CRITERION-CENTERED RESEARCH

ALEXANDER W. ASTIN
National Merit Scholarship Corporation

This paper is concerned with clarifying certain issues regarding criterion measures and their use in educational and psychological research. The problems to be considered will be discussed separately in three general categories.

1. The Nature and Role of the Criterion (definitions, common fallacies about criteria, and certain logical and technical considerations in developing criterion measures).
2. Criteria and Test Development (the function of criteria in the construction and validation of tests).
3. Criterion-Centered Research versus Construct Validity (similarities and differences between the two approaches, and the case for criterion-centered research).

The Nature and Role of Criteria

Since much of what has been written about the criterion is unclear and even contradictory, it is necessary to clarify the role of criterion variables in research, to point out several common fallacies about criteria, and to discuss some of the logical problems involved in criterion development.

Definitions

The following definitions help to indicate what is meant by the word criterion:

---

1 This paper is part of the research program of the National Merit Scholarship Corporation and was supported by a grant from the National Science Foundation.
“a comparison object, or a rule, standard or test for making a judgement . . . a behavior goal by which progress is judged . . . the variable, comparison with which constitutes a measure of validity” (English and English, 1958, p. 130);

“a behavior or condition which is or can be described in terms of an ideal . . . a goal . . . behavior which is considered desirable and toward which one works” (Jensen, Coles, and Nestor, 1955, p. 58);

“quantification of need-satisfaction” (Gaylord and Stunkel, 1954, p. 297).

Common to each of these definitions is the idea that the criterion represents something important or desirable. Bechtoldt (1959) has pointed out that to designate one of a set of variables as “the criterion” is to create a “status difference” between that variable and all others.2

In any area of applied research, the criterion measure is an operational statement of the goals or desired outcomes of the program under study. For this reason, investigators ordinarily have little interest in predictor or independent variables, except as they are related to the criterion. Antecedent variables are considered valuable only if they in some way forecast or modify the criterion; on the other hand, the importance of the criterion is assumed and is not contingent upon its relationship to any antecedent variable.

One distinguishing characteristic of the criterion which is seldom mentioned in the literature is its essentially ecological nature. In contrast to a purely psychological construct or trait, a criterion variable usually refers to a relationship between the person and his environment.3 It is, in fact, difficult to speak of “standards” of performance or of behavior as being “desirable” without also defining the social context in which the behavior occurs.

In a sense, many current social problems are criterion problems. For example, the reluctance to pay teachers salaries which are commensurate with their formal training probably stems from ignorance about the results of the teacher’s efforts. We cannot assess how “valuable” a teacher is because we have no criteria for evaluating what he does. Even in higher education, there are no criteria for assessing teaching proficiency; a professor’s competence is usually measured in terms of the number of articles he has published.

The field of medicine offers another example of the central role of the criterion. The surgeon is higher-paid and enjoys higher prestige than almost any other medical specialist. The psychiatrist has much less status. Part of this discrepancy can probably be traced to criterion problems: while the outcome or product of the surgeon’s efforts is easily observable and relatively unambiguous (removal of diseased tissue, elimination of pain, prolongation of life, etc.), it is difficult even to define what the psychiatrist is trying to do, much less to judge how well he does it. In short, there are as yet no good criteria (standards of performance) for the psychiatrist.

The Conceptual Criterion

A good deal of ambiguity results from the frequent use, sometimes interchangeably, of such terms as “criterion,” “criterion performance,” “ultimate criterion,” “criterion measure,” etc. A brief discussion of the procedures usually followed in criterion development may help to resolve these ambiguities.

Ordinarily the applied researcher is confronted with some “problem” which involves, at least implicitly, what may be called a conceptual criterion. This conceptual criterion is a verbal statement of important or socially relevant outcomes based on the more general purposes or aims of the sponsor.4 Typical problems and conceptual criteria, viewed in terms of the sponsor’s goals, are the following.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Conceptual Criterion</th>
<th>Relevant Goal(s) of Sponsor</th>
</tr>
</thead>
<tbody>
<tr>
<td>To compare different methods of teaching</td>
<td>Effectiveness of teaching Education, optimal development of human talent</td>
<td></td>
</tr>
<tr>
<td>To test the effectiveness of psychotherapy</td>
<td>Social adjustment Mental health</td>
<td></td>
</tr>
<tr>
<td>To train better pilots</td>
<td>Proficiency in flying National defense, maintenance of peace</td>
<td></td>
</tr>
<tr>
<td>To increase efficiency</td>
<td>Production/costs Profit</td>
<td></td>
</tr>
</tbody>
</table>

2 Some investigators tend to equate the terms “criterion” and “dependent variable.” In this paper, however, “criterion” is used only with reference to variables or events which are judged to have immediate social relevance or importance.

