EDUCATIONAL THEORY
Summer 1985, Vol. 35, No. 3

The Origins and Functions of Evaluative Criteria

By D. Royce Sadler

Although other constructions have been put on the term, an evaluation, according to primary definition, is basically a judgment about quality or value. If the purpose of an evaluation is simply to arrange a set of things in order of quality without any being necessarily designated as good or bad, the judgments are said to be *relative.* However, if the purpose is to decide which are good and which are not, the judgments are called *absolute.* If something deserves to be called good, it is because the set of criteria is accepted as a reasonable definition of what *good* means in the context and because the object exhibits certain minimum degrees of intensity or levels of performance on those criteria. Absolute judgments presuppose standards, which are fixed points of reference along the relevant dimensions.

Criteria and *standards* are often used as interchangeable terms. For example, something is said to "meet the criterion" when what is meant is that there is enough of the characteristic present to exceed the minimum required. It would be more correct to say that the object "meets the standard." Similarly, the Joint Committee's *Standards*[1] is for the most part a compilation of criteria, not standards. A few, however, are auditable without further specification and therefore function as standards. A third example relates to the criterion-referenced testing movement, where Glass[2] has traced a shift in meaning of the term criterion. Originally, a criterion was a dimension of performance, but from about the time of Glaser[3] onwards, a criterion has been used to mean a cutting score that designates competence or mastery. Frequently it does not matter which term is used because the context makes the meaning clear.

Whether something that is called a criterion refers to a dimension or to a minimum qualifying level can sometimes be determined from the form of words. The noun *consistency* clearly refers to a dimension, the adjective *consistent* implies a threshold or standard. Properties that are potentially useful for evaluating all members of a given class are generally criteria, while those relating to an appraisal that has already been made imply (perhaps unwritten) standards. In this article, criteria refer to dimensions relevant to an evaluation and standards to particular levels used as reference points.

Consider the evaluation of unsolicited manuscripts submitted to a journal for publication. Suppose an author asks an editor: "What criteria are used in deciding which articles to publish?" The editor replies by giving three criteria that specify the most salient dimensions, namely, an article must show familiarity with recent work, be original, and make a significant contribution to knowledge. Now suppose that a particular article has been recommended by referees for publication, and a different question is put: "What makes this article a publishable piece?" The response is that the article shows familiarity with the field, is original, and makes a contribution to the discipline. These are obviously the same three that answer to the first question. But what is meant by the second list is that the article takes *sufficient* cognizance of recent work, is *sufficiently* original, and is a *sufficient* contribution to knowledge for it to be worth

Correspondence: Department of Education, University of Queensland, St. Lucia, Queensland 4067, Australia.

1. Joint Committee on Standards for Educational Evaluation, *Standards for Evaluations of Educational Programs, Projects, and Materials* (New York: McGraw-Hill, 1981).

2. Gene V Glass, "Standards and Criteria," *Journal of Educational Measurement* 15 (1978): 237-61.

3. R. Glaser, "Instructional Technology and the Measurement of Learning Outcomes," *American Psychologist* 18 (1963): 519-21.

publishing. It is obvious that no article can ever be completely up to date (if only for publishing lag), but neither would an author be likely to ignore completely other developments in the field. Familiarity with recent work is not a property that an article unequivocally does or does not possess. The notion of sufficiency in any list of properties describing something implies qualifying levels.

Ordinarily, evaluations are substantiated by reference to criteria and, except for purely relative judgments, to standards as well. Lists of criteria for all manner of educational evaluations abound in the literature: for essays and assignments,[4] oral history interviews,[5] classroom activities,[6] journal articles,[7] academic promotions,[8] reference books,[9] school programs,[10] and evaluation agreements.[11] However, no attempt is made in this article to survey sets of criteria for different types of appraisal. Although criteria and standards constitute two of the key elements in evaluative discourse, it is possible for reliable judgments to be made even when no criteria are used explicitly. In such cases, evaluations are valid to the extent that the evaluator is accepted as authoritative and competent. This aspect is explored in more detail later in the article.

ORGANIZATION OF CRITERIA

It is useful both conceptually and practically to analyze the linkages between the various elements of a system of valuation. One way to organize criteria is by means of a hierarchical structure. In general, any given criterion can be expressed either as a component of some higher-level criterion (moving up the hierarchy) or in terms of a number of lower-level criteria. The discussion that follows uses a concrete example, starts at a middle level, and then works downwards. However, the principles are easily generalized.

