

# Issues of Validity for Criterion-Referenced Measures

Robert L. Linn

University of Illinois, Urbana-Champaign

It has sometimes been assumed that validity of criterion-referenced tests is guaranteed by the definition of the domain and the process used to generate items. These are important considerations for content validity. It is argued that the proper focus for content validity is on the items of a test rather than on examinee responses to those items. Content validity is important for criterion-referenced measures, but it is not sufficient. This claim is dis-

cussed and the case is made that interpretations and uses of criterion-referenced tests require support of other kinds of evidence and logical analysis. The inferences that are made should dictate the kinds of evidence and logical arguments that are needed to support claims of validity. Illustrations of aspects of the validation process are provided in two concrete examples.

Validity is widely acknowledged as the touchstone of educational and psychological measurement. Yet, within the context of criterion-referenced measurement, validity has received relatively little attention. Certainly it has not been the focus of as much attention as topics such as reliability, the determination of test length, and questions about the need for score variability. This apparent imbalance may be partially attributed to the fact that criterion-referenced measurement posed some new problems and/or alternate ways of formulating the issues in those areas that have attracted the most attention. The relative lack of attention to questions of validity may also be attributed to perceptions about what validation of criterion-referenced measures entails and about the inherent strengths of such measures.

One of the important contributions of the criterion-referenced testing movement has been an increased emphasis on content. The absolute interpretations of the measures are dependent upon clear specifications of the content domain and on the degree to which the measure is representative of the domain. These are, of course, the key components of content validity—ones that have often been espoused in other contexts but seldom taken as seriously as they are by proponents of criterion-referenced measurement. Thus, the content validity of a criterion-referenced measure may often seem less debatable than that of a test developed using more traditional methods of content specification and item selection. Furthermore, content validity commonly has been held to be the only, or at least the most important, type of validity that is needed for criterion-referenced measures.

---

*APPLIED PSYCHOLOGICAL MEASUREMENT*

*Vol. 4, No. 4 Fall 1980 pp. 547-561*

© Copyright 1981 West Publishing Co.

Recently, several authors (e.g. Hambleton, 1980; Linn, 1979; Messick, 1975) have pointed to the need for a broader conceptualization of validity for criterion-referenced measures. It is content validity, however, that remains the central focus in many discussions of the validity of criterion-referenced tests. Thus, a discussion of content validity seems a natural starting point for a paper such as this, which is concerned with issues of validity for criterion-referenced tests.

Following the discussion of content validity, the case will be made that despite its importance for criterion-referenced measures, content validity is seldom, if ever, sufficient. Common uses and interpretations of criterion-referenced measures require other kinds of evidence for support. Although these other kinds of evidence have traditionally been categorized under the headings of criterion-related and construct validity, it will be argued that there are advantages to a unified conceptualization of validity that starts with the inferences that are drawn and the uses that are made of scores on criterion-referenced tests. These uses and inferences dictate the kinds of evidence and logical arguments that are required to support judgments regarding validity. The section on inferences from criterion-referenced measures attempts to illustrate the way that inferences dictate the kinds of evidence that must be marshalled to support claims of validity. The final two sections of the paper provide concrete examples of parts of the validation process. The first example involves items for a relatively specific objective, while the second example illustrates the validation of a criterion-referenced measure covering a broad and complex domain.

### Content Validity

Content validity is one of the three kinds of validity that are formally recognized in the *Standards for Educational and Psychological Tests* (APA, 1974). However, few would consider content validity to stand on an equal footing with the other two types of validity in terms of the rigor of the evidence that is usually provided to support a claim of validity. Indeed, some well-known test theorists have argued that what traditionally goes under the heading of content validity should not even be called validity. For example, Messick (1975) said "call it 'content relevance,' . . . or 'content representativeness,' but don't call it content validity" (pp. 960-961).

Others besides Messick have expressed reservations about content validity. For example, Guion (1977) provided a number of reasons for his reservations regarding content validity. Primary among these reasons is his conclusion that "judgments of content validity have been too swiftly, glibly, and easily reached in accepting tests that otherwise would never be deemed acceptable" (1977, p. 8). Despite his reservations, Guion argued that the ideas contained under the notion of content validity are extremely important.

For the acceptance of a measure on the basis of content validity, Guion proposed a set of five minimal conditions. These conditions are

1. The content domain must involve "behavior with a generally accepted meaning" (p. 6);
2. The definition of the domain must be unambiguous;
3. The domain must be relevant to the purposes of the measurement;
4. "Qualified judges must agree that the domain has been adequately sampled" (p. 7); and
5. The measure must have reliability.

Although Guion's list is useful, it includes considerations other than those of content validity. Concerns for "meaning" and "relevance" involve inferences that go beyond those that are justified on the basis of content validity. They involve constructs or external criteria and require other types of evi-

dence. It will be argued that content validity is more limited and is derived from only two considerations: domain definition and representativeness. It also will be argued that the focus of content validity should be on the items rather than on the behavior of the examinees.

