THE GREAT CRITERION-REFERENCED TEST MYTH

Rodney W. Skager

CSE Report No. 95
January 1978

Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

THE GREAT CRITERION-REFERENCED TEST MYTH[1]

Rodney W. Skager

The word "criterion" may be the most overused term in the measurement vocabulary. This situation leads to a lack of clarity in conceptualization, especially in the notion of the "criterion-referenced" test. Probably no concept in measurement is more widely misunderstood by members of the wider educational public.

In the ordinary language the term "criterion" refers to "a standard on which a judgment or decision may be based." Webster tells us that it is derived from the Greek word kritērion, having to do with the making of judgments or decisions. In measurement we use a variety of such standards, especially when predictions about the future performance of individuals are made. Discriminating between individuals who are likely to be successful or unsuccessful amounts to what Harris (1974) referred to as a "sign," as opposed to a "score."

Another meaning of the term is more closely associated with measurement per se. It refers not to a standard or sign, but to a particular type of variable. Specifically, it is the variable against which scores on tests are compared in order to assess their predictive or criterion-related validity. Until recently, this particular interpretation of "criterion" has probably been the most common one within the measurement field. Still, the two separate meanings, sign and variable, existed in the vocabulary of measurement without leading to serious confusion.

---

[1]This is an expansion and revision of a paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April, 1975.

Our problems with the term criterion really began when Glaser (1963) separated tests into those that are norm-referenced (NRT) and those that are criterion-referenced (CRT) as a way of emphasizing the distinction between scores that can be interpreted in terms of what a person actually can do vs. how well a person does compared to other people. Glaser was concerned about the widespread failure to use tests or other assessment devices as tools in the instructional process. He called for types of tests whose scores would be directly interpretable in terms of some defined domain of educational content. This kind of test, as well as an ideological climate supportive of its use, Glaser saw as necessary for implementing what he and a colleague later referred to as "adaptive" instruction (Glaser & Nitko, 1971). This latter idea was also not really new; it was certainly embedded in the much earlier work of Washburn and Morrison. Within the field of measurement, the idea of direct measurement had also been advanced in the early paper of Cattell (1944) under the rubric of "interactive" measurement; that is, measurement that yields a score that is directly interpretable without reference to an individual's relative standing in a reference group. But the idea is even older than that. Buros (1977) has recently reminded us that a distinction between tests designed to "measure" and tests designed to "differentiate" goes back to the beginning of modern educational measurement.

Glaser's recycling of the notion of tests directly referenced to what a person can do in defined areas of educational content apparently advanced an idea whose time had finally (or once again) arrived. Developments in educational technology promised, according to Block (1971), to provide systems which could support the greater information demands of adaptive forms of

instruction such as learning for mastery, individually prescribed instruction (also associated with Glaser), and other variations on the theme. Moreover, the social context was right. The role of the school as a device for winnowing the educational chaff was to be sharply condemned, especially by spokesmen for groups that had in the past been winnowed to a disproportionate degree. Schools and teachers were even to be held accountable for making sure that the great majority of learners emerged equipped with the basic skills and competencies that comprise the curriculum.

Glaser chose to refer to his new/old type of tests as criterion-referenced. Yet, the Scholastic Aptitude Test, that arch tool of educational elitism, is clearly criterion-referenced when an admissions officer at a university uses predicted grade-point average to decide whether or not a given applicant should be admitted. This falls clearly within common usage of the term "criterion." A typing test is criterion-referenced when words per minute are used as a criterion for the employment of a typist. The latter meaning is quite close to Glaser's intention. The former is very far removed. The Scholastic Aptitude Test definitely cannot be directly interpreted in terms of a defined domain of educational content. It was built for quite different purposes.

The idea that by nature some types of tests are criterion-referenced and others are norm-referenced has carried the day. This idea is erroneous. It is only the interpretation of tests that can be criterion- or norm-referenced. Any test of educational achievement, whether generated from highly specific behavioral objectives (in the new style) or from highly inclusive and perhaps vaguely defined test plans or broad theoretical constructs (in the old style), can be interpretated from either approach. It is even possible, and the heresy

perhaps may be excused in the interest of discussion, for a test generated from the most impeccable behavioral objective, item form cell, transactional grammer, or whatever, to be norm-referenced.

