DOCUMENT RESUME

ED 128 386 TM 005 521

AUTHOR	Skager, Rodney
TITLE	Critical Characteristics for Differentiating Among Tests of Educational Achievement.
PUB DATE	[Apr 75]
NOTE	34p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Washington, D.C., March 30-April 3, 1975)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS	*Achievement Tests; Behavioral Objectives; Classification; Conceptual Schemes; *Criterion Referenced Tests; Curriculum Evaluation; Educational Diagnosis; Formative Evaluation; Measurement Goals; *Norm Referenced Tests; Scores; Student Evaluation; Student Placement; Teacher Evaluation; *Test Construction; *Test Interpretation; Test Validity
IDENTIFIERS	Content Process Matrix; Domain Referenced Tests

ABSTRACT

The corpus of descriptive terminology associated with achievement testing has expanded considerably in recent years, in large part due to the heightened interest in absolute and/or direct metrics for interpreting test performance plus the development of more rigorous strategies for specifying test content. Widely prevalent disagreement about terminology reflects a lack of conceptual clarification and may inhibit the development of theory and practice. Distinctions commonly made between criterion referenced and norm referenced tests turn out to be inaccurate, since it appears that both content and norm referenced interpretations can apply to scores on any type of achievement test. Rather, the particular manner in which a given test can and should be interpreted turns out to be a function of the mode by which test content is specified and the function for which the test is to be used. All approaches to the interpretation of achievement test scores are classified as either domain referenced or norm referenced, with reference to a criterion or standard viewed as a special case of the former. Finally, it is argued that normative interpretations can and in many instances should be made of scores which are referenced directly to content, including mastery scores. (Author/BW)

Critical Characteristics for Differentiating Among Tests of Educational Achievement*

Rodney Skager

Center for the Study of Evaluation and
Graduate School of Education, UCLA

The corpus of descriptive terminology associated with the characteristics of achievement tests has expanded greatly in recent years (cf. Alkin, 1974). Much of this expansion derives from the heightened interest in absolute and/or direct metrics for interpreting test performance as well as in the development of more rigorous strategies for defining test content and specifying item characteristics. Disagreement is widely prevalent in the field over the distinctions represented by these new, or sometimes resurrected, terms. Ebel (1971) has even argued that criterion-referenced measurement was tried out and abandoned early in the history of testing.

This paper will argue that recent trends in testing theory and practice reflect a serious attempt to build new types of tests that lend themselves to modes of interpretation referenced directly to content and/or performance. Unfortunately, our terminology often seems to blur important distinctions and equally important similarities between "new" and "old" approaches. A successful delineation of such critical differences and similarities should lead to conceptual clarification and perhaps contribute to the development of theory and practice in the measurement of achievement.

Glaser's (1963) discussion of norm (NRT) and criterion-referenced (CRT) testing emphasized the distinction between interpreting test scores in terms of what a person can do in terms of actual performance vs. how well a person does as compared to other people. The distinction has been

*Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

useful, but is now commonly applied in a way which suggests that a given test is invariably either in one category or the other. Inherent in Glaser's original formulation, and abundantly clear later on (Glaser and Nitko, 1971), is the fact that the distinction refers both to (a) the way in which test content is specified and (b) the kinds of interpretations that can be made of the resulting scores. Finally, in typical usage, the NRT vs. CRT distinction ignores a third type of interpretation that can be made of test scores--one that is referenced to content or performance directly, but which does not incorporate the notion of a "criterion."

All achievement tests, whether viewed by their developers as NRT or CRT, are in many situations interpreted both in terms of the "what" and "how well" question. By no means, for example, do we interpret the traditional standardized achievement test solely in terms of norms. To say that "Johnny scored at the 50th percentile for his age/grade group in terms of national norms" would immediately bring the response, "Scored at the 50th percentile on what?" The test turns out, of course, to have a title, but if the manual is adequate there will also be something referred to as a content-process matrix (C/P matrix) and even an index relating sets of items to categories in that matrix. Though obviously subjective and non-quantitative this information is just as relevant to the test's interpretation as the numerically expressed normative score.

The "what" aspect of the interpretation of the typical published standardized test unfortunately incorporates great areas of subjectivity and vagueness in (a) the way in which the content universe is specified via the content/process matrix (Cronbach, 1969), as well as in (b) the criteria used to develop items from that matrix (Ebel, 1962, Bormuth, 1972). In spite

of this, most users of achievement test data have been willing to take it on faith that publishers typically develop valid measures of educationally important universes of content.

Ebel (1962) demonstrated that content domains could be specified with greater rigor. His "content standard" score referred to the percentage of items answered correctly on a test made up of items sampled from such a domain. This formulation also anticipated contemporary developments in the construction and interpretation of tests of educational achievement in the sense of defining a numerical score referenced directly to content rather than indirectly to the performance of other individuals.

Just as content-referenced interpretations can (and often must) be applied to what are usually thought of as "norm-referenced" tests, so too are norm-referenced interpretations relevant to tests now being marketed as "criterion-referenced," "objectives-based," or "domain-referenced." There must be some basis for believing that such a test is appropriate for a given learner or group of learners. For example, Dahl, (1974 in preparation) observed that teachers often make major errors in leveling objectives for their students. This is not surprising, since many educational objectives are actually taught at differing levels of complexity at different grade levels, and teachers are also often unaware of the specific pattern to entry skills their students possess (Skager, 1969). Displaying sample test items is one way of helping the teacher level objectives more accurately. Providing appropriate normative information would be another, and probably simpler, method from the teacher's point of view.

More important, it is unrealistic to expect that a statement of the "what he can do" variety will in many circumstances be seen as sufficient.

Parents are likely to be interested in when (e.g., at what age or grade) the typical child "masters" a given universe of content. Evaluation reports, accountability studies, etc. cannot avoid referencing mastery interpretations to relevant comparison groups.