Although there is widespread agreement that content validity depends on domain definition and representativeness, there are (as recently documented by Fitzpatrick, 1980), rather sharp discrepancies in the positions that are found in the literature regarding "(1) what features of a measure are evaluated under the rubric of content validity, (2) how content validity is established, and (3) what information is gained from study of this validity" (Fitzpatrick, 1980, p. 2). A particularly important distinction that is highlighted in Fitzpatrick's review is the one between samples of items and samples of responses. Both have been proposed as the focus of concern for content validity, but they are clearly not equivalent.

Fitzpatrick (1980) has argued convincingly that content validity should "be thought of as referring to the outcome of judgments about the sampling adequacy of test content and that construct validity be considered the validity relevant to issues about the meaning of samples of test responses" (p. 19). Thus, for content validity, the concern is with the items rather than with the responses.

This position is consistent with that of Hambleton (1980), who recently argued that

Criterion-referenced test scores are used frequently to make descriptions of examinee levels of performance in content areas measured by a test and to make decisions. However, it is essential to establish the validity of the descriptions and decisions. Content validity is not sufficient since it pertains to the content of the test, whereas descriptions are made based on examinee *responses* [italics in original] to the test items. (p. 94)

Judgments about sampling adequacy or representativeness require clarity of definition of the item domain. Indeed, domain definition provides the key to item generation and content validity.

Item generation starts with the definition of a content domain. Ideally, the definition should be complete in the sense that all potential items of the domain could be enumerated, at least implicitly, as in the case of a set of item generation rules. This first step is both the most crucial and the most difficult. Indeed, some would argue that it is a hopeless task for other than trivial content domains.

There are a few examples of relatively complete domain specifications, such as the work of Hively and his associates (Hively, Maxwell, Rabehl, Sension, & Lundin, 1973) on the creation of item forms and item generation rules. There are also examples of computer-generated spelling items (e.g., Fremer & Anastasio, 1969). Tests of literal comprehension have been produced using randomly selected paragraphs with the cloze procedure, and Millman (1980) has provided an example of computer-generated test items of higher level cognitive skills using a special computer language designed for the purpose of item generation (Millman & Outlaw, 1978). Examples of completely specified domains remain rather scarce, however. Furthermore, they tend to be for rather narrowly defined skills. Arithmetic computation is still the primary example.

Bormuth's (1970) suggestions for developing a systematic approach to generating items through transformation rules is one of the most comprehensive attempts at stating a content theory on which achievement testing could be based. As demonstrated by Diederich (1970), however, some items generated following Bormuth's rules may be rather meaningless. Even where meaningful, the intuitive importance of the questions often appears rather trivial when compared to the ones created using the judgments of experienced item writers.

Many of the meaningless and seemingly trivial questions that were illustrated by Diederich's examples could be avoided using refinements of Bormuth's approach, such as the one developed by

Finn (1975). In Finn's approach, word frequency counts are used to identify target words. Sentences in which the target words appear are then transformed into questions. Although not foolproof, this approach increases the likelihood that "high information" sentences will be selected for the creation of questions. Such modifications enhance the utility of Bormuth's approach but may still exclude many questions that most judges would rate as better for purposes of assessing understanding than the questions that are automatically generated.

Without an exhaustive listing of items or satisfactory item generation rules, tests cannot be constructed automatically by random sampling. The reaction to an inability to achieve complete specification is usually to fall back on the familiar table of content specifications to guide a committee of content specialists and test development personnel in the artistic creation and selection of items. As Popham (1978) has suggested, however, there are intermediate possibilities between the traditional table of content specifications and the complete specification of an item domain. Guttman's facet theory (see Berk, 1978, for a recent discussion of potential applications) and Popham's amplified objectives (see Millman, 1974; Popham, 1978, 1980) are examples of intermediate positions that provide much more specificity than is found in a typical table of content specifications.

Even if a table of content specifications is the best that can be accomplished, much could be done to learn how well the specifications work and how they might be improved. Little is known about how well such tables work or if some sets of specifications are better than others. Cronbach's (1971) "duplicate-experiment" suggestion is one potentially useful way of evaluating the use of content specifications in test construction. According to his suggestion, two independent teams would use the content specifications to construct separate versions of the desired test. The adequacy of the specifications could then be judged by the degree of equivalence between the two forms.

The systematic collection and analysis of judgments of content specialists can also be used to evaluate the content validity of a measure. Hambleton (1980) has reviewed several techniques for collecting and analyzing judgments of the "match between the test items and the domains they are designed to measure" (p. 87). These techniques are of greatest potential value when they are used to refine the definition of the domain specification and item generation rules rather than merely selecting and discarding items (Rovinelli & Hambleton, 1977; Wardrop, Anderson, Hively, Anderson, Hastings, & Muller, 1978).

### The Need for More Than Content Validity

Content validity is certainly important for a criterion-referenced measure. Questions of validity cannot, however, be limited to those that are traditionally categorized under the heading of content validity (Hambleton, 1980; Linn, 1979; Messick, 1975).