Inherent in Glaser's original formulation, and abundantly clear later on in Glaser and Nitko (1971), is the fact that the CRT vs. NRT distinction refers both to (a) the way in which test content is specified and to (b) the kinds of interpretations what can be made of the resulting scores. Moreover, in typical usage, the 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."

Achievement tests, whether viewed by their developers as NRT or CRT, are in many situations interpreted in terms of both the "what" question and the "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 title should reflect what the test is supposed to measure, but if the manual is adequate there will be a test plan or a "content-process matrix" (C/P matrix) and perhaps even an index relating individual items to categories in that matrix. Though obviously subjective, this information is just as relevant to the test's interpretation as the numerically expressed normative score.

The "what" interpretation of the typical published standardized test is necessarily vague because of (a) the way in which the content universe is specified via the content/process matrix (Cronbach, 1969), and (b) the

considerable latitude left up to those who develop items from the matrix (Ebel, 1962; Bormuth, 1970). In spite of this, most users of achievement tests have been willing to take it on faith that publishers do develop valid measures of educationally important universes of content.

Ebel (1962) demonstrated that content domains could be specified with greater rigor than that provided by the C/P matrix. 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 by 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, in earlier work on the development of an objectives-based assessment system, my co-workers and I observed that teachers often made major errors in leveling objectives for their students. This is not surprising, since some types of educational objectives are taught at differing levels of complexity at different grade levels, and teachers are also often unaware of the specific pattern of entry skills their students possess (Skager, 1969). Displaying sample test items associated with a particular level of performance (as suggested by Ebel) is one way to help teachers level objectives more accurately.

Providing appropriate normative information would be another, perhaps simpler method, at least from the teacher's point of view.

The leveling problem is one sign that the very excellent concept of learning for mastery has been greatly overgeneralized with respect to its applicability to all educational content. Bloom (1968) shares at least some of the responsibility for this in his assertion at the beginning of a landmark paper that about 95% of the learners in elementary and secondary school could master the entire curriculum given enough time and favorable conditions for learning. This turns out to be a true statement, in my view at least, only if one restricts the notion of content to domains that can be defined very specifically, so specifically that it is possible to say with certainty that a given learner has achieved mastery. I have referred to such domains as "specified" domains (Skager, 1978), avoiding, on the advice of a number of colleagues, the negative connotations of the term "closed." By specified, I mean circumscribed content objectives that can be learned so well that further improvement is either impossible or pointless. Learners can answer virtually all of the questions on the test correctly.

A number of writers have been unwilling to accept the notion that virtually all of the school curriculum deals with specified (or specifiable) educational goals. In his analogy of the three stage rocket, Cronbach (1971a) saw specified objectives as the foundation of the curriculum. He also argued that the school curriculum deals with higher levels of achievement for which mastery is not an appropriate concept. Earlier, Eisner (1968) struck the same theme in his distinction between instructional and expressive educational objectives.

Many domains of educational content are "open-ended." Further improvement is always possible, at least for some learners. In such domains there may be no single "right" answer or unique approach to solving problems. We may be able to define knowledge of the calculus in such a way that many or most learners will be able to attain mastery. But the insightful application of the resulting knowledge is neither predictable nor limited in any way. Even reading comprehension, seemingly a basic skill, can be defined in an open-ended fashion. Readers can be asked to draw inferences, the difficulty and subtlety of the material can be increased, and so on.

Reading comprehension is a useful illustration, because it is easy to see how adopting different modes for specifying test content could lead either to specified or open performance domains. This is an important point. The way in which test content is defined and represented in assessment materials is directly related to the distinction between open and specified performance domains. For example, if one chooses to define reading comprehension as a psychological construct, as Cronbach (1971b) did in one example, there is really no limitation placed on the level of performance elicited by a test built according to the guidelines provided by the construct. On the other hand, if one were to define reading comprehension by means of the kinds of applified behavioral objectives illustrated in Millman (1974), reasonably precise limits could be drawn and the criterion of mastery would be applicable. That is, the difficulty of the vocabulary, the sentence complexity, and other parameters could be specified. One would have cut a circumscribed region out of the potentially larger domain of reading comprehension.

But even when the domain is carefully specified, it is unrealistic to expect that a statement of the "what he can do" variety will in many circumstances be accepted 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.

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 thus inaccurate, since such interpretations can be applied 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.