It is thus argued here that the notion that one type of test is necessarily interpreted comparatively in terms of other people and the other directly in terms of a universe of content is inaccurate, since such interpretations will be seen to apply to measures presently classified in both categories. What a number of researchers and theorists seem to be searching for are ways of formalizing and objectifying content-referenced interpretations to a degree that approaches the sophistication of existing comparative or normative interpretations. In other words, instead of a vague "content interpretation," it would be desirable as Ebel (1962) suggested, to have a score referenced to a content domain and expressed on a numerical scale.

The original distinction between criterion- and norm-referenced testing obviously anticipated certain pragmatic information needs arising in classrooms oriented toward what has come to be referred to as "mastery learning" (cf. Bloom, 1968). Theoretically justifiable procedures for formulating content-based decision rules relating to the management of instruction appear to be needed. Being able to determine with some degree of confidence whether a learner has mastered some domain of content would presumably contribute to the orderly and efficient movement of students through the curriculum in such classrooms.

A Classification System

A variety of distinctions can be made among contemporary achievement tests based on (a) the way in which content is specified and (b) the types of interpretations that can be made of the scores. However, an initial attempt to use these two characteristics to develop a comprehensive classification system, while useful as far as making distinctions was concerned, tended to obscure similarities that might exist between instruments falling in different categories. This problem was resolved by developing a third set of categories which reflected the various functions that tests might serve in the classroom.[1] It was then apparent that the specific way in which a test can be interpreted is determined by a particular combination of intended function and mode or strategy for specifying content. The table reflects these relationships.

Content Specification Mode

Modes for specifying the content of classroom tests fall into four general categories, the first being the familiar content/process matrix from which most achievement tests in use today originated.

The strengths and weaknesses of the C/P matrix are well known (cf. Cronbach, 1971). On the positive side, when properly utilized this approach does provide tests with broad content coverage and which are capable of making reasonably accurate distinctions between individuals. But the test developer really cannot know in advance what sorts of mental processes examinees will actually utilize in arriving at the answer, nor be confident that all examinees will use functionally equivalent processes. Partly as a

[1] I am indebted to Chester Harris for suggesting this approach. I am also greatly indebted to Robert Brennan, Robert Ebel, and my colleague Richard Shavelson for a variety of other pertinent suggestions.

result, concern may have shifted away from attempting to describe cognitive processes to a pragmatic emphasis on careful specification of the nature of the correct response and the conditions under which the response is to be elicited.

The C/P matrix is also an imprecise specification strategy in terms of its ability to define the limits of the intended content domain. In light of the uses for which most contemporary tests were designed, content coverage usually tends to be quite broad. Moreover, great latitude is left up to item writers in the determination of what the categories of the C/P matrix actually mean. Different item writers working independently might construct non-parallel tests from the same C/P matrix (cf. Cronbach, 1969). Finally, it may often be difficult to decide whether or not a given item belongs uniquely in a specific cell of a C/P matrix.

These problems have not inhibited the development of quantitatively meaningful norm-referenced score interpretations. They do, however, place severe limits on the kinds of content-referenced interpretations that can be made as well as on their precision. For example, if the content domain is not precisely specified, it is pointless to attempt to define mastery of that domain.

A second means of specifying test content is provided by the theoretical construct. This term is used in the usual sense--in reference to hypothesized personal characteristics, referenced to one or more psychological theories, which, in turn, explain consistencies in the behavior of individuals in a variety of situations. Classroom tests measuring constructs such as intelligence, aptitudes, and perhaps cognitive styles, are familiar. However, generalized patterns of achievement also may be formulated as

theoretical constructs. Cronbach's (1971, p.463) definition of reading comprehension, which either explicitly or by implication excludes vocabulary, reading speed, general information, etc. as irrelevant to the construct is a useful illustration.

Theoretical definitions of constructs, as specific as that developed by Cronbach, serve as guides for writing test items. But theoretical constructs are not likely to provide precise specifications in this regard, because they refer to generalized characteristics or traits which can be measured in a variety of ways. Item writers working independently from the same construct could easily produce non-parallel tests, especially in the sense of having scores influenced by different kinds of "method" variance (cf. Campbell and Fiske, 1959). Clearly defined constructs focus on behaviors representative of the construct. The theory in which the construct is embedded deals with cognitive or affective processes, but, unlike the C/P matrix, the construct focusses on what can be observed and measured.

The function of the theoretical construct is thus deliberately not that of defining a precise content domain. Because it is embedded in theory, it must relate to other constructs. No matter how many construct validity studies are done, there may always be another plausible interpretation of scores on a test measuring a given construct. Cronbach suggests,

> "It might sound as if construct validity is either present or absent, but most studies lead to an intermediate conclusion. The reading test may truly require comprehension, but it also makes demands on vocabulary." (p.465)

Interpretations referenced to a precisely delimited domain of content are not appropriate for tests derived from theoretical constructs. This is not a liability, as will be evident in the later discussion of the purpose for which such instruments are likely to be used.

The third approach to defining the content of an educational test is now commonly referred to as objectives-based. The behavioral or performance objective specifies (a) the conditions which will confront the examinee and (b) the observable behavior on his part which can be taken to constitute a correct response to those conditions. Skager (1974 , p.47, footnote) in an earlier paper suggested that the inclusion of a third element advocated by some--an arbitrary criterion of mastery--is probably inappropriate. A test built to measure a given objective may be of different lengths depending on the purpose for which it is to be used. Further, there is the already alluded to tendency to confuse the concept of a criterion or standard with the separate question of how test content is to be specified.

Objectives-based test materials are presently being marketed by several test publishers in commercial delivery systems. While these systems take different forms with different publishers, they represent a new generation of educational assessment instrumentation.