Consider the analysis by House[12] of the fairness of evaluation agreements. An evaluation agreement is the formal element in the contract between an evaluator and a sponsor that sets down the respective obligations and responsibilities of both parties. The concern here is not with the content of such an agreement, but with what makes an agreement a fair one. Beginning with the concept of justice-as-fairness as developed by Rawls,[13] House set up a three-tier hierarchy with justice at the apex. He argued that in relation to evaluation agreements, fairness could be judged in terms of twelve characteristics, among them rationality, noncoercion, and disinterestedness. That is, an agreement cannot be fair unless it is, among other things, rational, noncoercive, and designed so as not to serve vested interests. House then specified each of these characteristics in more detail. Rationality was defined in terms of three finer-grained properties: the ability to argue from one proposition to another, to give evidence for points of view, and to estimate and take into account the consequences of one's

---

4. Paul B. Diederich, *Measuring Growth in English* (Urbana, Ill.: National Council of Teachers of English, 1974).

5. F. J. Fonsino, "Criteria for Evaluating Oral History Interviews," *History Teacher* 13 (1980): 239-43.

6. James D. Raths, "Teaching without Specific Objectives," *Educational Leadership* 28 (1971): 714-20.

7. Duncan Lindsey and Thomas Lindsey, "The Outlook of Journal Editors and Referees on the Normative Criteria of Scientific Craftsmanship: Viewpoints from Psychology, Social Work, and Sociology," *Quality and Quantity* 12 (1978): 45-62.

8. Kala M. Stroup, "Faculty Evaluation," in *Issues in Faculty Personnel Policies,* New Directions in Higher Education no. 41, ed. J. W. Fuller (San Francisco: Jossey-Bass, 1983), 47-62.

9. B. C. Bell, "Choosing a Reference Book for Writing, *College Composition and Communication* 32, no. 1 (1981): 38-46.

10. Newton S. Metfessel and William B. Michael, "A Paradigm Involving Multiple Criterion Measures for the Evaluation of the Effectiveness of School Programs," *Educational and Psychological Measurement* 27 (1967): 931-43.

11. Ernest R. House, "Fair Evaluation Agreement," *Educational Theory* 29 (1979): 159-69.

12. Ibid.

13. John Rawls, *A Theory of Justice* (Cambridge, Mass.: Belknap Press, 1971).

actions. Similarly, noncoercion was defined in terms of power of one person over another, coalitions for the exercise of power, and desperate or stressful circumstances.

Developing a hierarchy downwards by specifying meanings for a primary criterion is a matter of interpretation and semantics, not of deduction. Furthermore, lower-level criteria frequently overlap because the meanings of words have fuzzy boundaries. This in most cases rules out the possibility of definitive decompositions. Equivalent decompositions can be developed, equivalent in the sense that two sets may "cover" the same content. This does not necessarily mean, however, that they would be equally useful. Adjudication between equivalent decompositions may be required, for example, when evaluations are intended to stimulate action. In evaluating an educational program, the people who have to make the day-to-day decisions in order to adjust the program while it is in operation may not accept an evaluation or recognize its practical implications if the criteria are couched in narrow psychological, philosophical, or sociological terms.[14]

While a criterion is given content by expressing it in terms of a number of simpler criteria, its connection with a higher-level criterion is brought to light by asking why the criterion is thought to be worthwhile. Taking one of House's middle-level criteria, for example, noncoercion is legitimized as a valid criterion because it is part of the concept of fairness. It is simply not fair to coerce people. But how is fairness itself validated as a criterion? For most people, fairness is adopted as a kind of fundamental criterion because it appeals directly to some deep-seated concept of decency or aesthetic sensibility. This means that regress in the upward direction at least is, for all practical purposes, not infinite. Although the notion of backstop or *zero-level* criteria is accepted by many (though not all) philosophers, such criteria go by different names. Najder[15] calls them *axiological values;* Rescher[16] calls them *underlying values* or *values proper*. In the limited context of program evaluation, House[17] examined the implications of three axiological values, namely truth, beauty, and justice. In the broad context of education, individual worth, growth, and freedom figure prominently as axiological values.

Suppose it is argued that fairness is not really a backstop criterion at all but that it can be justified pragmatically, in that unfair societies are unstable or violent (*bad* characteristics) and tend towards self-destruction. In this case, stability, nonviolence, and at a higher level survival, function as axiological values. The issue is not so much *where* one stops, but that one *has* to stop somewhere and adopt or choose a value as a personal conviction or at least as an operating principle. This is not to say that values ultimately have no empirical connections. They are more than simple matters of etiquette or taste, such as liking strawberry flavor, which are also accepted (or rejected) "just because." As abstract ideas, values are developed out of a complex interplay among experiences, social pressures, traditions, concrete valuations, and discourse about all of them. During their development, there occurs simultaneous clarification of (a) what the label of the value (for example, freedom) means and what it implies for action, and of (b) whether one approves, for reasons which may or may not be apparent at the time, of states which are classifiable as free. Ultimately, individual values are embedded in and vindicated by the rational choice of a way of life with a complete value system. It is this, says Taylor,[18] that rescues individual values from being merely arbitrary.