Likewise, it appears that the crucial variable distinguising between the two fundamentally different types of tests is the way in which the content domain is specified, rather than the use of norm- vs. criterion-referenced interpretations. Some modes of content specification lead to specified domains and mastery interpretations. Other modes of specification lead to open domains where the mastery concept does not apply. A conceptual scheme is needed for sorting out distinctions between different types of tests, especially with respect to the kinds of interpretations that can be derived from their scores.

A CLASSIFICATION SYSTEM

Educational achievement and aptitude tests can be described in terms of (a) the functions for which they are used, (b) the ways in which their content domains are specified, and (c) the ways in which they can be interpreted.[2] A fourth distinction, whether or not the test measures an open or a specified content domain, is a direct function of (b). Taking the function for which the test is to be used as the primary dimension, it turns out that certain content specification modes are more appropriate for certain functions than others and that particular types of score interpretations go with particular function/specification mode combinations. What follows is an attempt to show how this is the case. This discussion is limited to the use of educational tests in the evaluation of learners. It does not refer to the use of tests in the evaluation of the conditions under which learning occurs.

Functions of Testing

There are two groups of functions for tests in the evaluation of learners. These are referred to as "formative" and "summative" after Bloom, Hastings, and Madaus (1971).

---

[2] I am indebted to Chester Harris for suggesting that one should begin with function. I am also greatly indebted to Robert Brennan, Robert Ebel, and my colleague Richard Shavelson for a variety of other pertinent suggestions.

Formative Evaluation

Formative evaluation depends an assessment of learners for making decisions about (a) the instructional mode to be utilized, (b) the entry point in the curriculum, and (c) the progress of learners through the curriculum. The terms "diagnosis," "placement," and "diagnostic/progress" are used in Table 1 to refer to these three separate formative uses of tests.[3]

Insert Table 1 about here

How tests are used for (b) and (c) above is generally well understood, though neither as widely nor as systematically practiced as might be wished for. However, applying tests diagnostically to assign instructional modes optimal for given learners is at present more hope than reality. In his review of three of the most widely disseminated individualized instructional programs, Hambleton (1974) concluded, "...while nearly all developers of individualized programs describe this feature, there are few demonstrations of 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.

---

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

TABLE 1

Classifying Educational Tests by Function, Content Specification Mode, and Interpretation

FUNCTION	CONTENT SPECIFICATION MODE				
	Open Domains		Specified Domains		
	Content/Process Matrix	Theoretical Construct	Criterion Sample	Objectives-Based	Item Generation Rule
Formative					
Diagnosis	✓	✓			
Placement	✓			✓	✓
Diagnostic/ Progress				✓	✓
Summative					
Certification/ Crediting	✓		✓	✓	✓
Selection/ Prediction	✓	✓	✓		
DOMAIN-REFERENCED INTERPRETATIONS	Representative Item Cluster	Absolute Score		Content Standard Content Reference Domain Score Estimate	
CRITERION-REFERENCED INTERPRETATIONS	Expectancy/ Prediction	Expectancy/ Prediction Diagnostic/ Interpretation	Expectancy/ Prediction	Mastery	
NORM-REFERENCED INTERPRETATIONS	Percentiles Age/Grade Equivalents Standard Scores Arbitrary Scales	Percentiles Age/Grade Equivalents Standard Scores Arbitrary Scales		Domain Score Norm Mastery Norm	

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 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 (Cronbach & Snow, 1977). 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, Cronbach (1975) has warned that actual relationships may be considerably more complex than this simple first order interaction.

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 Learning for Mastery) 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). "Diagnostic/progress" tests are shorter instruments designed to assess one or more objectives within a unit of instruction.

Instruments used for placement and diagnostic/progress decisions would ordinarily be expected to assess relatively specific content domains. In

fact, variations on the adaptive instruction theme all appear to be directed primarily at the achievement of specified goals. Such instruments would ordinarily be constructed in a way that would make it possible to assess mastery of a domain. However, at least one variety of adaptive instruction, Project PLAN, uses predictions for placement purposes (Hambleton, 1975). This would require instruments that differentiate among learners rather than instruments that assess mastery.

Summative Evaluation

There are two basic kinds of summative decisions made about learners. The first is referred to as "certification/crediting," meaning that some end point has been successfully reached in a learning process. One associates the term "certifying" with the competency to perform some kind of relatively complex task such as a job. Most professionals are certified in some way. "Crediting" refers to a record of the fact that some kind of education has been completed. People obtain credit for courses or for completing high school or college, but the credit does not imply that they can perform some specific kind of job or other life-role.