The real question, however, is not whether objectives-based systems represent something new, which they most certainly do, but rather with how far the behavioral objective can take us in the direction of providing test scores which are susceptible to direct, content based interpretations. Millman (1974) has recently reminded us that behavioral objectives typically leave much latitude up to the item writer. The specificity of objectives currently in use also varies widely. Sample objectives from the National Assessment of Educational Progress listed by Wilson (1974, p.30) are behavioral in the sense of referring to observable actions (though in very general terms) without incorporating specifications about conditions. NAEP objectives are supplemented by "exercise prototypes" specifying response

mode and other conditions as well as by sample exercises designed to provide guidelines. These additional specifications, while made by committees of experts and extensively reviewed, are to some degree arbitrary since another panel of experts might have generated somewhat different specifications.

Dahl (1971) demonstrated that judges rarely if ever made errors when asked to classify randomly grouped items under the objectives they were written to measure. It seems unlikely that the level of accuracy would be the same if judges were asked to classify test items in the appropriate cells of a typical C/P matrix. The behavioral objective undoubtedly has a great advantage in terms of clarity of specification. Also, this particular content generation mode makes no attempt to specify the process by which an examinee is to obtain the correct answer. But it is still reasonable to argue that objectives defining content domains containing many items may not define those domains uniquely. Rational analysis must be used to derive sets or systems of interrelated objectives from broad subject-matter areas. Each and every objective represents a decision about what is important. The arbitrariness interwoven into this process is self-evident.

Millman (1974) describes Popham's attempt to provide practical but reasonably precise guidelines for generating items from objectives. This author's "amplified" behavioral objectives are supplemented by statements describing the testing situation, the characteristics of the response alternatives, and the criteria for scoring. One critical difference between Popham's amplified objective and Hively's item form to be discussed next is that the former does not include replacement stimuli. While the rules also are looser than those formulated by Hively and his associates, amplified objectives do appear to offer significantly more guidance to the item

writer than do ordinary behavioral objectives. The criticism that those rules were derived arbitrarily is still relevant.

It is also evident that amplification does not rule out the possibility that a given item or set of items will be defective from a technical point of view. If the rules for writing items are faulty the items will also be faulty. Thus, the amplified objective of Popham's provided by Millman (1974, p.34) for illustrative purposes (a) contains a specific determiner (correct answer inevitably a longer, less commonly used word than incorrect answer), (b) has the examinee putting an "X" (in effect, crossing out) through the correct rather than the incorrect word, and (c) has instructions to the examinee which may not communicate very accurately what is intended in the objective. It is perhaps easy to forget that the behavioral objective, even when amplified, does not circumvent the problem of technically defective items. Brennan (1975) has already called our attention to this issue as well as explored alternative procedures of item analysis appropriate to tests developed from objectives.

The last content specification mode incorporates procedures or models proposed by various authors, all of which involve the development and utilization of formal item generation rules. While diverse both in approach and specificity, all have the common intent of achieving a logical, systematic, and replicable means for generating t est items representative of a defined content domain. All in one way or another appear to devolve at least in spirit from Ebel's (1962) concept of the "content standard" test score. The latter was to be directly referenced to a set of tasks defined so systematically that "...independent investigators would obtain substantially the same scores for the same persons." (p.16) Ebel also described

a vocabulary test developed by applying systematic rules for sampling words from a dictionary by way of illustration, although the example itself admittedly did not go very far in exploring the potential of the approach.

Hively and his associates (Hively, et al, 1973) have developed perhaps the best known item-generation model within the context of a curriculum evaluation project. It is significant that a statement taken from an early project working paper written by Hively and quoted in the 1973 monograph reflects the goal of quantifying content interpretations quite explicitly.

> "The basic notion underlying domain-referenced achievement testing is that certain important classes of behavior...can be exhaustively defined in terms of structured sets of domains of test items...precise definition of a domain and its subsets makes statistical estimation (Italics mine) possible."(p.15)

Coming as it did out of the evaluation of a particular instructional program, the approach that Hively and his co-workers eventually developed involved an initial process of eliciting from developers of a mathematics curriculum statements about curriculum objectives. These statements ultimately were transformed into definitions of content domains which included (a) general descriptions of the task (sometimes in a form close to that of a behavioral objective), (b) statements about characteristics of the stimulus and response, (c) one or more "item form cells" defining each class of items in the domain (with classes grouped together because the same set of generation rules can be applied to each), (d) the "item form shell" which gives rules for constructing item variations from the one or more "replacement sets" of stimulus elements. Each of the latter, in turn, was referenced to a particular item form cell. Scoring specifications were also provided.

There is something arbitrary in this process of making decisions about the particular item form and the specific elements of the replacement

sets. This arbitrariness is at least analogous to the kind of decisions that are made (obviously less explicitly) by the item writer working directly from a behavioral objective without benefit of generation rules. This is certainly recognized by Hively, et al., 1973.

> "Even the simplest concept or skill has so many potential 'representative' behaviors that it is impossible to specify them all. Arbitrary limits to the population must be imposed." (p.15)

But one must credit Hively and his associates for developing a model which not only renders the results of such decisions open for all to examine (though not the reasoning behind them) and which is genuinely capable of objectifying the item generation process to the point where item writers working independently should be able to produce parallel tests. Defining a content domain that clearly, especially for tasks which appear to be non-trivial in the educational sense, is a significant achievement.

Obviously questions arise as to the appropriateness of the Hively model for content domains that are considerably less structured than mathematics as well as in developing tests measuring functions at a high level of the cognitive taxonomy. (The latter criticism has also been made of tests derived from the C/P matrix and the behavioral objective, e.g., Ebel, 1971). These questions are not the primary focus here, but they certainly bear on the extent to which the model will be used. Likewise, the sheer amount of work and expertize that must go into generating a significant number of item forms raises questions about cost effectiveness, although admittedly alternate forms of the test can be generated virtually automatically once the form is constructed.