While the path which leads to the adoption of an axiological value may in principle be retraceable, in practice it is difficult if not impossible. However, what can be verified in part is whether a person's conduct and valuations are consistent with the values the person claims to hold. But the values themselves cannot be unambiguously inferred from what people actually do, nor even from their efforts to bring about a desired state

---

14. Jo Anne Pagano and Lawrence Dolan, "Foundations for a Unified Approach to Evaluation Research," *Curriculum Inquiry* 10 (1980): 367-81.
15. Zdzislaw Najder, *Values and Evaluations* (Oxford: Oxford University Press, 1975), 63-64.
16. N. Rescher, *Introduction to Value Theory* (Englewood Cliffs, N.J.: Prentice-Hall, 1969).
17. Ernest R. House, *Evaluating with Validity* (Beverly Hills, Calif.: Sage, 1980).
18. P. W. Taylor, *Normative Discourse* (Englewood Cliffs, N.J.: Prentice-Hall, 1961).

of affairs, because persons may, through ignorance or incapacity, be unable to engage in proper courses of action. Discrepancies between the values people espouse and what they do are not at all uncommon. As Najder[19] puts it, axiological values cannot be "safely deduced from empirical data and the rules of logic: they have to be chosen."

Being highly abstract, axiological values have an almost universal appeal so long as their content remains unspecified.[20] Once their content is determined, they appeal more to some people and less to others. For a first example, consider freedom in relation to education generally. Both Cooley and Lohnes[21] and Goodlad and others[22] cite freedom as a key educational value, but their conceptions of schooling and the processes used to assess the quality of education in terms of freedom are quite different. The former develop (absolute) freedom (of thought) through a psychometric frame of reference, the latter develop a conception of freedom-based education through a humanistic frame.

A second example concerns competing interpretations of justice as applied to admissions in higher education. Uniform application of a publicly declared evaluation *rule* (in the form of a set of clear, measurable criteria) obviously treats all cases alike, with no exceptions. In practice, this superficially just rule conflicts with a desire to do justice to individuals by treating each case on its merits. When the formal admission rules are found systematically to disadvantage minority groups, a case is often made for special admission provisions. The argument is along the lines that differences between persons are relevant to the aims of distributive justice only to the extent that the persons can be held responsible for those differences, and that one cannot be held responsible for one's sex or race, or even (some say) for one's academic achievement.[23] If this argument is accepted, the concept of a just distribution of resources based on merit has to have superimposed on it the Aristotelian concept of rectificatory justice. Under this interpretation, a new admissions policy may be formulated in such a way as to help redress imbalances in the past, the expectation being that imbalances from such causes in the future will be reduced. Interpretive differences of this kind occur at a more fundamental level than the one at which it is asked whether the criteria to be used in program evaluation should be those of the program managers, the clients, the general public, or the evaluator.

The implications of a group of axiological values within a given context need to be balanced against one another because the criteria deriving from different axiological values are not necessarily operationally independent. Freedom in the sense of the right to self-determination cannot be pursued without limit, for instance. It soon conflicts with the rights of others to do the same, that is, with equality of opportunity, another axiological value. The boundary between the two is movable and essentially quantitative: how much of *this* is compatible with how much of *that*. Conflicts of this type are obviously not fatal. They are resolved and working balances achieved in countless classrooms every day without fanfare or much conscious thought.

FUNCTIONS OF LOWER-LEVEL CRITERIA

Criteria of lower levels serve five important functions. First, they are useful in clarifying and communicating. By proposing an interpretation for a criterion at the next highest level, they enable others to assess the adequacy of a decomposition. Thus if a person were to disagree with House's statement that a particular evaluation agreement

19. Najder, *Values*, 176.

20. Ernest R. House, *The Logic of Evaluative Argument* (Los Angeles: Center for the Study of Evaluation, University of California, 1977), 12.

21. William W. Cooley and Paul R. Lohnes, *Evaluation Research in Education* (New York: Irvington, 1976).

22. J. I. Goodlad, G. D. Fenstermacher, T. J. LaBelle, V. D. Rust, R. Skager, and C. Weinberg, *The Conventional and the Alternative in Education* (Berkeley, Calif.: McCutchan, 1975).

23. Lois T. Weinberg, "An Answer to the 'Liberal' Objection to Special Admissions," *Educational Theory* 29, no. 1 (1979): 21-29.

is fair, it is because either the set of subcriteria is regarded as defective (with at least one element irrelevant, improper, or missing), or different standards are employed. In general terms, lower-level criteria provide a mechanism for exploring the reasons for disagreements about evaluations. Some disagreements are about the facts of a matter: how data should be collected, analyzed, and interpreted. That is why secondary analysis is so important. However, many disagreements are about the admissibility of criteria. Indeed, a common way in which one researcher criticizes another researcher's work is to suggest that while the judgment on a particular criterion may be conceded, the evaluation should be discounted because the *wrong* criteria were used or the criteria were poorly articulated.[24] Seidman[25] cited several cases where evaluations of educational programs had been ignored because the criteria were rejected.