If a domain of items is exhaustively defined such that tests can be constructed using random sampling procedures, then probability theory provides a natural basis for supporting certain inferences from the test scores. This approach is often seen as a way of insuring content validity, which may be the only type of validity considered important for a criterion-referenced test. However, one must be willing to be restricted to simple descriptive statements (e.g., the person answered  $x$  percent of the items correctly, from which it is estimated that the person would answer correctly between  $y$  and  $z$  percent of items in the domain if all items could be administered under identical conditions). Note that nothing is said about the proportion of the items that the person *knows* or is *able* to answer correctly. Nor is any use of the test scores, such as the differential assignment of instructional materials, supported by the description. It should also be noted that much is covered by the phrase "if all items could be administered under identical conditions."

Many interpretations that are commonly made of criterion-referenced tests involve implicit predictions. Many of the interpretations also involve constructs that are invoked to account for performance on the test. A prediction about future performance is implicit when scores on a criterion-referenced test are used to decide which students should be given remedial instruction. A construct is implicit when it is concluded that students with low scores on a criterion-referenced spelling test are incompetent or that they lack the ability to spell an adequate proportion of words in the domain of interest. Statements regarding inability or incompetence require inferences, and “the very use of the term *inability* invokes constructs of attributes and process, whereas a content-valid interpretation would be limited to the outcomes” (Messick, 1975, p. 960).

Interpretations of scores that go beyond simple empirical summaries (e.g., 75% of the examinee’s answers were correct) to even the simplest level of inference (e.g., the examinee *can* answer correctly 75% of the items in the domain) need to be supported by evidence and logical argument. In traditional terms, evidence for criterion-related, as well as construct, validity will usually be needed in addition to evidence of content validity. The latter is necessary but rarely, if ever, sufficient to support the interpretations and uses of criterion-referenced test scores. As noted by Hambleton, Swaminathan, Algina, and Coulson (1978),

Content validity is a test characteristic. It will not vary across different groups of examinees. . . . However, the validity of test score interpretations *will* [italics in original] vary from one situation to another. For example, if a criterion-referenced test is administered, by mistake under highly speeded test conditions, the validity of interpretations based on test scores obtained from the test administration will be lower than if the test had been administered with more suitable time limits. (pp. 38–39)

Evidence that is associated with each of the three traditional types of validity (content, criterion-related, and construct) will usually be needed to support the interpretations and uses of a criterion-referenced test. The mix of kinds of evidence that are needed will vary as a function of the interpretations and uses that are made of the scores. The kinds of evidence that are most needed should be determined by the interpretations and uses and by the plausible alternatives.

### A Unified Conception of Validity

It is common to speak of the validity of a test as if it were a singular entity and an inherent property of the instrument, but it has long been recognized that a test may have many validities associated with it. As described by the *Standards for Educational and Psychological Tests* (APA, 1974), “Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores” (p. 25). There are obviously many different inferences that can be based on a test score, not all of which are equally valid. Indeed, some inferences based on a given test score may be quite invalid, whereas others are of relatively high validity.

Different inferences, interpretations, and predictions based on test scores may differ not only in their degree of validity but also may require different kinds of evidence. Hence, it is both common and natural to talk in terms of types of validity. Messick (1979) has made an important distinction between the need for different *kinds of evidence*, by which he means “both data or facts and the rationale or argument that cement those facts into a justification of test score inferences” (p. 6) and *types of validity*. As noted by Messick, “. . . there are a variety of validation methods available but they all entail in principle a clear designation of what is to be inferred from the scores and the pre-

sentation of data [and logical argument] to support such inferences” (p. 5). Compartmentalization of thinking about validity into the traditional types of validity tends to perpetuate the notion that there are alternative approaches or “roads to psychometric salvation” (Guion, in press, p. 4) and that one must merely pick one of them.

Messick is not the only one to have objected to the separation of types of validity. Rather, there is a growing consensus that a unified conceptualization of validity is needed (see, for example, Cronbach, 1980; Dunnette & Borman, 1979; Guion, 1978, 1980; Messick, 1979; Tenopyr, 1977). So-called types of validity should be viewed as approaches to accumulating certain kinds of evidence rather than as alternative approaches, any one of which will do. The *Standards for Educational and Psychological Tests (Standards; APA, 1974)* warned against the treatment of the three types of validity discussed in that document as alternatives. After the three types of validity—content, criterion-related, and construct—are introduced, the *Standards* go on to state,

These aspects of validity can be discussed independently, but only for convenience. They are interrelated operationally and logically; only rarely is one of them alone important in a particular situation. A thorough study of a test may often involve information about all types of validity. (p. 26)

The argument that the common types of validity should not be separated conceptually or operationally is, unfortunately, often ignored. As noted by Dunnette and Borman (1979), even the *Standards* tend to reinforce the compartmentalization in the discussion following the introduction in which each type of validity is taken up separately without much attention to an integration of the approaches. This approach to the topic has also been typical of the Division 14 Guidelines (APA, 1975) and many textbook and journal discussions of validity, including a recent one of mine (Linn, 1979). That is, the “types of validity” are discussed in separate sections, and insufficient attention is given to integration of the various sources of evidence for particular interpretations of test scores. The separation is reasonable if viewed merely as a way of providing focus on approaches that have been useful in accumulating certain “kinds of evidence” for particular interpretations. The danger of the separation is that it may contribute to thinking of the types of validity as alternatives.