Apparently even an approach to content specification as rigorous as Hively's can still result in items and tests with traditional types of

technical defects. The particular item form chosen by Hively (1973, p.24) to illustrate his approach may result in items which do not really assess (at least for some examinees) the competency identified in the general task description for the item form. In this particular instance it is possible that examinees could produce correct responses without understanding or being able to generalize the concept being assessed.

A second approach to the generation of items by systematic means has been advanced by Bormuth (1970), and there is a link between his and Hively's work. Bormuth has been especially concerned with tying the achievement test item as closely as possible to instruction by going directly from verbal instructional content without the intermediary of behavioral objectives and "idiosyncratic" decisions by item writers.

> "To develop a science of achievement testing, the procedures for deriving items from the instruction must be operationalized. One way to do this is to regard the test item as a property of instruction and the item as being obtained by performing some manipulation on the instruction. Thus, an operational definition of a class of achievement test items is a series of directions which tell an item writer how to rearrange segments of the instruction to obtain items of that type." (p.5)

Bormuth's approach utilizes linguistic principles to derive various item transformations from instructional content. Items are to have a logical relationship with instruction. It should be possible to state the "...exact manner in which the structure of the test item is related to the structure of the relevant segment of the instruction" (p.14). Empirical evidence that the item is sensitive to instruction is seen as superficial, in that it deals only with "...observations of responses"(p.14).

There is another interesting difference in approach which contrasts Hively and Bormuth with Popham and the developers of most objectives-based assessment systems. Hively and Bormuth derive test content directly from instructional materials and statements. Bormuth is especially explicit, even

militant, on this point. He strongly objects to contemporary evaluation systems which provide items measuring behavioral objectives derived from abstract analyses of content domains. Bormuth maintains, for reasons that are not entirely clear, that teachers should not be led to shape instruction in the direction of maximizing performance on such objectives. There is a difference of opinion here which would make for an interesting debate. Many educators have maintained for some time that much instruction in the schools goes on without clearcut objectives. Tests produced by analysis of actual instructional content might be content valid, but fail in many cases to meet the addition validity criterion of "educational importance" described by Cronbach (1969). Still, Anderson's (1972) point is well taken. If we are to measure whether or not the learner comprehends actual instruction, then, "...a system of explicit definitions and rules to derive test items from instructional statements..." is highly desirable (p.149). The general utility of approaches utilizing transformational and other grammars should continue to be examined, even if such approaches are limited to instruction presented via what Shoemaker (1975) refers to as the "natural language" (p.134).

Bormuth's formulations are also subject to questions about efficiency and practicality, as well as about generality of application. But he does suggest another path toward the precise definition of content domains which yields rigorous and direct (non-comparative) interpretations of performance.

It is now appropriate to relate the four basic modes of specifying content to the functions for which tests are used in the classroom.

Functions of Testing in the Classroom: Managerial

There are two major functions for which tests are used in the classroom--

poorly on a unit posttest, the teacher may make an evaluative comparison with another child who answered all of the questions correctly. Summated over the time such information may be used for evaluating either the instruction or the learner. But this kind of evaluation has nothing to do with deciding whether to assign extra practice in the same learning mode or to select a different approach to instruction for the unsuccessful child. Performance alone is sufficient. It is directly interpretable.

The use of tests for (b) and (c) above is generally well understood, though neither as widely or as systematically practiced as might be hoped. However, applying tests diagnostically to assign instructional modes optimal for given learners is at present more hope than reality. Hambleton (1974) in his review of three of the most widely disseminated individualized instructional programs concludes, "...while nearly all developers of individualized programs describe this feature, there are few demonstrations of

---

[2]The extensive use of tests in the schools for purposes not directly related to instruction is irrelevant to this discussion. However, testing for guidance or clinical diagnosis is analogous to diagnostic testing in the classroom context. Using tests for purposes of selection has its analogue in learner evaluation.

[3]Usage of the terms "diagnosis" and "placement" follows that of Glaser and Nitko (1971). Other authors, e.g., Cronbach (1971) have used these terms differently.

should be interpreted turns out to be a function of (a) the mode by which

significant interactions between aptitudes and instructional modes" (p.393). But the function itself is potentially of great importance, even if knowledge lags behind instructional theory.

Aptitude tests in the past have been seen as likely candidates for diagnostic use. Measures of cognitive styles, falling in the region between aptitude and personality, may also be promising, and Cronbach (1975) has argued recently that pure personality measures may have greatest promise of all. It is even conceivable that achievement tests could be used for diagnostic purposes, not in the sense of establishing entry skills for placement, but rather as indicators of potential transfer effects from a different learning domain that might interact with an instructional mode. Competency in the English language, for example, is a relevant basis for assigning children to monolingual or bilingual classrooms.

Tests used for diagnostic purposes have to be constructed so as to differentiate among groups of students. Determining whether or not a test will be useful for diagnosis involves prediction studies, specifically the search for regression lines (achievement on predictor) which cross for different instructional modes. However, recently Cronbach (1975) has warned that actual relationships may be considerably more complex than simple first order interactions. The theoretical construct is the most likely content-generation mode for diagnostic tests, although the C/P matrix cannot be ruled out, particularly if any generalized measures of achievement turn out to be useful in this function.

We can now turn to the formative and placement use of tests in instructional management. Placement tests are likely to be relatively long because they typically cover a spectrum of instructional objectives. The three

well-known instructional models reviewed by Hambleton (1974) (IPI, PLAN, and Mastery Learning) all were organized around "...a curriculum defined in terms of behavioral objectives arranged into small clusters or units around a common topic or theme" (p.392). Formative tests, (referred to as "diagnostic-progress" tests by Hambleton), are shorter instruments designed to assess one or more objectives within a unit of instruction.