Clearly, there are limits to how far the process of decomposition should proceed. While on the surface it may appear that the more detail we have, the better, in practice we are in danger of becoming swamped with atomistic detail, at the same time losing sight of what the overall evaluation is all about.

Second, criteria simplify the processes of judgment. By restricting the dimensions on which comparisons are possible, a value claim is made easier to establish. It is easier to judge whether a particular introductory course in calculus is a good one than to judge whether the mathematics curriculum of which it is a part is good, and this in turn is easier than judging whether the whole curriculum constitutes good education. But although the processes of judgment are simplified and made more objective,[26] a judgment made within a particular genre is not a substitute for and should not be confused with an appraisal *of* the genre itself. That a particular calculus course is a good one is not a judgment that calculus is a legitimate component of the curriculum.

The third function of lower-level criteria is that they set up an evaluation policy for the assessment of other cases in the same class (in House's example, evaluation agreements) with which other policies may be compared. Quality in education is ultimately defined in philosophical terms. Educational evaluations differ in degree, if not in kind, from evaluations of consumer products, health services, and business organizations. For this reason, the links between classroom experiences and fundamental human values have to be established and kept visible. In Najder's words, "Disclosure and presentation of the components of a system, and of its metaphysical underpinnings, will put that system to the test of one's beliefs, inclinations, and feelings."[27]

Connections up and down a hierarchy are generally easier to make than those among criteria at the same level. It is easier to see the connections between noncoercion and fairness, and between disinterestedness and fairness, than to see a direct connection between noncoercion and disinterestedness. Direct lateral connections are nevertheless sometimes possible, being especially interesting if they turn out to be circular, such as is occasionally the case with reputational criteria. Graduates from a good program can be expected to be superior to those from a poor program. The caveat is, of course, that the quality of a program should not be based primarily on the status and presumed quality of the graduates.

The fourth advantage of hierarchical structuring is that it allows the sufficiency of a set of criteria to be tested. Suppose *appropriateness* is a criterion for choosing a textbook for a course, along with cost, availability, and durability. Suppose further that appropriateness is broken down into a set of lower-level criteria, say, coverage of subject-matter content, depth of treatment, reading level, and pertinence of examples. The question is whether such a list can make claim to being exhaustive. This could be

---

24. Wilbur B. Brookover, "Review of *Why Do Some Urban Schools Succeed?* (The Phi Delta Kappa Study of Exceptional Urban Elementary Schools)," *Harvard Educational Review* 51 (1981): 439-42.

25. William H. Seidman, "Goal Ambiguity and Organizational Decoupling: The Failure of 'Rational Systems' Program Implementation," *Educational Evaluation and Policy Analysis* 5 (1983): 399-413.

26. Michael Scriven, *Primary Philosophy* (New York: McGraw-Hill, 1966), 53-55.

27. Najder, *Values*, 160.

tested by trying to find (or perhaps write) a text which satisfies all of the lower-level criteria but which would have to be disqualified on grounds other than those on the list. The set of lower-level criteria is satisfactory, and acceptable as a substitute for the higher-level criterion, to the extent that exceptions are rarely or never found.

Whether a set of subcriteria is adequate determines whether the overall judgment is essentially qualificatory of disqualificatory. When something measures up on an exhaustive set of subcriteria, it also (by definition) measures up on the main criterion. However, when the set of subcriteria is deficient, a weaker judgment is inevitable. All that can be determined is whether the phenomenon *fails* to meet the main criterion by failing on one or more components. It cannot be used in the opposite direction to confirm compliance with the main criterion. Regular checking up the hierarchy is necessary if an evaluator is to remain alert to the possibility that the whole may be more than the apparent sum of its parts. In the textbook example, the higher-level reflexive question would be: Given that a particular textbook is satisfactory with respect to coverage, depth, reading level, and examples, is it, for all that, *appropriate?*

The fifth and final practical advantage is that a hierarchy suggests where tradeoffs may be possible. Continuing with the textbook evaluation, it is a matter of simple economics that one cannot maximize both coverage and depth of treatment and still keep costs within reasonable bounds. Similarly, depth may be attainable only at the expense of a rise in reading level. What is necessary is a view of these criteria not as separate, independent entities, but as elements of appropriateness. In practice, the level on one criterion has to be traded against levels on others in such a way that reading level, depth, and coverage are not separately maximized (a physical impossibility) but that the *mix* is optimized, subject to the constraints. The target is to choose or to write the *most appropriate* textbook, given the circumstances. In this sense, the higher-level criteria act as coordinators and synthesizers of the lower-level criteria. A plethora of simple judgments, each of restricted scope, can then be replaced by fewer higher-level judgments which can be unpacked at will, should the need arise.