The second summative function of testing in learner evaluation has been labeled "selection/prediction." This function has to do with selecting people for special opportunity, mainly though an implicit or explicit prediction about future performance. Certification/crediting is analogous to a "sign" interpretation referred to earlier. People are either qualified or unqualified. As long as a distinction can be made between those who do, and those who do not,

meet certain standards, further differentiation among individuals is unnecessary. In order to make predictions, however, instruments or other modes of assessment which differentiate among individuals along a continuum are usually required. This difference between the two types of summative evaluation functions will be shown to have implications for the selection of content specification mode.

Modes for Specifying Test Content

Modes for specifying the content of educational tests fall into five categories, the first being the familiar test plan or C/P matrix from which most achievement tests in use today originated.

The Content/Process Matrix

The strengths and weaknesses of the C/P matrix are well known (cf. Cronbach, 1971b). When properly utilized this approach does provide tests with broad content coverage capable of making reasonably accurate discriminations between individuals. But the test developer cannot know 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 result, concern may have shifted 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. It defines the limits of the intended content domain only in a very loose way. 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 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 meaningful types of norm-referenced score interpretations. They do, however, place severe restrictions on the kinds of content-referenced interpretations that can be made as well as on their precision. If the content domain is not precisely specified, it is pointless to attempt to define mastery of that domain.

The Theoretical Construct

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, perhaps referenced to a psychological theory, which in turn explain consistencies in the behavior of individuals in a variety of situations. 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 (1971b, p. 463) definition of reading comprehension, which either explicitly or by implication excludes vocabulary, reading speed, general information, etc., as irrelevant to the construct was already cited.

Theoretical definitions of construct as specific as that developed by Cronbach serve as guides for writing test items. But theoretical

constructs are not likely to provide precise specifications, 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 statistically non-parallel tests, especially in the sense of having scores influenced by different kinds of "method" variance. 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 the construct points to what can be observed and measured.

The function of theoretical constructs is not one of defining precise content domains. Because constructs are the building blocks of theory, they 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 has suggested,

> "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, 1971b)

Criterion Sampling

A third approach to the definition of test content was advanced by McClellan (1973) in the notion of "criterion sampling." This approach is presumably suited to the assessment of highly complex performance domains such as those represented by most occupational roles. The idea is to develop measures of significant aspects of such roles in order to assess relevant skills and abilities, especially in areas where the predictive validity and possible cultural bias of typical cognitive paper and pencil tests is

in question. Shavelson, et. al., (1974) have described how criterion sampling might be applied to the selection of policemen. Their paper illustrates how criterion sampling is really a device for maximizing content validity in areas where the predictive validity of traditional tests is doubtful.

Just how one goes about determining which performance requirements of an occupational role are most important does not appear to have been fully worked out. Would it really be possible to sample randomly in any meaningful sense, or would some kind of conceptual analysis be necessary first? One senses that developers of tests or assessment situations that sample criterion performance would develop some sort of structure much like a traditional test plan, except that the categories would reflect performance in situations that mirror real-life situations rather than academic achievement. But the source of the content is qualitatively different than that of the school achievement test, and criterion sampling deserves to be classified separately.

Objectives-Based

A fourth way of specifying the content of an educational test is commonly described 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. Some advocate the inclusion of a third element in the objective -- an arbitrary definition of what constitutes mastery. This appears to be inappropriate (Skager, 1974). Indicating how many items must be answered correctly to achieve mastery confuses the specification of the content domain with the setting of an interpretive standard or criterion.

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, especially when tied to the computer.

The real question is not whether objectives-based systems represent something new, which they certainly do, but rather how far the behavioral objective can take us in the direction of providing test scores which are amenable to direct, content-based interpretations. Millman (1974) has reminded us that behavioral objectives still 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 (NAEP) listed by Wilson (1974) 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 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) described Popham's attempt to provide practical but reasonably precise guidelines for generating items from objectives. "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 provide the "stimuli" to be used in constructing the items. 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 Millman (1974, p. 34) reproduced 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 and has explored alternative procedures of item analysis for tests developed from objectives.

Formal Item Generation Rules

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 test items representative of a defined content domain. All in one way or another appear to have evolved 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 quite explicitly the goal of quantifying content interpretations.

> "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 [emphasis mine] possible" (p. 15).