The most appropriate content specification modes for these two types of instruments would be behavioral objectives or formal item generation rules, since precision in the definition of the content domain is highly desirable. The question being asked in the classroom is whether or not the learner has mastered the domain in question. While formal item generation rules give more precision in the sense that the particular form selected for the items is explicit, this does not necessarily mean that a clearly stated objective accompanied by a sample item would not provide a definition of the domain adequate for the typical test user.

Relationships between managerial functions and content specification modes are shown by "X's" in the upper portion of the table. Thus, the C/P matrix and the theoretical construct are identified as the most likely content generation modes for tests used for diagnosing which instructional treatment is most appropriate for given learners. Objectives or formal item generation rules are seen as appropriate modes for generating test content in the case of placement and formative functions. There is no intent here to portray one content generation mode as superior to the others. Each has its uses, but there are ties between function and how test content is likely to be specified.

Functions of Testing in the Classroom: Evaluation

There are really two different evaluative functions within the classroom.

The first involves evaluating the learner for grading, promotion, awards, and the like. The second involves evaluating the instruction, and in the decade of "accountability," perhaps even the teacher. This differentiation is made in the left hand margin of the Table. Realistically, it must be admitted that in most classrooms evaluation tends to focus on the former. The same types of measures are used for both, although obviously the tests themselves might differ in certain characteristics depending on whether or not group or individual data is needed, whether matrix sampling is appropriate, etc.

In contrast to the situation for managerial decision-making, tests developed under any of the four basic content generation modes can be used for evaluation as indicated by the "X's" in the Table. This by no means implies that content generation mode is irrelevant in developing evaluation instruments. The particular evaluative question to be asked in a given situation will vary depending upon the philosophy of evaluation held by whoever will interpret those data. The nature of the question in turn relates directly to content generation mode. For example, if the evaluation focusses on individual learners and the question to be answered is how well those learners stand on the generalized objectives of the course or how well they are able to transfer what they have learned to new situations, then the C/P matrix would probably be chosen because of its simplicity and generality. If, on the other hand, the question is phrased in terms of how many of the specific objectives of the curriculum have been mastered in a given period of time, or in terms of Brennan's (1975) notion "instructional time" (how long it takes the learner to master a given objective or set of objectives), then the C/P matrix or the theoretical

construct are clearly not appropriate content specification modes. Tests based on objectives or formal item generation rules are needed if these kinds of questions are to be answered. Clearly, while any of the four types of tests can be used for evaluation of learners or of instruction, there is a close relationship between the nature of the evaluative question posed in a given situation and the mode by which test content is most appropriately defined.

Having looked at the relationships between content generation modes and the functions for which tests are used, it is now time to turn to the matter of how scores on the four types of tests may be interpreted. This is undoubtedly the area in which contemporary terminology and the underlying conceptions which it represents leads to the greatest confusion about distinctions between different types of tests.

Interpreting Test Scores: Domain vs. Norm-Referenced Interpretation

There appear to be two fundamentally different ways of interpreting test scores. The first is labeled here as a domain-referenced (DR) interpretation and refers directly to content or performance without regard to comparisons among individuals. In contrast, norm-referenced (NR) interpretations derive their meaning from the relative standing of individuals compared to one another. This distinction is obviously not new, although it should be noted that the term "criterion-referenced" has not been used at this level of generality.

The last two rows of the table list various types of DR and NR interpretations that have either been in use for some time, or whose use is conceivable given the newer approaches to the generation of test content. The types of score interpretations have been listed in relationship to

the columns of the table corresponding to content generation modes.

Perhaps the most important observation to be made here is that DR interpretations have long been available for traditional types of tests whose content is derived from C/P matrices or theoretical constructs. The DR row under these two content specification modes lists the familiar expectancy score which, as Cronbach (1970) suggests, refers to actual performance rather than to comparative standing. Even the predicted grade point averages provided by some college admissions testing programs are thus subject to direct, rather than comparative, interpretations, e.g., the probability of having a "C" average or better at the end of the freshman year at college X. Likewise, the representative item cluster score describes Ebel's (1962) proposition to the effect that normative test scores should be supplemented by content-based interpretations based on displays of representative items typically passed by individuals obtaining various scores on the test.

These two types of DR interpretations can also be applied to tests generated from theoretical constructs. In addition this later mode is susceptible to interpretation in terms of absolute scores of the Guttman or Rausch variety. Tucker's (1953) proposition IV on the characteristics of an "ideal" test minimizing the importance of reference groups makes this clear.

> "The scores (on such a test) indicate extent or degree of some trait which exhibits homogeneity in the behavior of examinees" (p.27).

Angoff's (1971) discussion of Guttman, Rausch, and Tucker's models does not reflect any particular interest on the part of any of these

theorists as to how test content is to be generated initially. The emphasis is rather on whether a given set of items meets the various criteria of scalability. But Guttman's early work was in attitude measurement, again suggesting the theoretical construct. Absolute scales, while referring to difficulty in the case of achievement tests, do so independently of any population of examinees.

Finally, allowance should be made for the fact that diagnostic interpretations for the purpose of selecting instructional mode may be made from tests generated from these first two content specification modes. Here almost any kind of numerical score scale might be used, since the intent is to divide learners into two or more groups.

Several new types of DR scores are pertinent to content specification modes based on objectives or formal item generation rules. Ebel's (1962) content standard score, while proposed some time ago, remains the progenitor of this category of interpretations. Ebel used the term "domain" and his content standard score can be taken as a point estimate of the examinee's competency with respect to that domain. Cronbach's (1970) content reference score refers to "...level of performance on content that is like the test" (p.85). A score indicating how many words an examinee can type over a given period of time without making errors lends itself directly to incorporation into precise decision rules for training or selection.