3 Criterion variables appear to be what Allport (1960) has termed “transactional” phenomena in personality theory.

4 The term “sponsor” usually designates the organization which underwrites a study (Bellow, 1941; Fiske, 1951; Vallance, Glickman, and Sue, 1953), but the sponsor may be any organization or individual, including the investigator himself.
Since the sponsor's goals are in each instance based on considerations of value, different conceptual criteria can be judged in terms of their relative goodness or desirability. More abstract, higher-order goals might include "the happiness and long life of the individual" or "the welfare of humanity." The conceptual criterion represents the lowest level of abstraction in the sponsor's hierarchy of relevant goals.

A criterion performance is any observable event which is judged to be relevant to the conceptual criterion. For example, clinical ratings of social adjustment, number of job changes, self-ratings, etc. are performances which might be judged relevant to "social adjustment." Such performances provide the raw data from which criterion measures are derived. In some instances the method used to observe the performance provides data which are already in the form of measures, while in others the observations must be combined or manipulated to obtain satisfactory criterion measures.

Although operationally we cannot differentiate between the conceptual criterion and the criterion performance, the construct of conceptual criterion is important for at least two reasons. First, the relevance of any criterion performance and the history of its selection are better understood if the conceptual stage is documented. Second, because the conceptual criterion usually implies something more than the actual criterion performance, it may prove to be a valuable source either of future criterion measures or of improvements in existing ones.

The principal requirement for any criterion performance is, of course, that it be relevant to the conceptual criterion. Since conceptual criteria are rational rather than empirical, the relevance of a criterion performance can be judged only on rational grounds. To illustrate: changes in a student's achievement test score (criterion performance) constitute a measure of "effective teaching" (conceptual criterion) only if one is willing to assume that the student's score is an important or socially relevant variable to be manipulated by the teacher. To test this isomorphism empirically would require an independent measure of the teacher's effectiveness, and the relevance of that measure to the conceptual criterion would first have to be established in the same rational manner. Similarly, the comparative relevance of any performance would have to be determined non-empirically: e.g., a student's score on an achievement test is more relevant to "effective teaching" than is, say, the teacher's knowledge of educational theory, only if we believe it to be so on rational grounds. The question of whether or not a criterion performance is "immediate" or "ultimate" (Thurstone, 1949; Cuthton, 1951), therefore, is basically a judgmental—rather than an empirical—matter.

Normally it is the sponsor's responsibility to designate the conceptual criterion and the researcher's responsibility to develop criterion measures. Since the researcher's task becomes easier as the meaning of the conceptual criterion is made clearer, it is of central importance to define the conceptual criterion unambiguously. In particular, the sponsor should define what he sees as the limits of his responsibility. For instance, the educator must decide if the teacher's effectiveness includes his possible influence on the student's home life, social life, and mental health; the psychiatrist who wants the researcher to measure "social adjustment" must decide if he is to be in any way responsible for the patient's social adjustment after treatment is terminated. In these and many other examples the definition of the conceptual criterion depends heavily on the breadth of responsibility assumed by the sponsor.

Fallacies about Criteria

Perhaps the most common misconception about criterion measures is the notion that they can be "validated." It should be clear from the preceding discussion that the only method for "validating" a criterion measure is a logical analysis of its relevance to the conceptual criterion. Although some authors have alluded to this point (e.g., Anastasi, 1950; Brogden and Taylor, 1950), it is frequently overlooked in discussions of criterion development. One commonly held view is that the "validity of sets of criterion measures is evaluated in terms of . . . intercorrelations among the measures" (Bechtoldt, 1951, p. 1252). However, for this assertion to hold true, the investigator must first make a complex set of assumptions about the relevance of common and specific factor variance among the measures to the conceptual criterion. (These assumptions will be discussed in the ensuing paragraphs.) Similarly, the sugges-
tion that criterion measures can be "improved" by studying their correlations with predictors (e.g., Gulliksen, 1950, pp. 512–513) involves circular reasoning or, at best, misuse of terms. Only when predictors cease to be predictors and become (more relevant) criterion measures, can they be used in this way. In other words, sooner or later in his search for measures to approximate the conceptual criterion, the investigator must settle for something. Once the criterion performance is selected, it has, by definition, validity.

