DEVELOPMENT PROCESS

The remainder of this paper contains an outline of the process that CCSD has followed in the development of math and language arts tests for grades 1-8 and area exams for required high school courses. Decision points with which CCSD staff have been forced to grapple will be highlighted in order to set the stage for their practical resolutions as presented in the accompanying case studies.

IDENTIFICATION OF OBJECTIVES

It seems apparent that in order to assess what has been taught and learned, one must first have a statement of what the student is to obtain from instruction. This is the role of objectives. Measurement texts often illustrate this concept with diagrams like the one below.



It appears, however, the the notion of an objective is one of the least "objective" in education. Objectives can probably best be described as dimensional continua. Objectives sometimes range from narrowly defined behavioral objectives to broadly defined goals. In other situations they are conceptualized hierarchically from "process" objectives to "product" or "terminal" objectives. Sometimes objectives which constitute required portions of a curriculum are designated "core" to differentiate these from optional or enrichment objectives.

Objectives are necessary to a strong instructional program, and there are places in the district's curricula for the many different types of objectives noted above. In our experience, it seemed as if the manner in which objectives were written was directly related to the writer's graduate school education. For example, curriculum staff who were trained to write narrowly defined behavioral objectives produced objectives for which only one or two test items could be written. In contrast, broadly defined objectives could generate a sizable item pool. Of course, both extremes have implications for instruction as well as assessment.

Does the District have stated objectives? CCSD's test development processes have begun with already extant stated objectives. This is due, in large part, to the separation of the Office of Evaluation and Research from the content area instructional offices. If objectives do not exist for a given content area or grade level, then test development does not begin until objectives are written and field-tested by the Instructional Office.

If objectives are stated, are they testable? Since CCSD's CRTs are end-of-year or end-of-semester exams, objectives must be testable within this context before test development begins. For example, in foreign language courses, the same objectives may be repeated throughout the year--with each re-introduction set in the context of a different cultural exposition. This is a clever way

to reinforce learning while maintaining interest, but not entirely necessary from the assessment point of view. In this particular case, a set of "test objectives" were derived from the "situational objectives" which covered the language skills.

Can objectives be made testable? Many objectives can be made testable in numerous ways. Using the foregoing language example, situational-free test objectives can be written. Another example comes from math where the breadth of objectives was a concern. A useful instructional objective might be, "The student will be able to identify the center of a circle." However, whether or not this type of objective is important enough to warrant a large number of items on an end-of-year test may be questionable. Should it be combined with other related objectives to form a geometry domain? The answer to this question is dependent upon the purpose of the test as well as a host of other factors including the number of objectives tested and the intended length of the test. With regard to purpose--will the test be used to certify course mastery, or will it be used to provide rather specific instructional feedback on individual objectives or domains?

In summary, the test development process begins with a statement of instruction. This statement is most often called an objective. Objectives range in type and structure according to instructional need, design, or possibly, writer's graduate school training. Instructional objectives may be testable in an end-of-year or end-of-semester context or they may be made testable. Some of the methods of making objectives testable are revision, creation of new objectives, or grouping them into domains. The guiding force in making objectives testable should be the test purpose and the kind of feedback the test should provide. This may or may not be synonymous with the original design of objectives, but certainly should not be contrary to the intent of the curriculum. Since test information will in some fashion be used as instructional feedback, decisions

about making objectives testable should be made jointly by instructional and test development staff. Hence a healthy rapport between departments is necessary.

BLUEPRINTING

Blueprinting refers to outlining the content to be tested. Simply stated, blueprinting requires that the test developer decide (a) which objectives to test and (b) the number of items needed to test each objective. Thus blueprinting is highly dependent upon decisions made previously: the purpose of the test (e.g., curriculum mastery versus objective/domain mastery); type of objective (e.g., process versus product); the length of testing time; and test format. (The latter two decisions may limit the number of items that can appear on the test.) If curriculum mastery decisions are to be made, the test should be weighted to reflect course content. If information at the objective or domain level is required for diagnostic or evaluative purposes, then the test should contain a sufficient number of items to determine mastery of the objective(s) or domain(s).

For some curriculum areas, the test blueprint can be a content/process matrix which reflects the weight given to various areas and sub-areas. Charts like these are known as Tables of Specifications. A section of the blueprint developed for the district's Spanish I area exam is reproduced in Figure 1.