The view that content, criterion-related, and construct validity are alternatives, albeit not equally desirable ones, is most strongly reinforced by the “Uniform Guidelines on Employee Selection Procedures” (Equal Employment Opportunity Commission, 1978). The stress on types of validity reinforces narrowness in conceptualizations about the kinds of evidence that is needed or desirable. As noted by Dunnette and Borman (1979), “the implication that validities come in different types leads to confusion and in the face of confusion, oversimplification” (p. 483).

In order to avoid the type of oversimplification caused by thinking of kinds of validity as alternatives, it is desirable to start with a focus on interpretations and uses that are made of test scores and on the inferences that are based on the scores rather than on the kind of validity that is needed. A variety of kinds of evidence and logical arguments will usually be involved when a particular interpretation is subjected to close scrutiny and evaluated against alternative interpretations. Hambleton (1980) provides a review of some of the many potentially relevant methods that may be useful in obtaining evidence of validity (see also Cronbach, 1971).

### **Inferences from Criterion-Referenced Measures**

Attempts are sometimes made to carefully delimit the nature of the inferences to be made from criterion-referenced tests. For example, it may be argued that only “low level inferences” (i.e., those

that are closely tied to the definition of an item domain and rely upon well-established principles of sampling theory) are appropriate for criterion-referenced measures. Harris, Pearlman, and Wilcox (1977) provide a clear exposition of this position, although they do not use the criterion-referenced measurement label to designate the type of achievement testing that they discuss.

Harris et al. (1977) require that "a universe of items exists either implicitly or actually in such a form that it will be feasible to draw random or stratified random samples from this universe" (p. 2). The work of Hively et al. (1973) is cited as an example that satisfies this requirement. Given an achievement test satisfying their requirement, Harris et al. (1977), would limit inferences to generalizations "based on sampling principles." In their words, "An achievement test constructed on this item sampling principle yields a proportion correct score which is an unbiased (and maximum likelihood) point estimate of the proportion of the items in the domain which the student can handle adequately" (p. 3). The validity of the inference about the proportion of items in the domain that the student can handle adequately is presumed to rely only on well-established sampling theory and, of course, on the use of random or stratified random-sampling procedures to select items.

The inference that Harris et al. (1977) wish to make from an achievement test score is very limited and, consequently, the validation requirements are clearly defined and straightforward. The major hurdle in the process comes at the point of defining the universe in such a way that the sampling requirements can be satisfied. It should be recognized, however, that the inference that can be legitimately supported in this fashion is actually much more restricted than is connoted by the phrase "can handle adequately."

Low proportion-correct scores on a test may be seriously biased estimates of the proportion of items in the domain that students *can* handle, not because of any violations of sampling principles, but because the students were not motivated to try very hard on the test. Thus, it is important to emphasize that inferences about what students can handle requires more than the definition of the universe of items and the use of appropriate sampling procedures. The inferences must be restricted to particular situations and conditions of administration. This needed restriction was recognized by Harris et al. (1977) who argue that the test specifications should include a definition of the "types of situations in which the behavior is expected to be elicited" (p. 2). In other words, the estimate of the proportion of items in the universe that students can handle adequately is situation specific. If the situation does not adequately motivate students for the sample of items on the test, then the inference is about domain performance under conditions of limited motivation. If there is time pressure or if there are other factors that reduce the performance of high anxious examinees on the sample of items, then inferences about what they can do apply only to those anxiety-inducing conditions. As can be seen, even the limited type of inference envisioned by Harris et al. (1977) is more complicated than it may at first appear, and validation of the inference requires more than item generation rules and sampling theory. Furthermore, rarely, if ever, is it either feasible or desirable to limit inferences that are made from test scores so severely.

The restricted inference that a low test score implies—that a student cannot adequately handle items of that type, given the situation and conditions of administration—is of limited utility. Greater utility is provided by inferences about the students' current capabilities that rule out alternative explanations, such as poor motivation. Inferences about future performances on items in the domain, as well as about behavior in other situations, also contribute to the utility of the test scores. Actions that are taken on the basis of test scores depend upon inferences, albeit often implicit ones. For example, decisions to assign students with relatively high scores to instructional materials covering a new content area while assigning those with low scores additional, possibly remedial, materials in the content area corresponding to the domain of the test rest upon inferences about the future behavior of the stu-

dents. Those with high scores are judged to be ready to tackle new materials; and those with low scores are judged to need additional work first, that is, that they will be better off if they receive additional work in the area before moving on than they would be if allowed to proceed at that point.

Such inferences about future behavior and the differential effects of different experiences as a function of level of performance on the test need supporting evidence of validity. By themselves, domain definition and use of random sampling do not provide sufficient evidence to support the validity of the use of test scores to determine the assignment of instructional material. Results from an experiment with students randomly assigned instructional materials are apt to be more compelling but expensive, and sometimes infeasible to obtain. At a minimum, a logical analysis, not only of the item domain, but of the contents of the instructional materials and the relationship between the item domain and instructional materials is needed. Hambleton (1980) has discussed the other types of evidence that are needed under the heading of decision validity, which he views as "a particular kind of construct validity" (p. 98). As is clear from his discussion, the type of evidence that is needed to justify the use of a criterion-referenced test to make instructional decisions involves more than domain definition and random sampling of items.