The approach that Hively and his co-workers evolved involved an initial process of eliciting statements about curriculum objectives from developers of a mathematics curriculum. These statements utimately 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 when the same set of generation rules can be applied to each), and (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 was certainly recognized by Hively:

> "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" (Hively, et al., 1973, p. 15).

But one must credit Hively and his associates for developing a model which 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 nontrivial 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 for developing tests measuring open-ended performance domains. These questions 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 may not be worth the trouble, although alternate forms of the test can be generated virtually automatically once the form is constructed.

Even an approach to content specification as rigorous as Hively's can apparently still result in items and tests with traditional types of technical defects. The particular item form chosen by Hively (Hively, et al., 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 may be 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). 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 intermediate devices such as behavioral objectives and avoiding "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" (Bormuth, 1970, p. 5).

Bormuth's approach utilizes linguistic principles to derive items from various transformations of 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 the approach Hively and Bormuth take compared to the developers of most objectives-based assessment systems. Hively and Bormuth derive test content directly from instructional materials and statements. Bormuth is militant to the point of belligerency 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 <u>not</u> 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 clear-cut objectives. Tests produced by analysis of actual instructional content might be content valid, but fail in many cases to meet the additional 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 these approaches is 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 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.

Types of Score Interpretations

There are at least three distinctly different ways in which test performance can be interpreted. These involve referencing an examinee's test performance to (a) a content domain, (b) a criterion or standard, and (c) relative position in some defined reference population of persons.

Domain-Referenced Interpretations

"Domain-referenced" interpretations refer to level of performance in a content domain. While such scores are referenced to content rather than to the relative position of examinees, they would differentiate among people along a continuum of competency. These kinds of interpretations have mainly been talked about, but little used up to now. An ideal type of domain-referenced measure is listed in Table 1 as a domain score estimate under the specified content domain column of the table. Derived from performance on a test composed of items sampled from a content domain, the domain score estimate would indicate the number of items an individual would be likely to get correct if he or she could attempt all of the items in the domain. This type of score, which refers to competency with respect to content, carries with it no implication of a criterion or standard.

Other domain score interpretations, not necessarily quantitative, have been proposed. Ebel's (1962) content standard score must be the progenitor of this category of interpretations; the content standard score can be taken as a point estimate of the examinee's competency with respect to a domain. Cronbach's (1970) content reference score refers to the "...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 (though probably arbitrary) decision rules for training or selection. Finally, Ebel's (1962) representative item cluster is not a score, but rather a device for interpreting a score in terms of content. Representative item clusters are groups of items representative of those items typically passed by individuals obtaining a given total score on the test.

Scores that can be interpreted on some absolute scale have always incited a great deal of interest among the technically oriented, but have not yet been of much significant practical use. It does appear that such scores must be assigned to open rather than specified content domains. They also appear to derive from theoretical constructs. Tucker's (1953) proposition IV on the characteristics of an "ideal" test (minimizing the importance of reference groups) suggests this is the case in the use of the word "trait."

> "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 models does not reflect any particular interest on the part of any of these theorists as to how test content is to be specified 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.

Criterion-Referenced Interpretations

It has already been suggested that "criterion-referenced" interpretations, as the concept is currently applied in educational measurement, refer almost exclusively to "sign" rather than score interpretation. Examinees do, or do not, belong to some group of interest. But the traditional meaning of this type of score interpretation refering to predicted performance on a second, independently obtained, criterion measure seems to be just as much criterion-referenced as is a sign interpretation.

Indeed, Glass (1977) in a recent paper reported that this older meaning was really what Glaser had in mind when he coined "criterion-referenced" in 1968.

Glass argued rather convincingly that all criterion-referenced scores proposed to date, no matter how sophisticated mathematically, are based on thoroughly arbitrary decision rules. Tenable rationales for establishing mastery criteria do not yet exist. Glass also suggested that serious problems arise when grand schemes are based on a "fundamental, unsolved problem" necessary for their implementation. The accountability movement in education is a salient example. We attempt to make teachers "accountable" by demanding that they bring all of the learners in their charge up to some arbitrary performance standard which has no defensible educational, psychological, or technical rationale, or to hold learners accountable by insisting that they achieve some equally arbitrary standard of "minimum competency." The negative social fallout from rigorous attempts to enforce teacher accountability may already be evident in a willingness on the part of teachers to assume an adversary role vis a vis management and the public in the attempt to protect what they see as their rights and prerogatives. Other kinds of problems can readily be envisioned if and when parents discover that their children did not receive high school diplomas because they did not surpass arbitrary standards of performance in reading and mathematics. Given sufficient pressure, arbitrary standards imposed by states or school districts will turn out to possess remarkable elasticity. In the meantime, damage may have been done, more adversary relationships created.