From a strictly logical point of view, hierarchical structuring has a weakness. On the one hand, axiological values have to be decomposed for their substance to be known (that is, criteria are necessary in order for content to be specified). On the other hand, axiological values develop only in the context of experiences, traditions, and valuations, that is, from the ground up. The arguments seem to run counter to each other. What saves the situation from inconsistency or tautology is an understanding of the nature of the concepts. Although the label of an axiological value is compact and convenient, it applies to a rich and generalized idea whose power lies in its ability to transcend particular cases. In any concrete situation, a meaning appropriate to the context has to be generated. There is no logical necessity for meanings to be expressed in identical terms in different contexts. Even if there were a genuine logical difficulty, this would not necessarily be a barrier to usefulness. The concept of cause is not satisfactorily resolved in philosophy, neither is probability adequately defined without axiomatization, yet both are splendidly useful concepts.

ORIGINS OF CRITERIA

Criteria are not arbitrary inventions and cannot be logically deduced from values for the reason that the latter are highly abstract and require their content to be specified. As stated earlier, this specification is a matter of interpretation and semantics and not, strictly, of logic. How, then, are criteria derived? The proposition about to be argued is that they are inferred from *valuations* as fundamental acts. To use Dewey's terms, valuations "*spring from* the immediate and inexplicable reaction of vital impulse and from the irrational part of our nature. . . . [V]aluations have their roots *in an existence* which, like any existence *taken in itself,* is *a*-rational."[28] "Human beings are continuously

28. John Dewey, "Theory of Valuation," *International Encyclopedia of Unified Science,* vol. 2, no. 4 (Chicago: University of Chicago Press, 1939), 18.

engaged in valuations. The latter supply the primary material for operations of further valuations and for the general theory of valuation."[29]

In other words, recognition, not reason, is the primary evaluative act and predates any criteria. It is common knowledge that the quality of things can often be recognized as excellent or beautiful before the rules for excellence or beauty are formulated explicitly, hence the observation, "Quality may be impossible to define but one knows it when one sees it." Criteria are generated as part of the process of reflecting about and providing rationalizations for judgments already made. In this sense, criteria are initially descriptive. But in standing between past and future evaluations, they begin to function normatively. That is, although criteria may be used to facilitate and substantiate evaluations, they are not absolutely necessary.

In practice, the fundamental valuations that give birth to criteria often begin with feelings of indignation at, or admiration for, the ways things are being done or are turning out. But this is not to suggest that personal valuations are a sufficient source of criteria. Some can arise indirectly through the work of others. Chen and Rossi,[30] for example, argue that basic theory-oriented research in the social sciences can sensitize program evaluators to possible effects that would otherwise go unnoticed. These effects are potential sources of criteria.

Evidence that global judgments are somehow fundamental acts and that criteria are derived from them follows from a principle found in many fields of inquiry. In 1759, for example, the philosopher-critic Edmund Burke tried to determine the "properties on which beauty depends; properties that operate by nature, and are less liable to be altered by caprice, or confounded by a diversity of tastes, than any others."[31] By looking for the properties that beautiful objects have in common, he finally settled on seven criteria of beauty. What is more interesting than the content of Burke's list or the quaintness of his analysis is his clearly inductive approach. A modern parallel is the computer analysis of spatial and auditory patterns. What is called feature extraction in the field of pattern recognition corresponds closely to the identification of criteria in evaluation.

Criterion identification in evaluation is but a special case of a more general procedure in which classes are identified first and the discriminating characteristics defined and described later. When conflict arises between holistic valuations and feature-by-feature appraisals, the priority of the former calls the criteria for the latter into question. Rarely is it the other way round. Even where rational valuations (which are made and argued for in terms of levers called *criteria*) demonstrate utility in separating difficult cases, the function is not so much to provide the separation as to remove the appearance of arbitrariness when superiority is not immediately obvious and a judgment must be made.

The existential basis of criteria underlies Scriven's[32] concept of *goal-free* evaluation, in which the objectives of a project are deliberately set to one side while the evaluation proceeds. In the particular evaluation described by Scriven, it was obvious that one of the best projects scored poorly on the prespecified criteria but possessed other positive qualities (*side-effects* to the project directors) which were too important to be ignored.