Even if test scores are used only to describe the proportion of items that a student can handle adequately and if no actions are based on the scores, inferences will usually be of a more general nature than can be strictly supported by appeals to definition of the domain, the testing situation, and the procedure used to sample items. Consider, for example, the clearly defined domain of the division of two-digit integers by a single-digit integer ranging from 2 through 9. All possible items in the domain can be listed, and random or stratified-random samples of items can be selected for tests. The format used to present the items and the administration conditions can be specified. When proportion-correct scores are discussed, however, the interest will be not only in the proportion of problems presented in a particular format that the student can answer correctly but in the student's ability to divide two-digit integers by single-digit integers. As recently shown by Alderman, Swinton, and Braswell (1979), however, seemingly minor variations in item format can substantially influence performance. Thus,  $42 \div 7$ , 42 divided by 7,  $7/\sqrt{42}$ ,  $42/7$ , and  $\frac{42}{7}$  represent alternative formats that might be used to define the domain. Based on the results reported by Alderman et al. (1979), the proportion-correct score would be expected to differ from one domain to the next. Furthermore, the relative performance of different groups of students on the different domains is apt to depend upon the match between the format of the item domain and the problem format used in instruction.

Differential effects due to item format and the match between the format used in instruction and on the test are quite consistent with the formulation presented by Harris et al. (1977). It merely emphasizes the fact that item format may be a critical component in the definition of a universe of items. In general, however, inferences about a student's ability to divide are much more interesting than inferences about a student's ability to solve division problems when they are presented in a particular format. If a student fails to answer correctly items in one format but could answer them correctly if they were presented in another format, then the inference that the student cannot divide is invalid. Only a statement about items in the format used can be justified.

### **Objective, Domain Definition, and Score Interpretation: An Example**

Construction of a criterion-referenced test typically starts with some stated educational objective. The objective is not sufficient to generate test items, however. This was illustrated above by the different formats for division problems, any one of which is consistent with the highly specific objective that students can divide a two-digit dividend by a one-digit divisor. Another example is provided by the re-



sults of the validation studies conducted as part of the Beginning Teacher Evaluation Study (BTES; Filby & Dishaw, 1977).

Filby and Dishaw (1977) used individual testing as part of an effort to validate the group-administered achievement tests used in the BTES. In addition to differences in administration conditions, the group and individual tests also differed in the nature of the task. For example, constructed responses might be used in the individual administration in place of the multiple-choice format used in the group-administered tests. Both types of test items were intended to be appropriate for a common objective.

It is hardly surprising that differences in the proportion correct would be found between results for multiple-choice items in a group-administered test and those for constructed-response counterparts on an individually administered test. Psychological studies of memory long ago distinguished between recall and recognition. In addition, the multiple-choice format provides the opportunity for a student to guess the correct answer. Nonetheless, the comparisons made by Filby and Dishaw (1977) were quite informative and provided a good basis for judging the validity of particular interpretations of scores on subsets of test items.

A small example from the Filby and Dishaw (1977) results may serve to illustrate the gap that often exists between objectives and test items, and the need to investigate the validity of interpretations of scores. A sample objective that was presumed to be measured by some of the multiple-choice test items was that students should be able to identify common geometric figures (e.g., square, rectangle, circle, triangle). The proportion correct for a sample of fifth-grade students was only .73 for the multiple-choice item that asked students to identify a square and .67 on one that asked students to identify a rectangle. Not only are these proportions smaller than expected for fifth-grade students, they are also smaller than the proportion of students who were able to satisfactorily draw the figures when named by an interviewer in the individual testing. The proportion correct in the individual production tasks were .93 for the square and .87 for the rectangle.

As expected, almost all students know the two simple figures, as evidenced by their ability to draw them when asked to do so. The poorer performance on the multiple-choice questions was attributed to characteristics of the items that require knowledge other than that in the objective. The multiple-choice item for a rectangle is shown in Figure 1. As can be seen, the question requires a student to identify a rectangle imbedded in a larger figure. It also requires the use of geometric notation. Based on the results of student interviews, Filby and Dishaw (1977) concluded,

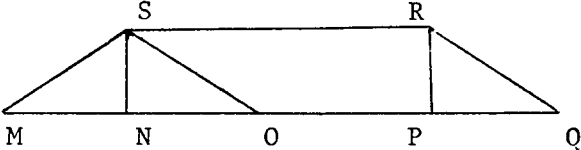
Almost all students know these two basic figures. But students are relatively unfamiliar with geometric notation, as became apparent during individual administration of group-test items. In approaching the test item with the imbedded rectangle some of the children viewed each part of the figure as separate and were confused by the extra lines. Within the rectangle SPRN there is the triangle SON. Since the letter O was within the rectangle some students thought that it must be accounted for in their answer. One student thought that the answer choices were words and that the object of the question was to find the word that was in the picture! Another student had difficulty since he did not realize that the starting point is also assumed to be the ending point. Therefore in tracing SRNP his figure was incomplete . . . . As time progressed the interviewer started instructing the children to trace the figure prior to answering the question and this of course biased the results. None of the students who were asked to find the rectangle, trace over it, and then determine which letters named it had any problem with the question. Also the order in which the letters were read did not appear to present any difficulties. Those students who were initially incorrect were able to solve the problem when asked to follow the

above procedure. It therefore appears that items that emphasize geometric notation introduce a test-taking factor that is separate from knowledge of geometric figures and is probably inappropriate for fifth graders. (p. 54)

**Figure 1**  
Example of an Item from the BTES

---

Which of the following names a rectangle?