We do have to set standards in many situations, arbitrarily or not. Resources allow for only so many new students to be admitted to the University

of California in a given year. A criterion or standard has to be set, even though many individuals who do not meet that criterion could have performed successfully, and many who do surpass the criterion will not perform successfully. In adaptive approaches to instruction, decisions have to be made about whether some type of content has been learned well enough to advance to other, related types of content. This latter use of arbitrary standards seems particularly benign. Mistakes can be made, but their costs are low or virtually zero in a positive learning environment. Thus it is possible to agree with Glass' observation that arbitrary standards are being seriously misrepresented and misused in the attempt to solve serious social problems, but at the same time to believe that even arbitrary standards are useful and even indispensable in certain types of decision situations.

Traditional types of criterion-referenced scores have been expressed as expectancies (probability of falling in each of several arbitrarily determined categories on the criterion measure) or as single point or predictions. In the table these two closely related types of criterion scores are identified with content generation modes yielding domains of the open type. While variability among individuals is necessary for deriving such score interpretations, Cronbach (1970) suggested that the scores themselves refer to actual performance rather than to comparative standing.

Clinical diagnosis is also a type of sign interpretation implying membership in one of two or more groups which differ in some significant way. Diagnosis may be a qualitative, judgemental process in which a precise criterion score is not explicit, but is nevertheless implied conceptually. Diagnostic interpretations generated in order to select the most appropriate mode of instruction

seem likely to be derived from tests assessing theoretical constructs, since interpretation implies a theoretical rationale rather than blind prediction. However, it has been noted that at least one system of adaptive instruction (Project Plan) apparently arrives at diagnostic and placement decisions through prediction from a variety of measures, not necessarily theoretically based.

Mastery interpretations are the object of great interest currently and apparently taken by some to be the only type of criterion-referenced score interpretation. It should be clear by now that this is not the case and that mastery interpretations, while quite arbitray in the light of current theory and technical knowledge, are a special kind of criterion-referenced interpretation identified with content domains of the specified variety. This is not to demean the special importance of the mastery concept with respect to current instructional theory. But the fact that there are limits to the applicability of the mastery concept as well as the arbitrariness involved in determing mastery cutoff scores suggests that reasonable wisdom and caution be exercised in deciding where and when to apply the concept.

Norm-Referenced Interpretations

Scores which are interpreted comparatively in terms of relative standing in some reference group have been with us for some time and hardly need elaboration. Percentiles, standard scores, and age/grade equivalents (no endorsement of the last often misleading interpretations intended) have been with us for a long time. Angoff (1971) adds arbitrary scales for scoring systems tied to a convenient reference group rather than a representative sample. The

Scholastic Aptitude Test, referenced to the population of persons taking the test in 1941, is a familiar example. A number of other types of scores are described in Angoff's exhaustive treatment of this topic, but discussing them would not contribute to the distinctions made here.

Finally, it was suggested early in this paper that norm-referenced interpretations can and probably will be attached to tests developed primarily to assess mastery. Here again 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, does not hold water. 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 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.

For many purposes either norm-referenced or criterion-referenced interpretations provide incomplete information when taken alone. A couple of years ago this situation was illustrated in a newspaper article critical of a report issued by the evaluation branch of a large city school district. It appears that the report demonstrated 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 taken as a sign that the

district was at the very least holding its own, and in many 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, resulting in the district moving upwards in relative standing. The reporter had obviously stumbled on the need for some non-comparative type of interpretation, although formal means for making such interpretations were not available for the tests in question. But answering the "What does a person know?" question does not necessarily answer the "How well is the person doing?" question. Mastery of the objectives might be achieved only because goals are accidentally or even deliberately set too low. Scoring above the 50th percentile could actually conceal the fact that there has nevertheless been an absolute decline over time.

Common Content Specification Mode/Function Combinations

Table 1 helps make it apparent that certain content specification modes are more likely to be used than others, and that for each combination of mode and function some types of interpretations are more likely to be more useful than others. Looking down the C/P Matrix column, it appears that this rather flexible specification mode can be used for all functions except the diagnostic/progress function. One might question its use for placement in the curriculum, but it has been noted that such use is made (via prediction) in one system of adaptive instruction. Likewise, in not being susceptible to mastery interpretation this specification mode might not seem adaptable to the certification/crediting

function. Yet, advanced placement examinations in use for some time probably fit best under this specification mode.