Indirect evidence for the existential basis of criteria is to be found in the discretionary clauses written into valuation policies. It is commonly accepted that an appraisal system (for selecting students for higher education or for promoting faculty, for instance) should be open and applied uniformly, unless there are extraordinary circumstances. The inclusion of clauses that allow for exceptional cases is tacit recognition that although

---

29. Ibid., 58.

30. Huey-Tsyh Chen and Peter H. Rossi, "The Multi-goal, Theory-driven Approach to Evaluation: A Model Linking Basic and Applied Social Science," *Social Forces* 59 (1980): 106-22.

31. Edmund Burke, *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful,* 2nd ed. (London: Dodsley, 1759), sect. 18 (facsimile edition, New York: Garland, 1971).

32. Michael Scriven, "Prose and Cons about Goal-free Evaluation," *Evaluation Comment* 3, no. 4 (1972): 1-4; and Scriven, "Goal-free Evaluation," in *School Evaluation: The Politics and Process,* ed. Ernest R. House (Berkeley, Calif.: McCutchan, 1973), 319-28.

a clear policy is necessary in the interests of justice and efficiency, special cases sometimes warrant special criteria. If these nonstandard criteria could be anticipated and set out in advance, they could in principle and should in practice be incorporated into the general assessment policy.

The transition from reactive or holistic valuations to rational valuations (that use criteria) may be assisted by empirical research. The study of human judgmental processes has a large literature of its own[33] and is not pursued at any length here, but the basic principles are important. If a number of characteristics can be coded for a large number of cases, criterion usage can be explored statistically. For example, a great deal of research has been directed toward trying to track down the characteristics of good practitioners in a variety of professions, including teaching. Even though it may be possible to identify excellent teachers, it is often not intuitively obvious what the operative criteria are, nor how they interact. Using expert judgments of teaching quality as the dependent variable, a multiple regression equation can be constructed to find those criteria which weight strongly in the overall assessment. (A variation of this approach is to use some exogenous variable such as student achievement in place of expert assessments of teaching excellence.) What such studies have shown is that occasionally a particular criterion which an evaluator has claimed to be important in a judgment weights zero in the regression. This indicates that the criterion was either not being used at all or that it correlated so highly with the other criteria that it was redundant.

However, these studies must be recognized for what they are. They do not set out so much actually to discover criteria as to confirm or disconfirm conjectures. The regression equation is a heuristic for determining the relative importance of criteria that are already considered potential contributors. Between the global evaluation (called the *criterion* in the psychometric literature) and the empirical validation of the criteria which appear to be important (the *predictors*) is interposed a constructive or creative act in which potential criteria are hypothesized. This stage is neither fully inductive nor fully deductive. More importantly, the set of criteria proposed need not be exhaustive. There may well be other criteria which are important but which are not included in a particular research study.

Regression studies sometimes lean toward blind empiricism. Although criteria are identified by searching valuations for recurring patterns, mere concurrence of a property with valuations of a given class is generally insufficient to warrant calling that characteristic a criterion of quality. Logically, there must be some epistemological connection between the property and the valuation as well. Thus hair color in a teacher, even if it were correlated with performance, would not normally be acceptable as a criterion unless a line of argument could be developed linking the two.

A more serious limitation of such studies is that they rely on the analysis of many similar cases and search for criteria that are appropriate to all cases in the genre. Commonality is the principle that is taken to validate a set of criteria. This is feasible when investigating the evaluation of, say, student essays that are all on the same topic.[34] But as the variety of cases expands, the probability that a common set of criteria will be uniformly applicable decreases, simply because the number of potential criteria which could be drawn upon is large. For purely practical reasons, not all of them can be used at the one time unless in the form of a checklist. In other words, the highest-common-factor approach is certain to do injustices to particular cases except when the cases are similar. Perfectly valid evaluations can be made using little-used criteria which do not appear often enough to figure in formal statements of evaluation policy. In student assessment, a certain tension exists between reliability

33. Paul J. Hoffman, "The Paramorphic Representation of Clinical Judgment," *Psychological Bulletin* 57 (1960): 116-31; and Robyn M. Dawes and Bernard Corrigan, "Linear Models in Decision Making," *Psychological Bulletin* 81 (1974): 95-106.

34. S. W. Freedman, "How Characteristics of Student Essays Influence Teachers' Evaluations," *Journal of Educational Psychology* 71 (1979): 328-38.

and validity, especially in tasks calling for originality and creativity. In multisite program evaluations, it appears as a tension between achieving consistency across sites and site-specific validity. One of the criticisms levelled by House[35] at the evaluation of the Follow Through program was that the evaluators were forced to use a small number of quantitative measures, mostly based on achievement test scores. The apparent need to scale the effects of different realizations of the program so that comparisons could be made drew the evaluation away from the program's original set of broad social goals. The scope of the evaluation was thereby narrowed dramatically.