In most applied studies, more than one kind of performance is identified as relevant to the conceptual criterion. In order to deal with multiple criterion elements, investigators sometimes attempt to combine the several elements into a single criterion measure. Although the use of a single criterion score simplifies the mechanics of prediction, it is subject to the following error: if the separate elements measure relatively independent criterion-relevant factors, the use of a single score will confound these factors, thereby attenuating the validity of any predictor and making less apparent the effect of any independent variable.

Another approach to the problem of multiple criterion elements is to identify the common factor variance among the separate elements by means of factor analysis (e.g., Super and Overstreet, 1960; Taylor, Smith, and Ghiselin, 1959). Criterion elements which have the highest factor loadings are retained as the best measures of the "criterion factors," and criterion elements with relatively low factor loadings are either given less weight or discarded. In employing such a procedure, however, the investigator runs the risk of defeating his purpose unless his data can satisfy the following assumptions.

1. That no significant amount of variance relevant to the conceptual criterion is specific to any of the elements. (This assumption is especially difficult to defend where the number of separate elements is relatively small.)

2. That the proportion of common factor variance among the separate elements which is "error" variance (i.e., variance which is not relevant to the conceptual criterion) is negligible.8

---

8It has been suggested that ratings be weighted in proportion to their correlations with an "over-all" rating (e.g., Cureton, 1951, p. 638); if this technique is used, there is the danger that correlated error variance will be treated as if it were variance relevant to the conceptual criterion.

In short, the belief that multiple criterion elements should be combined "empirically" on the basis of their intercorrelations fails to deal directly with the fundamental problem of relevance. In the final analysis, some judge, whether it be the investigator himself or a panel of "experts," must decide how relevant each element is to the conceptual criterion. Any judge who undertakes to weight or otherwise combine multiple criterion elements should be aware of the following considerations.

1. The comprehensiveness of each element with regard to the conceptual criterion; i.e., are there aspects of the conceptual criterion which are unexpressed or only partially expressed by the element?
2. The extent to which the element contains non-relevant (error) variance.
3. The extent to which the intercorrelations with other elements are a function of variance in the conceptual criterion.

The "objectivity" of statistical intercorrelations in no way relieves the judge of the responsibility of making such decisions. (Perhaps the only limitation imposed by intercorrelations is that the judge cannot legitimately conclude that two elements whose correlation is unity or near unity are measuring relatively independent aspects of the conceptual criterion.) In other words, intercorrelations do not necessarily provide a reasonable basis for combining several criterion elements as long as the differential relevance of these elements is the basic consideration.

Criteria and Test Development

Of the four types of test validity outlined in the American Psychological Association's "Technical Recommendations for Psychological Tests and Diagnostic Techniques" (1954), only concurrent and predictive validity involve the use of criteria. Tests which have content validity need no criteria, since to assign content validity to any test is in essence to treat it like a criterion performance. Content validity, then, is established in precisely the same rational fashion as any other criterion performance. For the purposes of construct validity, measures of socially relevant behaviors or events no longer assume the role of criteria, since they are not actually accepted as any kind of goal or standard of performance.
against which the test is compared. (This point is discussed more fully in the next section.)

In developing a test which is to be used primarily in an applied setting, investigators frequently construct, “standardize,” and market the test first, and look around for criteria against which to “validate” it afterward. It is an understatement to say that this practice is an anachronism. Indeed, the earlier discussion about the primacy of criteria vis-a-vis predictors would imply that such tests have no raison d'être in the absence of meaningful criteria.