Theoretical constructs as content specification modes appear to be primarily useful for the formative function of diagnosis and the summative function of selection/prediction. This does not mean that test content specified by constructs could not be used for certification/crediting, although this does not appear to be a common application. The criterion sample mode, connected as it is with occupational competencies, seems most applicable to the two summative functions.

Modes for generating specified content domains appear to be particularly appropriate for the two formative functions of placement and diagnostic/progress and for the summative function of certification/crediting. Domain score estimates are scores rather than signs and as a result yield more information. They would be useful for research purposes and for calculating domain score norms. The applications of mastery interpretations in adaptive programs of instruction should be self-evident.

CONCLUSION

This paper was not written to downgrade the importance of developments that have occured over the past decade. Interpretation of achievement measures in terms of content has been seriously neglected, especially in relation to the instructional process. The interest that has been stimulated for finding new ways to specify test content is laudable and progressive, as is on-going work on the development of new content-based means of interpreting test scores.

Above all, this paper makes the point that not one type of test is solely criterion-referenced as compared to another type of test that is solely norm-referenced. The really important differentiating factors have to do with the function for which the test is to be used and the mode by which the test content is to be specified. Once this distinction is made, criterion- vs. norm-referencing becomes a matter of the type of score interpretation which is likely to be most useful. In many situations both types of interpretations are likely to be useful, whatever the mode of content specification.

The mastery interpretation, so critical to adaptive instructional systems and procedures, is a criterion-referenced interpretation that can be attached to special content specification modes. Development of relevant theory and technology in this area is certainly one of the most significant areas of contemporary work in psychometrics. But at the same time, the applicability of mastery interpretations is not universal. There are important educational domains that can be described as "open" which are not susceptible to mastery interpretation and which are associated with their own set of content specification modes.

REFERENCES

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.

Block, J. H. (Ed.), Operating principles for mastery learning. New York: Holt, Rinehard & Winston, 1971.

Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 2(1).

Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill, 1971.

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.

Buros, O. K. Fifty years in testing: Some reminiscences, criticisms, and suggestions. Educational Researcher, 1977, 6(7), 9-15.

Cattell, R. M. Psychological measurement; Normative, positive, interactive. Psychological Review, 1944, 51, 292-303.

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

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

Cronbach, L. J. Comments on mastery learning and its implications for curriculum development. In Eisner, E. W. (Ed.), Confronting curriculum reform. Boston: Little, Brown, 1971, 49-55. (a)

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

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

Cronbach, L. J., & Snow, R. E. Aptitudes and instructional methods. New York: Irvington Publishers, 1977.

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

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

Eisner, E. Instructional and expressive educational objectives: Their formulation and use in curriculum. In Popham, W. J., et. al., Instructional Objectives. Series on Curriculum Evaluation, American Educational Research Association, No.3, 1-18.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike, (Ed.), Educational Measurement. (2nd ed.) Washington: American Council on Education, 1971.

Glass, G. V. Standards and criteria. Occasional Paper No. 10, Evaluation Center, College of Education, Western Michigan University, Kalamazoo, Michigan, 1977.

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, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.

Hively, W., Maxwell, G. R., Sension, D., & Lundin, S. Domain referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST PROJECT. CSE Monograph Series in Evaluation, No. 2. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1973.

McClellan, D. C. Testing for competence rather than for intelligence. American Psychologist, 1973, 28, 1-14.

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

Shavelson, R. J., Beckum, L. C., & Brown, B. A criterion sampling approach to selecting patrolmen. Police Chief, 1974 (Sept.), 55-61.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45 (1), 127-147.

Skager, R. W. Lifeling education and evaluation practice Oxford: Pergamon Press, 1978.

Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assessembly, and interpretation. In Harris, C. W, Alkin, M. C., & Popham W. J. (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. 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. [ED 054 232]

Tucker, L. R. Scales minimizing the importance of reference groups. In Proceedings of the 1952 invitational conference on testing problems. Princeton, N. J.: Educational Testing Service, 1953, p. 22-28.

Wilson, H. A. A judgmental approach to criterion-referenced testing. In Harris, C. W., Alkin, M. C., & Popham W. J. (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, Los Angeles, 1974.