Finally, measures are necessary for statistical analysis. Conceptualizing a criterion as an idea is different from devising a measure that is both intuitively satisfactory and useful for distinguishing degrees of quality. Criteria for which adequate measures are difficult to define are obvious candidates for neglect in quantitative studies.

OBJECTIVES AND CRITERIA

A goal or objective is a statement about an intended outcome or a desired state of affairs. Because one normally wants to achieve something considered potentially valuable, objectives are themselves derived by way of previous valuations and are as deeply rooted in experience as are criteria. The difference lies in the fact that objectives are criteria that have been arrested and formulated in such a way as to allow them to function as ends-in-view. The conception that evaluation is the process of measuring the extent to which prespecified goals have been attained[36] is therefore a natural one. It is simply the criteria-objectives relationship come full circle.

Objectives as ends-in-view are the links between past experience and future action. Actions which are still in the future may eventually increase the pool of experience. Objectives should not therefore be viewed as fixed but as tentative and fluid. Dewey considered it a "sign of immaturity when an individual fails to view his end as also a moving condition of further consequences, thereby treating it as *final* in the sense in which 'final' signifies that the course of events has come to a complete stop."[37] It is a bold step to claim that all of the potentially good aspects of an educational program will have already been identified and encapsulated in the objectives. A thoroughgoing goal-based evaluation leads to premature foreclosure of the set of criteria, which is one reason why such writers as Scriven[38] and Stake[39] have objected to it.

Tentativeness of objectives also makes it possible to evaluate objectives as *means.* Because the significance and value of a goal becomes clear only at the end, an objective can be evaluated in terms of whether it leads to a revised statement of intent which is better formulated (more realistic, more feasible, clearer, or more comprehensive) than the original. Furthermore, it almost never happens that "only the specially selected 'end' held in view will actually be brought into existence by the means used, something miraculously intervening to prevent the means employed from having their other usual effects."[40] It follows that objectives may also be evaluated according to the degree to which they guide actions which lead to valued (expected or unexpected) outcomes.

---

35. Ernest R. House, "The Objectivity, Fairness, and Justice of Federal Evaluation Policy as Reflected in the Follow-Through Evaluation," *Educational Evaluation and Policy Analysis* 1, no. 1 (1979): 28-42.

36. Ralph W. Tyler, *Basic Principles of Curriculum and Instruction* (Chicago: University of Chicago Press, 1949); and Malcolm Provus, *Discrepancy Evaluation* (Berkeley, Calif.: McCutchan, 1971).

37. Dewey, *Theory of Valuation,* 44.

38. Scriven, "Prose and Cons" and "Goal-free Evaluation."

39. Robert E. Stake, *Evaluating Educational Programmes: The Need and the Response* (Paris: Centre for Educational Research and Innovation, OECD, 1976); and Stake, "Program Evaluation, Particularly Responsive Evaluation," in *Rethinking Educational Research,* ed. W. B. Dockrell and D. Hamilton (London: Hodder and Stoughton, 1980), 72-87.

40. Dewey, *Theory of Valuation,* 42.

This approach to the evaluation of objectives is in contrast with Scriven and Roth's[41] needs-assessment (in which the objectives are compared with the demonstrable needs of the consumer) and with Stake's[42] proposal for evaluating objectives and priorities according to the values held by all of the parties who stand to benefit.

Needs assessments, the concerns of participants, and statements of objectives are important starting points. Without implicit criteria to guide proceedings, it is impossible to collect any relevant data because relevance would be undefined. That is why the suggestion that evaluators should collect data and values separately and somehow put the two together to make an evaluation[43] is misleading unless *values* is understood to mean not criteria, but standards or expectations.

The argument here is not that no criteria should be specified in advance. There must be some basic concerns in order for an evaluation to begin to get under way. Some criteria (especially those to do with literacy, numeracy, and principles of democracy) are so demonstrably fundamental to the idea of formal education that they constitute the very reason for the existence of schools. Schooling for the masses sprang from dissatisfaction with the nonexistence of provisions for the common people. Some criteria can therefore be safely taken as fixed; there is no need continually to rehearse the arguments for them.[44] However, it would be a mistake to *limit* inquiry to those dimensions specified in advance. There always remains the possibility of alternative conceptions of what is to count as good in education. A new quality may be discovered or conceived, or an old criterion may turn out to be irrelevant. The possibility of surprise features always exists, and the incidence of these is likely to be much higher in evaluating educational programs than in judging the quality of a piece of hardware such as an audio amplifier. The point is that the parties who have an interest in the evaluation need to be aware that the set of criteria needs to be left somewhat open, so that new criteria may on occasion be negotiated into an evaluation or inappropriate ones negotiated out.