With precise specification of a content domain one can envision two kinds of DR scores with very useful properties. The first is a score estimating the proportion of items in the content domain that would be passed by the examinee were all of the items to be administered. This is

referred to in the Table as a domain score estimate, after Millman (1974). The second would provide a "sign" interpretation as Harris (1974) has characterized the much-talked about notion of mastery, and is labeled mastery, domain-referenced. Here (at last) the much used concept of a "criterion-referenced" score obviously applies, although the term "mastery" seems much more descriptive of the particular type of criterion desired.

Whatever one's preferences may be with respect to terminology, it should be duly noted that the concept is one of a criterion-referenced score rather than a criterion-referenced test. Cutoff points reflecting decision criteria could be established for all of the DR scores discussed up to now, including those applicable to tests generated from C/P matrices. Millman (1974) even discusses what he terms the "criterion-referenced differential assessment device" or CRDAD. This is an objectives-based test, but one in which items have been selected for discriminating power. Scores on this type of test could be referenced to a criterion or standard but would no longer represent the content domain defined by the objective, since some proportio of the items in the domain have been eliminated.

However, only the last two content generation modes appear to be amenable to concept of a mastery criterion. Millman's (1972) paper used the test score as a point estimate of the domain score and applied the binomial theorem to get at the probabilities of correct and incorrect classification of examinees for different mastery criteria and for tests of different lengths. Harris (1974) illustrated the relevance of sequential testing procedures for fixed length tests which can be regarded as samples of items from a defined domain. Novick and Lewis (1974) applied the Bayesian approach to the same problem. Lewis, Wang, and Novick (1973)

applied Bayesian principles to estimating domain scores.

Davis and Diamond (1974) point out that "...mathematically, we regard it as impossible for an examinee who has complete knowledge of all items in the population to mark incorrectly any item drawn from that population" (p.134). In the abstract, mastery is perfect performance, nothing less. Practically, of course, we know that individuals who are in fact masters of some domain of content may answer items incorrectly because of carelessness, distractions, and the like. These authors all view mastery in terms of a domain of possible items. Hence the concept of a domain-referenced mastery criterion seems useful.

Statistical methods described by these authors could be applied to tests generated by any of the four content-specification modes in the Table. However, the resulting estimates would be seriously misleading in the case of the first two modes. Being able to pass even 90% of the items in the "domain" represented by a test generated through the use of a C/P matrix still leaves open the possibility that there are some skills measured by the test on which the examinee has no competence at all. Domain-referenced interpretations only become meaningful when there is an appropriate degree of specification of that domain.

Unfortunately, the typical rule for determining mastery on tests being published presently is quite arbitrary in the sense of being confined to the test itself rather than being referenced to the domain. "Eighty percent mastery" means getting eight out of tem items correct on the test, period. Because such interpretations are now commonly utilized, a second "sign" interpretation, mastery, test-referenced has been added.

NR type interpretations are mainly familiar, especially as they apply to traditional types of tests. The common varieties, percentiles, age/grade

equivalents, standard score scales, are listed under the first two columns of the table along with arbitrary scales. The latter term was used by Angoff (1971) to describe score systems tied to a convenient reference group rather than a systematic sample. The Scholastic Aptitude Test, referring to the 1941 examinee population, is a familiar example. A number of other types of scores are described in Angoff's exhaustive treatment of the topic, but discussing them would not contribute to the distinctions being made here.

Turning to the last two content-specification modes, it is apparent that NR, as well as DR, interpretations can be applied to tests derived from objectives or by means of item generation rules. Here again we must contradict the popular notion that there are two types of tests, one always interpreted in terms of norms, and the other in terms of a performance based criterion. Of particular interest would be (a) domain score estimates referenced to a normative scale (domain score norm), and (b) a mastery norm providing a comparative interpretation of mastery such as, "objective _____ is mastered by 50% of the _____ population at grade level 5.3." This kind of normative interpretation (whether in the form of a percentile or an age/grade norm) applied to a rigorously specified content domain, would be very useful. It would reflect what schools are accomplishing in a way that is tied both to instructional content and to relative standing. It could also summarize the teacher's evaluation of student performance in a manner that is far more informative than the maligned, but tenacious, letter grading system.

The choice of one kind of score interpretation over another obviously depends in part on the particular function for which the information is

to be used. However, the conceptions of management and evaluation held in a particular time and place also are determining factors. Modern approaches to the individualization of instruction such as those reviewed by Hambleton (1974) certainly stress the use of domain-referenced interpretations for placement and formative decision-making in the classroom. Indeed, instructional philosophy has provided the primary stimulus to the development of newer approaches to the generation of content and interpretation of test scores. Traditional approaches to classroom management can be expected to continue to be associated with a preference for NR interpretations, in spite of the fact that this type of information does not seem to be as useful for these two managerial functions.

In the case of evaluation, whether focussed on the pupil or the instruction, both DR and NR interpretations appear to be relevant because the two provide different and, taken alone, incomplete information. This observation was nicely illustrated in a recent newspaper article. It appears that the research branch of a large school district had published a report demonstrating that median percentile ranks on state mandated achievement tests for students in the district had remained at the same level or risen somewhat over the last few years. This was of course taken as a sign that the district was at the very least holding its own, and in some cases improving. The reporter, however, noted that raw score medians over the same period had actually gone down at most grade levels. In other words, students on the average were getting fewer questions correct, but the decline was not as precipitous as that occuring in other districts. Comparatively speaking the district came off rather well. The reporter had obviously stumbled on the utility of a DR type interpretation, although

no formal means for making such interpretations were available for the tests in question.

It should be clear that, taken alone, both NR and DR interpretations can be misleading, and equally so. In the case of evaluation one should be interested in both the "What?" and "How well?" questions. Mastery of all the goals and objectives might be achieved merely because those goals and standards were deliberately set low. Scoring above the 50th percentile might conceal real declines in achievement.