A. SRPO  
B. MSO  
C. SRPN  
D. SRQO

---

It might quite reasonably be argued that one should never have considered interpreting poor performance on items such as the one in Figure 1 to imply that a substantial fraction of fifth-grade students are not able to identify rectangles. Adherence to the principles espoused by Harris et al. (1977) certainly would have avoided such an invalid interpretation. One might imagine the statement of a set of rules for constructing embedded figures that would include the item in Figure 1 as one possible instance. Alternatively, a universe of such items might be defined by a listing of acceptable items. A random sample of items could be drawn from the universe and the proportion-correct scores used as estimates of "the proportion of items in the domain which the student can handle adequately" (Harris et al., 1977, p. 3). The latter type of inference may be valid, whereas the inference regarding the objective that students should be able to identify common geometric figures was not.

Validity is not solely a property of the item or set of items: It depends on the inference. The lack of validity for purposes of assessing attainment of the figure recognition objective is due to a lack of "item-objective congruence" (Berk, 1980). In this example, achieving the objective is a necessary condition for getting a large proportion of the items correct, but it is not a sufficient condition.

### **An Example of a Broad Gauged Criterion-Referenced Test**

The push toward complete definitions of domains that allow the use of sampling procedures to construct tests often results not only in improved specificity but in very narrowly defined domains. A score on a typical norm-referenced achievement test and a score for a specific skill on a criterion-referenced test differ in a number of ways. One of the most immediately obvious differences, however, often is in the breadth of the tests. Breadth of coverage of a criterion-referenced test need not be narrow, however. As suggested by Hambleton (in press), the breadth of the domain for a criterion-referenced test should depend on the purpose of the test. Domains can vary greatly in breadth and complexity. The essential feature for a criterion-referenced test is clarity of definition rather than breadth or complexity (Hambleton, in press). Nonetheless, highly specific but rather narrow domains, dis-

cussed above, are common for criterion-referenced tests. For example, the Basic Arithmetic Skill Evaluation (BASE; May & Hood, 1973-1974) reports scores for skills such as

Given any 4, 5, or 6 digit numeral the pupil can round off the numbers to the nearest hundred and

Given 2 unlike fractions with denominators of 2, 3, 4, 5, 6, 7, 8, 9, 10, or 12 the pupil can find their sum or their difference

(from Student Profile of Arithmetic Skills, *Basic Arithmetic Skill Evaluation*, May & Hood, 1973-1974)

The high degree of specificity for skills such as those listed for the BASE tests is very desirable for some purposes. They are linked much more clearly to particular instructional activities than is possible with a global score for arithmetic computation. For other purposes, however, the narrowness of the skills tested and the proliferation of scores limits their utility. Koslin, Koslin, and Zeno (1979) have argued that

Criterion-referenced tests are typically (although they do not necessarily have to be) limited in scope to the mastery of specific units of instruction. While the tests may adequately measure whether students have or have not mastered some particular instructional objective (such as learning to blend initial diphthongs), they are *not* [italics in original] good measures of complex long-term educational outcomes, such as having learned to read well enough to function as a competent adult. (p. 313)

Koslin et al. (1979) go on to suggest that an alternative type of measure is needed for purposes of measuring long-term educational outcomes. The alternative that they suggest is called an "effectiveness measure," and they clearly articulate the requirements for such measures in their description of the development and validation of the Degrees of Reading Power (DRP) test. According to Koslin et al. (1979), an effectiveness measure is "a measure of outcome which provides a directly interpretable assessment of the extent to which an enduring adult ability such as reading with comprehension has been attained" (p. 311).

An ability such as "reading with comprehension" is certainly much broader than the specific skills that most criterion-referenced tests are intended to measure. However, the DRP satisfies the primary requirements of a criterion-referenced test. Indeed, it is in many ways an outstanding example of a test that is designed to measure a broad and important ability over a wide range but that has a well-defined domain and results in scores that are directly interpretable in terms of the difficulty of the prose that an examinee can read with comprehension. As noted by Koslin et al. (1979), the fact that criterion-referenced tests are typically fairly narrow in scope does not imply that they must necessarily be so limited. On the contrary, the DRP seems to provide an outstanding example of such a test with broad scope that may reasonably be labeled a criterion-referenced test.

Whether the DRP is best categorized as an effectiveness measure or as a criterion-referenced test is not particularly important, at least not for purposes of this paper. Regardless of the classification, the DRP provides an example of a validation that is quite relevant to a discussion of validity of criterion-referenced test scores.