In the Table of Specifications (Figure 1) the percentage listed in each cell identifies the weight assigned to the Content Domain/Language Skill Area. The Xs represent the location of the Spanish I objectives within the matrix. Marginal percentages summarize the content of the test for each content domain and skill area. Matrix entries should match the amount of instructional time or importance allocated to the topics identified by the cells in order to ensure the validity and fairness of a test designed to assess course mastery.

For end-of-year CRTs, this notion of test/instructional time match may not be straightforward. Suppose, for example, that process and terminal objectives are included for testing. Both are listed in the blueprint. In one content area, instruction demands attention toward the process objectives for half of the year, but the terminal objective is the true culmination of the process objectives and mastery of the terminal objective indicates mastery of this content area. Should assessment be equally divided between process and terminal objectives?

Completion of a Table of Specifications may also highlight a discrepancy between real and ideal. Perhaps curriculum guides are written with 25% of the objectives requiring the "evaluative" skills of Bloom's taxonomy. Do teachers actually spend 25% of instructional time requiring students to evaluate? Is it fair to test students on these skills? Or suppose a survey of the teachers shows that they spend no instructional time eliciting higher level cognitive processing. Should the assessment device and test objectives be used to force the inclusion of higher level skills into instruction?

Because the blueprint is derived from the curriculum and can potentially alter instruction, teachers and curricula/instructional staff should participate in the blueprinting step. Surveys of instructional time and coverage should be conducted, if possible, to validate blueprint assumptions.

TEST AND ITEM SPECIFICATIONS

Specifications provide guidelines for writing items for a given objective. CCSD specifications are based on Popham's amplified objectives. An amplified objective is an enlargement or fuller description of an objective for testing purposes. Amplification is accomplished by providing a generalized description of the objective as it is to be tested, a sample item, and descriptions of the stimulus and response attributes of the item. A specification supplement, listing content eligible for testing, is an optional component.

Detailed specifications are necessitated by CRTs since CRTs are expected to result in fuller descriptions of student behavior than NRTs.

Popham's model of test specification has not been employed in its precise form for any of CCSD's tests. In most cases, an adaptation of the amplified objective model has been employed. The nature of the modification has been dependent upon the type and structure of objectives. CCSD specifications range from checklists (e.g., Number of Options: __3 __4 __5), to prose (e.g., The student will read a question or incomplete statement and select from four alternatives the one which best answers the question or completes the statement.) Component portions of the specifications range from "Descriptions" (Description of the Stimulus) to "Restrictions" (Stimulus Restrictions). For some content areas the specifications are designed as guidelines and for others as strict rules to follow. All of the specifications focus on items, but in some cases test restrictions (e.g., item location on a test) are also given.

Who should compose the test specifications? Specifications should be constructed jointly by measurement staff and instructional staff under the guidance of the measurement staff. Variations on this approach include (but are not limited to) employing external consultants to prepare specifications from input given by district staff or writing specifications which are reviewed by instructional personnel.

Each option has advantages and disadvantages which must be weighed in light of district needs and resources. For example, including teachers in the specifications process helps instill a feeling of ownership which may be critical to the acceptance of the testing program. Teacher inclusion also provides insights into the way in which objectives are actually interpreted for classroom practice. On the other hand, teachers may lack the skills to compose specifications efficiently. When district measurement staff lack the time and/or skills to write specifications for a large-scale project, a contractual

arrangement for specifications development can be initiated in which a district measurement staff member acts as a liaison between the district and the contractor.

Who should review the specifications? Specifications should be reviewed for clarity, completeness and curricular validity. Reviewers indicate whether or not each specification provides complete and unambiguous directions for writing items. Reviewers may also respond to the accuracy and practicality of the specification content, to its congruence with instruction, and to its fairness for students.

Reviewer possibilities include content specialists, measurement specialists and potential item writers. Reviewers may be teachers, former teachers, university professors, district staff members, staff members from other districts, and other professionals in the measurement field. The nature of the test and the way in which the specifications were developed should be taken into account when selecting reviewers. For example, specifications which have been developed by measurement specialists may need to be monitored more carefully in terms of content than form whereas specification written by content area experts may need to be thoroughly reviewed by the measurement community.