Content Specification Modes and Item Selection Strategies

In traditional test assembly both judgmental and empirical considerations enter into the determination of item quality. First, items are scrutinized for violations of traditional rules relating to the construction of achievement test items, such as those summarized by Gronlund (1968). Items found to contain specific determiners, for example, are modified or discarded.

It was suggested earlier that these familiar rules are also applicable to tests generated by the newer approaches to content specification. Traditional principles of item construction should be applied with care to items derived by means of the newer content generation modes, since additional empirical checks on item quality provided by difficulty and discrimination indices may not be appropriate. Certainly eliminating some items because they are "too easy" or "too difficult" for a given population, would invalidate a test as a measure of the domain defined by the objective or item generation rule.

Brennan (1974) has presented a variety of approaches to the analysis of what he terms as "criterion-referenced and mastery items." But where

items are derived from item generation rules these kinds of data may suggest that there is something wrong with the rules themselves. In other words, the domain itself might require redefinition. How this problem (should it arise) would be dealt with under an approach utilizing grammatical transformation of actual subject matter is an interesting question.

Items included in tests developed from C/P matrices and theoretical constructs obviously must discriminate among individuals. Technical decisions based on traditional procedures of item analysis, while appropriate and useful, do modify the content domain defined jointly by the content-specification mode and the item writer. Cox (1965) provides convincing empirical evidence of this fact.

Items measuring theoretical constructs should correlate with other items written to measure the construct. In factor analytic studies such items should also show a reasonable degree of independence from items presumably measuring different constructs. In the case of constructs defining domains susceptible to absolute interpretations, items must meet whatever scalability criteria are imposed.

It appears that none of the above statistical criteria are invariably relevant to the study of items from content domains defined by objectives or item generation rules. When the items in a given domain vary in difficulty for a given population of examinees, a high degree of homogeneity among items derived from these two specification modes would probably be observed. But as long as the items are congruent with the domain specification, high inter-item correlations are not a requirement, as Cronbach (1969) points out.

It appears, then, that determination of the quality of items based

on objectives or derived by means of item generation rules still calls for judgmental application of long-established principles of item construction.

Probably we have not yet had enough experience with tests based on objectives or derived from generation rules to write the final word how item quality is to be determined for tests of this type. It appears that procedures for assessing the quality of objectives-based or rule-generated tests and the domains from which they are derived are the next area to be explored in relation to the classification system proposed here. We refer, of course, to the analogues to traditional concepts of validity and reliability. Brennan's (1974) report as well as the ongoing investigations of Chester Harris and his students into approaches to item analysis which assess "sensitivity to instruction" are both highly relevant to this issue.

SUMMARY

The corpus of descriptive terminology associated with achievement testing has expanded considerably in recent years, in large part due to the heightened interest in absolute and/or direct metrics for interpreting test performance plus the development of more rigorous strategies for specifying test content. Widely prevalent disagreement about terminology reflects a lack of conceptual clarification and may inhibit the development of theory and practice.

Distinctions commonly made between "criterion" and "norm-referenced" tests turn out to be inaccurate, since it appears that both content- and norm-referenced interpretations can apply to scores on any type of achievement test. Rather, the particular manner in which a given test can and

should be interpreted turns out to be a function of (a) the mode by which test content is specified and (b) the function for which the test is to be used. Four content specification modes were discussed in this paper in conjunction with five functions, the latter classified as either managerial or evaluative.

All approaches to the interpretation of achievement test scores are classified as either "domain-referenced" or "norm-referenced," with reference to a criterion or standard viewed as a special case of the former. Finally, it is argued that normative interpretations can and in many instances should be made of scores which are referenced directly to content, including mastery scores.

****

List of References

Alkin, M. C. "Criterion-referenced measurement' and other such terms." Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation (3). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Anderson, R. C. "How to construct achievement tests to assess comprehension." Review of Educational Research, 1972, 42, 145-170.

Angoff, W. H. "Scales, norms, and equivalent scores." In R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971.

Bloom, B. S. "Learning for mastery." Evaluation Comment, Center for the Study of Evaluation, 1968, 2(1).

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Brennan, R. L. "A model for the use of achievement data in an instructional system." Instructional Science, 1975, 3, 1-24.

Brennan, R. L. "Psychometric methods for criterion-referenced tests." Final Report to the Research Foundation of the State University of New York. Albany, New York, 1974.

Cox, R. C. "Item selection techniques and evaluation of instructional objectives." Journal of Educational Measurement, 1965, 2, 181-185.

Cronbach, L. J. "Validation of educational measures." Proceedings of the 1969 invitational conference on testing problems. Princeton: Educational Testing Service, 1969.

Cronbach, L. J. "Test validation." In R. L. Thorndike (Ed.), Educational Measurement. Washington, D. C.: American Council of Education, 1971.

Cronbach, L. J. Essentials of psychological testing. (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. J. "Beyond the two disciplines of scientific psychology." American Psychologist, 1975, 30, 116-127.

Dahl, T. A. "Field trial of SOBAR materials: final report." (In Preparation), Center for the Study of Evaluation, University of California, Los Angeles.

Dahl, T. A. "The measurement of congruence between learning objectives and test items." Unpublished doctoral dissertation, University of California, Los Angeles, 1971.

Davis, F. B. and Diamond, J. J. "The preparation of criterion-referenced tests." CSE Monograph Series in Evaluation (3). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Ebel, R. L. "Content-standard test scores." Educational and Psychological Measurement, 1962, 22, 15-25.

Ebel, R. L. "Criterion-referenced measurements: Limitations." School Review, 1971, 79, 282-297.

Glaser, R. and Nitko, A. J. "Meas nt in learning and instruction." In R. L. Thorndike, (Ed.), E: onal Measurement. (2nd ed.) Washington: American Council ducation, 1971.

Gronlund, N. E. Constructing achievement tests. Englewood Cliffs, N. J.: Prentice-Hall, 1968.