The DRP is similar to a cloze test in that words are deleted from a prose passage and the examinee has to select the deleted word. It differs from the typical cloze passage, however, in the choice of words for deletion and in the number of words deleted. Fewer words are deleted on the DRP (e.g., sev-

en deletions for a passage of several paragraphs) than is found on the typical cloze test. Also, only familiar words (ones that occur with high frequency in prose) are deleted or used for distractors. The purpose of limiting deletions to highly familiar words is to minimize the dependence of the DRP scores on the particular vocabulary of the response options. A third, and probably the most important feature of the words selected for deletion, is "that processing surrounding prose is both necessary and sufficient to choosing the right answer" (Koslin et al., 1979, p. 316).

The scaling of the DRP is accomplished by scaling the difficulty of the passages. That is, passages are initially indexed by a readability formula, and difficulty ordering of the passages is then verified after the test is administered.

Koslin et al. (1979), identify three propositions underlying their work, which, if satisfied, result in a measure of the difficulty of prose that can be used to "specify the ability that someone will need in order to process it" (1979, p. 317). Conversely, the likelihood that a person can read particular prose selections with comprehension can be estimated from the person's ability score and knowledge of the difficulty of the test. These are strong claims that require supporting evidence. That is, they must be validated.

Validation of the DRP has involved a wide array of activities, including experimental and correlational studies, accumulation and interpretation of evidence from the literature, and theoretical and logical arguments that piece the various sources of evidence together. For example, a requirement that answers to items should depend on the intersentential context was investigated experimentally by using the identical sets of response options with the same deleted words in passages. The sentences containing the deleted words were also the same, but the surrounding sentences differed in readability. Differences in performance on the difficult and easy versions of the test were interpreted as support for the claim of dependence on the intersentential context. In contrast, however, the requirement that comprehending the passage is a sufficient condition for answering questions correctly is supported primarily by logical argument.

Several kinds of evidence were obtained to validate the proposition that texts can be arrayed in difficulty and examinees arrayed in ability to comprehend prose on a common scale. Surface linguistic variables such as average word length and average sentence length were used to predict mean and median Rasch (1960) difficulties for passages. Bormuth's (1969) mean cloze forecasts were also used to predict passage difficulties (mean and median Rasch difficulties). The obtained correlations were all quite high (e.g.,  $r = -.98$  between Bormuth's readability measure and Rasch difficulty on the DRP).

The validity of the DRP for predicting the probability that students can successfully read particular prose passages was investigated experimentally. The study demonstrated that "giving students the intact text to read just before they took a comprehension (hard cloze) test led to 'mastery' level performance *if and only if* the readability of the text was in range relative to subjects ability as predicted by the DRP (i.e., if the probability of success equalled  $p = .75$ )" (Koslin et al., 1979, p. 327).

As is true of any test, the validation of the DRP is incomplete. Whenever an alternative interpretation or use of the scores is introduced, there will also be additional validation needs. But the DRP does provide an unusually good example of a systematic approach to validation. Propositions underlying the approach are explicated and evidence is presented to support the reasonableness of those propositions. Inferences to be made from the test scores are identified and tested against plausible alternatives. The process is not narrowly confined to a single type of validation or to a particular kind of evidence. Rather, the process typically starts with a definite statement of the proposed interpretation and a contrasting alternative interpretation. Evidence is then sought that will distinguish between the competing interpretations.

Cronbach (1971) noted in his discussion of validity that "the procedures cannot be cataloged exhaustively and no guide can tell just how to meet the requirement of hard-headed reasoning from data to conclusions" (p. 483). The activity is purposeful, however, and proceeds best when proposed interpretations are clearly stated and contrasted with competing interpretations. Although Cronbach's conclusion was made within the context of a discussion of construct validity and without specific reference to criterion-referenced tests, it applies to such tests and to the validation of interpretations and uses of tests that do not invoke constructs.

As stated by Messick (1979), "Validity is the overall degree of justification of test interpretation and use" (p. 7). Many methods may be used to accumulate the evidence required to justify the interpretation and use, "but they all entail in principle a clear designation of what is to be inferred from the scores and the presentation of data to support such inferences" (Messick, 1979, p. 5).

### Needs for the Future

Possibly the greatest short-coming of criterion-referenced measurement is the relative lack of attention that is given to questions of validity of the measures. The clear definitions of content domains and well-specified procedures for item generation of some of the better criterion-referenced measures places the content validity of the tests on much firmer ground than has been typical of other types of achievement tests. Content validity provides an excellent foundation for a criterion-referenced test; but, as was argued above, more is needed to support the validity of inferences and uses of criterion-referenced tests. Unfortunately, the accumulation and reporting of evidence to support the uses and interpretations of criterion-referenced tests is the exception rather than the rule.

In their review of 12 commercially prepared criterion-referenced tests, Hambleton and Eignor (1978) did not find a single one that had a test manual that included satisfactory evidence of validity (Hambleton, 1980). Validity has too often been assumed by both developers and users of criterion-referenced tests. This is no more acceptable for a criterion-referenced test than it is for any other test. It is time that questions of validity of the uses and interpretations of criterion-referenced tests be given the attention they deserve.