What is the best form for a review? Reviews may be conducted orally in group sessions or may be solicited in written form from individuals. A highly structured form may be employed for either groups or individuals, or guidelines may be given and reviewers may be allowed to react informally--by speaking freely or writing comments on the actual specification. When specifications are lengthy and detailed, a logistically sound approach may be administering individual reviews followed by group sessions to allow for discussion and "piggybacking" of ideas.

Who should make revisions? Specifications may be revised by the initial writers, or they may be revised by someone else. The measurement specialists who are ultimately responsible for the test may prefer to make the revisions, or the original specifications committee may wish to study other professionals' reactions and make the necessary adjustments. The educational benefits of the latter option are especially important for committees which may continue to write specifications. Both measurement and content area staff should approve specifications.

When a committee of teachers is employed to write and revise specifications, reviewers comments should be presented anonymously to the committee. Comments may be typed or re-written by a "third" party. With few exceptions, comments should be transposed verbatim.

ITEM WRITING

Objectives identification and specifications development provide the foundation for item construction. When the former processes are conducted properly, item writing becomes a well-defined task. CCSD's item writing assignments have been almost exclusively multiple-choice, but the decision points explicated below would apply to a wider range of item formats.

Who should write items? Like specifications, items may be written by district measurement or instructional staff or may be contracted to outside professionals. Consideration must be given to financial resources, logistics and staff expertise. The advantages of involving teachers include the resulting feelings of ownership and the benefits of item-writing training sessions which may generalize to other teacher endeavors. Potential problems encountered by employing teachers include teachers' lack of item writing expertise or experience and the necessity to accomodate teachers' schedules. This latter problem may be resolved by training groups of teachers on item writing techniques and then giving them independent item writing assignment.

Pointers on item writing for teacher groups. Teachers should be given a report on the test and test development process, a thorough explanation of the test specifications, and training in writing items. To help ensure an equal distribution of style and quality across objectives, items should be assigned in such a way that no one person has full responsibility for writing all items for a given objective.

Initial item writing attempts should be monitored by measurement staff members. Feedback on a sampling of items should be provided during the training sessions and item writers should be responsible for initial revisions. Requiring item writers to address specification issues (e.g., explain why distractors were selected) helps the writers to focus on their tasks.

A combination of group training with some practice en masse, followed by independent assignment, is an approach found to be useful in our district.

Who should review and revise items? The primary purposes of item review are to check for item clarity, content accuracy, bias and face validity. Reviewers should include measurement and content area specialists as well as someone who is familiar with the students who will be responding to items. The people who fill these roles may be district staff members, university professors, or other professionals in the content area or measurement field. Our district's preference is to include teachers, central office staff, and outside experts.

Guidelines for the review form and revisions parallel those given for specifications review and revision with one exception. Revision of items by the original writers quite often is not efficient.

THE PILOT TEST CYCLE

The purpose of the pilot test cycle is to try out items and administration procedures, to obtain empirical data which can be used in item evaluation and

test form composition, and to obtain performance data from students. In CCSD the pilot test cycle has consisted of a pilot administration and a field test administration. The pilot test has been used to try out items and administration procedures. From pilot test data, several forms of the test are compiled. These forms are then field-tested and if necessary, items are re-evaluated and re-calibrated. The field test also provides student performance data which are used for setting student performance standards on the tests. In CCSD, a pilot is conducted the first year, a field-test the next, and a test can become operational the third year.

Advancements in item evaluation methodology and analysis have made the field test phase unnecessary; however, field-testing does provide for a double-check of pilot test data and an acclimation to the test by teachers and students.

How are items selected for the pilot? In piloting, a sufficient number of items are "tried out" to create several operational forms of the test. An overage is included to compensate for "poor" items which may be eliminated by the pilot. Given the testing time, the number of operational test forms desired, and the sample size, the number of pilot forms and items on a pilot can be calculated.

Item writing assignments may have been based on these calculations. In many cases, more than the necessary number of items will have been written since an overage is usually incorporated into the item writing scheme. In selecting items for the pilot forms, poor items are discarded. Then, several factors may need to be considered.