In some respects the “Technical Recommendations” support the notion of “tests first, criteria later.” Although it avoids making suggestions about how tests should be constructed, this manual frequently cautions the test-builder about how he should validate his test once he has constructed it. The manual expresses the hope that such recommendations will “not interfere with innovation,” but even if they do not, they do tend to reinforce the notion that the investigator need not concern himself directly with the practical uses for which his instrument is intended (and, hence, with criteria) until after he has developed it. However, if a test is designed for some purpose—to accomplish certain concurrent or predictive ends in an applied setting—there is little point in constructing the test without first defining those ends in terms of a set of criteria. For this reason, specification of conceptual criteria and some attempt at criterion development appear to be important preliminaries to the construction of any test which is designated for applied use.

The endorsement of criterion development as a first step in test development should not be construed as an endorsement of “empirical” over “rational” methods. On the contrary, it has already been proposed that criterion development is fundamentally a rational, non-empirical procedure. Even if criterion development never gets beyond a precise statement of the conceptual criteria, such a statement will aid in the development of a useful instrument.

It is perhaps unfortunate that the use of criteria in test development has come to be associated with “dust bowl” empiricism. Kelly (1958) has observed that empiricism “takes its criteria for granted . . . [so it can be] tough-minded about its predictors” (p. 335). However, whether or not this attitude constitutes “simple-minded credulity,” as Kelly puts it, depends upon the adequacy (relevance) of the criterion measures which are used. Investigators who object to empirical keying as a method of test construction might reflect for a moment: is the empirical method really the issue, or is their quarrel rather with the relevance of the particular criterion measures? Empiricism is simply a “trying-out” of potential test elements on the criteria, and irrelevant or poorly defined criteria are always a liability, no matter which method one follows in test construction.

There are undoubtedly many reasons why investigators have tended to be more test-oriented than criterion-oriented. For one thing, tests are usually much easier to build than criterion measures. For another, the conceptual criteria have never been adequately defined in some applied fields. In educational and clinical psychology, for instance, a plethora of tests masks profound inadequacies in existing criteria. Until more satisfactory criteria are developed, the use of “standard” tests (Sundberg, 1961) would appear to be premature.

It is likely that the number of new tests designed primarily for applied use would be reduced drastically, if each author were required to state, in terms of a set of criterion measures, some of the ways in which his proposed test would be used. A thorough analysis of the concurrent uses, in particular, to which some tests are put might reveal that the test is unnecessary. For example, in clinical psychology, many hours are spent on differential diagnosis and classification. The patient is tested and interviewed to determine how he should be categorized according to a number of classificatory schemes—“ego strength,” “reality contact,” psychiatric nosology, etc.—which represent concurrent “criteria” assessed by tests and other techniques. Even if we assume that the tests make these discriminations efficiently, we must face a more fundamental issue: are these concurrent variables themselves related to the conceptual criteria toward which clinical activity is ultimately directed (“cure,” “rehabilitation,” etc.)? The time which the clinical researcher now spends improving the techniques used to make these concurrent discriminations might be better spent developing criterion measures which are more relevant to the ultimate goals of the clinic. More-
over, such measures would serve as standards by which concurrent variables, tests, and even therapeutic techniques could be evaluated.

Criterion-Centered Validity versus Construct Validity

Most of the opposition to the criterion-centered approach in test development has come from the advocates of construct validity (Cattell, 1946; Jenkins and Lykken, 1957; Jessor and Hammond, 1957; Loevinger, 1957). In order to clarify the issues involved, some discussion of the similarities and differences between criterion-centered research and construct validity would seem to be indicated.

The Conceptual Criterion

The conceptual criterion is a verbal statement of the events or class of events which the sponsor deems important. The task in developing criterion measures is to state these events in operational terms. It is impossible to determine empirically whether or not the final criteria "really" measure the conceptual criterion; instead the investigator must accept the assumption that the particular behavior or event is relevant to the conceptual criterion. At this point of acceptance, the verbal statement (conceptual criterion) is replaced by a set of operations (criterion measure). The role, then, of the conceptual criterion is to provide a conceptual framework from which operations acceptable to the investigator or sponsor as having intrinsic importance can be derived. The conceptual criterion is not a theoretical proposition which can be tested empirically. There is nothing necessarily "psychological" about the conceptual criterion; it is, in fact, more sociological in nature. Nor is it assumed that the operations are homogeneous or unidimensional: for example, there may be several statistically independent behaviors which are judged relevant to the conceptual criterion of "social adjustment."