### References

- Alderman, D. L., Swinton, S. S., & Braswell, J. S. Assessing basic arithmetic skills and understanding across curricula: Computer-assisted instruction and compensatory education. *The Journal of Children's Mathematical Behavior*, 1979, 2, 3-28.
- American Psychological Association. *Standards for educational and psychological tests*. Washington, DC: Author, 1974.
- American Psychological Association, Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. Washington, DC: Author, 1975.
- Berk, R. A. Item analysis. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, 1980.
- Berk, R. A. The application of structural facet theory to achievement test construction. *Educational Research Quarterly*, 1978, 3, 62-72.
- Bormuth, J. R. *Development of readability analyses* (Final report, Project No. 7-0052, Contract No. OEG-3-7-070052-0326). U. S. Department of Health Education and Welfare, Office of Education, March 1969.
- Bormuth, J. R. *On the theory of achievement test items*. Chicago, IL: University of Chicago Press, 1970.
- Cronbach, L. J. Test validation. In R. Thorndike (Ed.), *Educational measurement*, (2nd ed.). Washington, DC: American Council on Education, 1971.

- Cronbach, L. J. *Validity on parole: How can we go straight? New directions for testing and measurement*. San Francisco: Jossey-Bass, 1980, 99-108.
- Diederich, P. B. Review of *On the theory of achievement test items* by J. R. Bormuth. *Educational and Psychological Measurement*, 1970, 30, 1003-1005.
- Dunnette, M. D., & Borman, W. C. Personnel selection and classification systems. *Annual Review of Psychology*, 1979, 30, 477-525.
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. Uniform guidelines on employee selection procedures. *Federal Register*, 1978, 43 (166), 38290-38315.
- Filby, N. N., & Dishaw, M. M. Construct validation of group-administered achievement tests through individual testing (Beginning Teacher Education Study Technical Note BTES III-A). San Francisco: Far West Laboratory, 1977.
- Finn, P. J. A question-writing algorithm. *Journal of Reading Behavior*, 1975, 7, 341-367.
- Fitzpatrick, A. R. *The meaning of content validity*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.
- Fremer, J., & Anastasio, E. J. Computer-assisted item writing—I (Spelling items). *Journal of Educational Measurement*, 1969, 6, 69-74.
- Guion, R. M. Content validity—The source of my discontent. *Applied Psychological Measurement*, 1977, 1, 1-10.
- Guion, R. M. "Content validity" in moderation. *Personnel Psychology*, 1978, 31, 205-213.
- Guion, R. M. On trinitarian doctrines of validity. *Professional Psychology*, 1980, 11, 385-398.
- Hambleton, R. K. Test score validity and standard setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, 1980.
- Hambleton, R. K. Advances in criterion-referenced testing technology. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology*. New York: Wiley, in press.
- Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement*, 1978, 15, 321-327.
- Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, 48, 1-47.
- Harris, C. W., Pearlman, A. P., & Wilcox, R. R. *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation, No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977.
- Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation* (CSE Monograph Series in Evaluation, No. 1). Los Angeles: University of California, Center for the Study of Evaluation, 1973.
- Koslin, B. L., Koslin, S., & Zeno, S. Towards an effectiveness measure in reading. In R. W. Tyler & S. H. White (Eds.), *Testing, teaching, and learning: Report of a conference on research on testing*. Washington, DC: National Institute of Education, 1979.
- Linn, R. L. Issues of validity in measurement for competency-based programs. In M. S. Bunda & J. R. Sanders, *Practices and problems in competency-based education*. Washington, DC: National Council on Measurement in Education, 1979.
- May, L. J., & Hood, V. R. *Basic arithmetic skill evaluation*. Kankakee, IL: Imperial International Learning Corporation, 1973-1974.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, 30, 955-966.
- Messick, S. *Test validity and the ethics of assessment*. Paper presented at the annual convention of the American Psychological Association, New York City, September 1979.
- Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, CA: McCutchan, 1974.
- Millman, J. Computer-based item generation. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, 1980.
- Millman, J., & Outlaw, W. S. Testing by computer. *AEDS Journal*, 1978, 11, 57-72.
- Popham, J. S. *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall, 1978.
- Popham, J. S. Domain specification strategies. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: Johns Hopkins University Press, 1980.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests* (Studies in Mathematical Psychology I). Copenhagen: Danmarks Paedagogiske Institut, 1960.
- Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-

referenced test item validity. *Dutch Journal of Educational Research*, 1977, 2, 49-60.

Tenopyr, M. L. Content-construct confusion. *Personnel Psychology*, 1977, 3, 47-54.

Wardrop, J. L., Anderson, T. H., Hively, W., Anderson, R. I., Hastings, C. N., & Muller, K. E. *A framework for analyzing reading test characteristics* (Technical Report No. 109). Urbana: Univer-

sity of Illinois, Center for the Study of Reading, 1978.

#### **Author's Address**

Robert L. Linn, College of Education, University of Illinois, Urbana, IL 61801.