First, even in pilot form composition, attention to face validity is warranted. Recent statistical technology makes it possible to estimate a student's ability on an objective that was not included on the test administered to the student. This concept is not easily understood, and teachers and students viewing the test items are likely to mistrust such test magic. On a pilot, it may be important to give the impression that all objectives are sampled.

Second, in selecting items, one must take care not to include items which provide clues to the answers for other items on the same test form.

Third, the items must all be put on a common scale in order to create equally difficult forms from the items which are being piloted. To do so requires linking pilot test forms with common items. There are several ways to link test forms. Each method must be considered in light of the content area, the number of items piloted, sample size and test time restrictions. Methods used by CCSD alone or in combination include (a) pairwise linking of items between two forms, (a) anchor linking of a group of items across all forms and (c) administration of a separate anchor form which contains items from other pilot test forms. Selection of linking items requires a priori assumptions since the items should be of average difficultly.

How should items be positioned on the pilot? Sometimes it is important to format a test by logical categories. For example, in foreign language, listening tests may be separated from reading tests. History tests may contain chronologically sequenced items. Each test is an individual case, requiring independent consideration for item positioning.

Linking items, which appear on more than one form, may be rotated throughout the test to mediate or ascertain the effects of test position. Linking items should not be placed at the very beginning or end of a test form.

Pilot test administration. "Pilot test administration" encompasses a host of logistical, political and educational considerations. The following discussion is limited to a few of the logistically complicating factors.

In order to randomly assign forms to students, pilot test forms can be "spiralled" prior to distribution. In this system, a Form A booklet is stacked on top of a Form B booklet which is stacked on a Form C booklet, etc. Booklets are distributed to students such that the first student takes Form A, the second takes Form B, etc.

This spiralling method can not be employed when the tests are administered orally (e.g., listening comprehension tests). When whole-class administration is necessitated, a stratified random sampling method is useful. Classes are stratified by average achievement test scores, and forms are randomly assigned within strata.

ANALYSIS OF TEST DATA

Test data may be analyzed in-house or contracted to measurement specialists. Portions of the analysis, ranging from keypunching item numbers and answer keys to initial item calibrations, may be completed by district staff. The more complicated parts may be contracted out. The answer to the question of who should perform the data analyses should take into account district budget, computer hardware and software, staff expertise and staff time.

A worthwhile next step is to study the items in light of statistical analyses and instructional feedback from teachers. For example, aberrant items may be explained by lack of instruction, unusual or varying objective difficulty, or poor item quality. Decisions to eliminate or retain items for future test forms depend upon this information and the preferences of the instructional staff.

CREATION OF TEST FORMS

Sophisticated computer programming is now available to generate euallly difficulty test forms, given restrictions designated by the test blueprint. However, since these programs employ item statistics only, they can occasionally generate test forms lacking in face validity. Sometimes handsorting of items and recalculation of domain difficulty is necessitated.

Spanish I Area Exam Blueprint

CONTENT DOMAIN	LANGUAGE SKILL AREA					DOMAIN TOTALS
	Listening Comprehension	Reading Comprehension	Language Usage: Manipulate Structures	Language Usage: Respond to Conversational Sit.	Recognition of Facts	
VERBS			15%	15%		30%
Regular present & stem changing			X	X		
Regular preterite			X			
Irregular present			X	X		
Reflexive present			X	X		
Contrast ser/estar, conocer/saber			X			
Me/te gusta(n)			X			
NOUNS			10%			10%
Adjective/noun agreement			X			
Article/agreement			X			
Poss. adj./noun agreement			X			
Dem. adj./noun agreement			X			
Personal 'a'			X			
Comp/super adj.			X			
PRONOUNS				5%		5%
Object pronouns				X		
Familiar/polite				X		
Prepositional pro.				X		
VOCABULARY	10%			10%	10%	30%
Adverbs					X	
Prepositions					X	
Interrogatives				X		
Telling time	X					
Calendar/weather	X					
Numbers to 100	X					
Greetings/expres.				X		
Tener idioms					X	
Verb + infin.					X	
Nouns and verbs					X	
General vocabulary					X	
INTEGRATED LANGUAGE COMPONENTS	10% X	10% X				20%
CULTURE					5% X	5%
SKILL TOTALS	20%	10%	25%	35%	15%	100%

Figure 1. Charleston County School District blueprint for Spanish I area exam.