The Psychological Construct

In contrast to the conceptual criterion, the "construct" as described by Cronbach and Meehl (1955) is most certainly psychological in nature (synonyms used by these authors include such words as "trait," "psychological process," "quality," and "attribute"). The construct and its associated net represent a kind of working hypothesis about some attribute which is possessed by people; this hypothesis is in turn subject to confirmation, refutation, or modification through empirical research.

Methods Used by the Two Approaches

There is one basic similarity between criterion-centered and construct-oriented research: both attempt to account for variance in some variable. In criterion-centered research the variable is the criterion; in construct validity the variable is the test. However, the two approaches differ markedly with respect to the roles which they assign to theory, tests, and criterion measures.

Use of Theory

For the criterion-centered investigator, theory is merely a tool. He uses theoretical propositions as conceptual guides in the selection of potential predictors or experimental treatments. In particular, he is likely to make use of environmental constructs because of the sociological nature of the criterion. Since he is primarily interested in improving prediction of and control over the criterion, the criterion-oriented researcher is not necessarily committed to testing any particular theory or to investigating any specific psychological or environmental constructs.

Some construct-oriented investigators also regard psychological constructs as tools: they are interested in discovering what constructs account for variation in test performance. For others a particular construct is the center of interest ("what is anxiety?"). Environmental constructs, however, in general constitute an undesirable source of variance which the investigator usually wants to eliminate.

Use of Tests and Criteria

Criterion-centered investigators use tests in much the same way that they use theory: to aid in accounting for variance in the criterion measures. In some instances, tests may be accepted as criterion performances (i.e., in the case of "content" valid tests or tests

---

8 Cronbach and Meehl (1955) suggest that there are two types of construct validation studies: (a) those in which a particular test is the focus of concern, in which case the investigator may want to determine what accounts for variance in the test scores; and (b) those in which the construct underlying the test is the focus of concern, in which case the investigator may be more interested in testing those parts of the theory associated with the construct.
which are known to predict other criterion measures), but the criterion-centered researcher is usually unwilling to accept test performances \textit{per se} (e.g., answers to ambiguous true-false questions, drawings, sentence completions, etc.) as being relevant to the conceptual criterion.

The use of criterion measures in construct validity is similar to the use of tests in criterion-centered research. For example, in attempting to test certain parts of his nomological net, the construct-oriented researcher may utilize a particular criterion measure, but only if he expects his construct to be exhibited in some way in the criterion performance. However, his choice of criteria is determined by the particular construct(s) in which he is interested, \textit{not} by his recognition of the social relevance of the criterion. Thus, in construct validity a criterion measure, in a sense, loses its status.

\textit{The Case for Criterion-Centered Research}

The advocates of construct validity have two basic criticisms of criterion-centered research. First, they maintain that because test-criterion correlations are always relative to the specific practical situation which determines the criteria, such correlations have limited generalizability and, hence, are of limited usefulness ultimately. Measures of "fundamental" or "pure" traits, they say, will eventually prove to be more useful in a wider variety of situations. Second, they believe that criterion-centered research, because of its focus on practical, decision-making situations, is atheoretical or at best of limited usefulness to psychological theory; construct validity, on the other hand, is of more theoretical value because it attempts to test and elaborate psychological theories.

Before discussing these objections further, some points should be clarified. First, the purposes of criterion-oriented research do not preclude the development and use of trait measures. As already noted, once criterion measures are developed, researchers can and usually do attempt to assess those relatively static personal attributes which are likely to be related to the criterion performances. Second, even

\footnote{Cronbach and Meehl (1955) and others sometimes use the term "criterion" in reference to the \textit{trait} under study: e.g., "For most tests intended to measure constructs, adequate criteria do not exist" (p. 291). It should be noted that, unless the \textit{trait} itself is judged to be a socially relevant event (conceptual criterion), even a perfect test of that trait would not be a "criterion measure" in the sense that the term is used in this discussion.}

the construct-oriented researcher must sooner or later resort to the use of criterion measures if he hopes to demonstrate that "[practical] decisions . . . will also be best served by a predominantly trait-oriented psychometrics" (Loevinger, 1957, p. 640). The distinction, then, is one of focus: whereas construct validity involves conceptualizing at the level of human needs, values, abilities, etc., criterion development first requires conceptualizing at the level of \textit{in situ} behavior which is judged to be of societal importance.