Hambleton, R. K. "Testing and decision-making procedures for selected individualized instructional programs." Review of Educational Research, 1974, 44, 371-400.

Harris, C. W. "Problems of objectives-based measurement," in C. W. Harris, M. C. Alkin, and W. J. Popham, (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation (3), Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Harris, C. W. "Some technical characteristics of mastery tests." Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation (3), Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Hively, W., Maxwell, G. R., Sension, D., and Lundin, S. Domain referenced curriculum evaluation; a technical handbook and a case study from the MINNEMAST PROJECT. CSE Monograph Series in Evaluation (2). Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1973

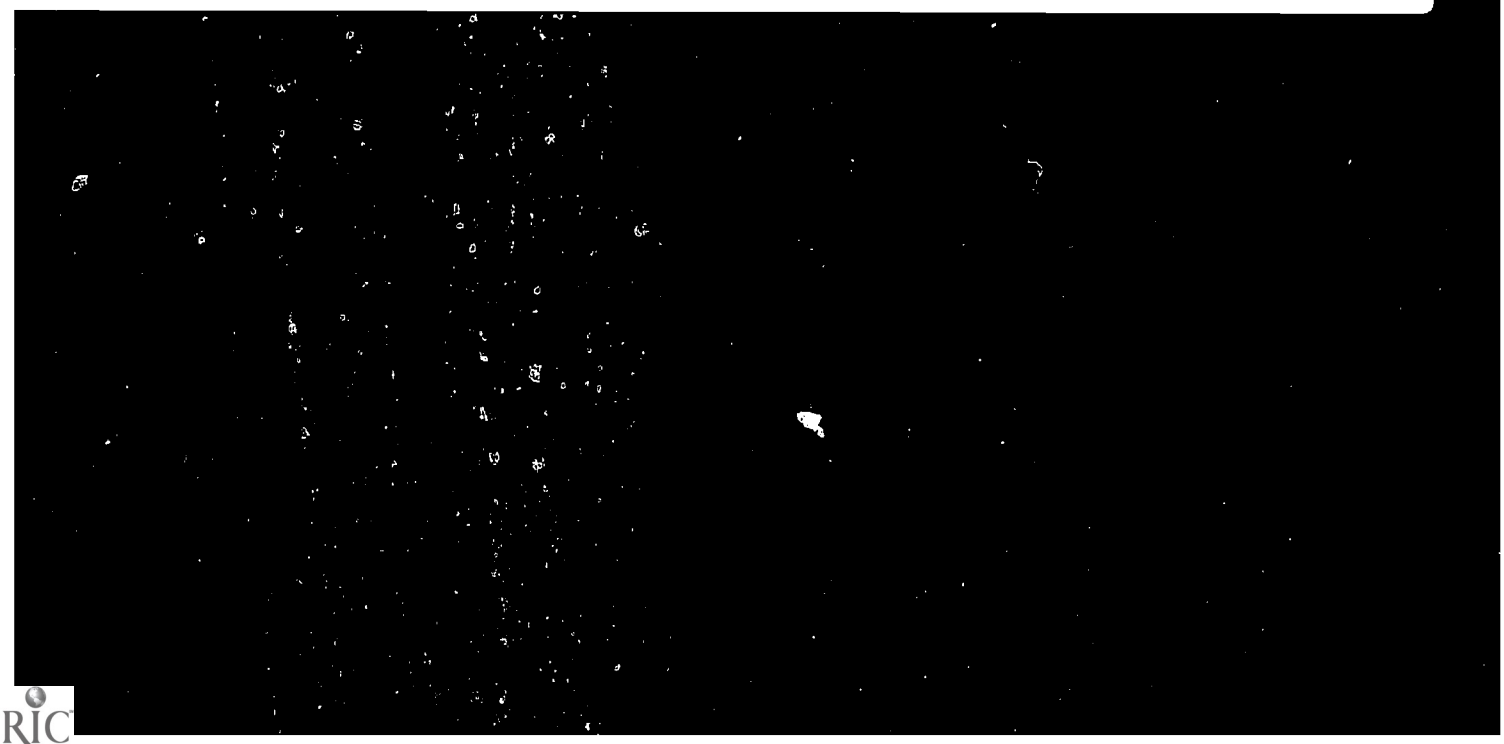
Lewis, C., Wang, M. and Novick, M. R. "Marginal distributions for the estimation of proportions in m groups." The Technical Bulletin, No. 13, Iowa City: American College Testing Program, 1973.

Millman, J. "Criterion-referenced measurement." In Popham, J. (Ed.), Evaluation in education: current applications. Berkeley: McCutchan, 1974.

Shoemaker, D. M. "Toward a framework for achievement testing." Review of Educational Research, 1975, 127-148.

Skager, R. W. "Generating criterion-referenced tests from objectives-based assessment systems: unsolved problems in test development, assembly, and interpretation." Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Skager, R. W. "Student entry skills and the evaluation of instructional programs: a case study." CSE Report No. 53, Center for the Study of Evaluation, University of California, Los Angeles, 1969.



A Classification System for Tests of Educational Achievement

CLASSROOM FUNCTIONS FOR TESTS	CONTENT SPECIFICATION MODE			
	Content/Process Matrix	Theoretical Construct	Objectives Based	Formal Item Generation Rules
Managerial				
Diagnosis	X	X		
Placement			X	X
Formative			X	X
Evaluative				
Learner	X	X	X	X
Instruction	X	X	X	X
DOMAIN-REFERENCED INTERPRETATIONS	Expectancy Representative Item Cluster (?)	(Same) (Same) Absolute Score Diagnostic Interpretation	Content Standard Content Reference Domain Score Estimate Mastery (Test-Referenced) Mastery (Domain-Referenced)	
NORM-REFERENCED INTERPRETATIONS	Percentiles Age/Grade Equivalents Standard Score Scales Arbitrary Scales		Domain Score Norm Mastery Norm	