One objection to criterion-centered tests probably stems from the well-known fact that the standards of performance on which criterion measures are based can be changed arbitrarily from one time to another. It is feared that any criterion-keyed test is of only temporary value and has very limited generalizability. The issue of generalizability can be recast in the form of a question: which kind of test is more likely to predict socially relevant behavior in a variety of situations—one which was built to predict socially relevant behavior in one specific situation, or one which was built to measure a specific psychological trait?

To say that test-criterion correlations are always "relative" to the particular applied situation may be a statement of fact, but the same statement probably applies equally well to any empirical relationship. In other words, the investigator's orientation toward "pure" or "fundamental" traits does not make his empirical findings any less relative to the circumstances under which they were obtained.

When the actual operations of criterion-centered research are scrutinized, the second objection, that it is atheoretical, also appears to be questionable. Once the criterion-centered researcher has developed criterion measures, it is only through extensive preliminary theorizing that he has any reasonable hope of conducting successful research, "success" in this case meaning the efficiency with which he is able to predict or control variation in the criterion. As indicated earlier, this theorizing involves the investigator's knowledge and intuitions about those personal attributes and environmental conditions (and their interactions) which are likely to be related to the criterion performance.

Criterion-centered researchers often fail to elaborate or even to make explicit their reasons for selecting potential independent variables. Even though "shotgun" studies may produce useful infor-
In short, criterion-centered studies are not intrinsically atheoretical; they rely heavily on the correctness of the investigator's theorizing. If the researcher's methods are objective, and if the theory underlying his selection of independent or predictor variables is made explicit, his data have as much potential theoretical value as the data of any other researcher.

The assumption that construct validity is the best or the only method for educational and psychological research is a dangerous one, since it may prevent the researcher from coming to grips with the socially relevant problems in his field. He reassure himself that he will contribute more in the long run by studying "basic science" or by being "theory-oriented." The fact is that theories about socially relevant performances (e.g., Barker and Wright, 1954; Brunswik, 1949; Murphy, 1958) are needed just as much as theories about psychological processes or traits.

Some researchers contend that criterion-centered research is more within the province of sociology or ecology than of psychology (Littman, 1961). This sort of objection, however, ignores the fact that most researchers working in applied fields are implicitly committed to certain social goals, and that the psychological profession as a whole is explicitly committed to "contributing to the welfare of man" (American Psychological Association, 1953). By requiring that the goals of research be stated in operational terms, criterion development offers the investigator the only means of assessing how close he has come to attaining those goals.

REFERENCES


Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory." Psychological Reports, III (1957), 635–694. (Monogr. Suppl. 9.)


INTERDEPARTMENTAL VARIABILITY AND STUDENT EXPECTATIONS OF COLLEGE GRADES

LEWIS R. AIKEN, JR.
The University of North Carolina at Greensboro

It should be clear to any thinking person that academic evaluation is not an extremely objective procedure and that college grades are not fixed or immutable but have different meanings under different sets of conditions. For example, it has been shown that there is great variability in grading practices among instructors of the same subject (Thompson, 1958), among departments or subject matter areas (Bass, 1951; Morris, 1953), and even among instructors and departments combined across years (Webb, 1959; Aiken, 1963). There are many possible reasons for such variations and changes in grading. Morris (1953) discussed nine suggested causes for differences in grading practices among college departments or schools, but only the factor of class standing of students (lower versus upper division) was studied empirically. This factor was found to have a significant effect upon the grading practices of several schools and departments, viz., higher average grades being assigned to upper division students.

Acknowledging the fact that variations in grading standards exist, to what extent are students aware of these differences? Certainly they must have some notion of the difficulty level of a subject or of how well they will do before taking a course. Such "notions" or expectations are modifiable as experience in the subject accumulates, but even so they undoubtedly play an important role in individual "pace setting," i.e., the student does the amount of studying which he feels is necessary to obtain a given grade. The few re-