CONNOISSEURSHIP

Like all other abstract concepts, criteria are mental constructions which cannot be experienced or enjoyed directly. It is therefore necessary to consider the reasonableness of demands that all evaluations be rational. Two facts bear upon this issue. The first has already been discussed: global evaluations can be, and often are, made in the absence of formal criteria because valuations are the primary acts in which criteria are grounded. The second is that the universe of human discourse is not coextensive with the universe of human experience. It follows that the two cannot be isomorphous. It is therefore not surprising that from time to time situations arise in which it is impossible to identify some or all of the criteria necessary to explain an evaluation fully. While people undoubtedly differ in their abilities to translate evaluative images into communicable form, the issue cannot be dismissed simply by requiring perfect powers of articulation. The lack of direct correspondence between verbalization and experience is a problem that besets all forms of empirical social inquiry.[45] The simple though heavy-handed solution is to ignore this aspect or deny its existence altogether and limit

41. Michael Scriven and Jane Roth, "Needs Assessments: Concepts and Practice," in *Exploring Purposes and Dimensions,* New Directions in Program Evaluation no. 1, ed. S. B. Anderson and C. D. Coles (San Francisco: Jossey-Bass, 1978).

42. Robert E. Stake, "Objectives, Priorities, and Other Judgment Data," *Review of Educational Research* 40 (1970): 181-212.

43. D. L. Stufflebeam, W. J. Foley, W. J. Gephart, E. G. Guba, R. L. Hammond, H. O. Merriman, and M. M. Provus, *Educational Evaluation and Decision Making,* Phi Delta Kappa National Study Committee on Evaluation (Itasca, Ill.: Peacock, 1971).

44. Philip G. Smith, "Evaluation and Practical Deliberations in Education," *Educational Theory* 28 (1978): 279-85.

45. Peter White, "Limitations on Verbal Reports of Internal Events: A Refutation of Nisbett and Wilson and of Benn," *Psychological Review* 87 (1980): 105-12.

investigation to those aspects which are observable and explicitly describable. An alternative approach is to recognize the subtlety of the problem and try to get at a partial solution indirectly.

Connoisseurship is usually associated with artistic forms of expression but is not such an esoteric activity as is often supposed. It is of wide significance, applying equally to livestock judges, radiologists interpreting X-ray plates, wine tasters, art critics, and teachers monitoring moment-by-moment exchanges in the classroom. It is an aspect of what Polanyi called *tacit knowledge,*[46] which everyone uses in the ordinary course of living. The important question in practical educational evaluation is: Under what circumstances is it reasonable to expect those who are not judges to accept an evaluation when clearly defined criteria cannot be produced? Consumers of evaluations, unless they are insiders who share the evaluative schemata of the connoisseur, have several options open to them: (a) they can reject the evaluation altogether, referring to such evaluations as "merely subjective"; (b) they can limit their acceptance of the evaluation to those dimensions associated only with formal, explicit criteria; (c) they can interpret the necessarily metaphorical language of the connoisseur in the light of their own experiences, on the assumption that they share with the evaluator sufficient experiences of an essentially similar nature for their interpretations to be valid; or (d) they can simply accept the authority and competence of the judge.

The third of these options is one of potential interest for educational evaluation. If the experience from other fields can be translated into an educational setting, the tasks of a connoisseur-critic include not only trying to convey a particular evaluation, but also helping people attune their expectations, suggesting ways in which a phenomenon might be approached. In other words, the critic's job is as much to provide a rich and personally coherent framework through which to perceive something as it is actually to make a valuation. As Najder puts it: "The more limited and vague the expectations, the more accidental the aesthetic experience, and the more poorly substantiated the evaluation."[47] To take a simple analogy, to judge a day's fishing by the catch rate or the total weight of fish caught is to assess by clear criteria. If, however, a friend says that fishing is "an opportunity to engage in unhurried serenity, in which to catch a fish is a bonus," the experience is approached, engaged in, and reflected upon in quite a different way, a way that almost certainly goes beyond the simple substitution of one set of criteria for another.

One of the key differences between educational evaluation and, say, film criticism is that the vast majority of educational transactions take place in private classrooms, so that one can never be sure whether the experiences "shared" are very much alike. Activities are sometimes accessible to researchers, less commonly to other teachers, and hardly at all to the public. There is little opportunity for public appreciation or adjudication, which is the essence of what is meant by objectivity.[48] Similarly, decision makers may find themselves making heavy use of their personal experiences and recollections in interpreting evaluations. This is why in many evaluations, discovering the characteristics of potential audiences is part of the evaluation problem.[49] The twin concepts of *connoisseurship* (appreciation) and *criticism* (disclosure) as approaches to educational evaluation are currently being developed and refined by Eisner,[50] Willis,[51]

---

46. Michael Polanyi, *Personal Knowledge: Towards a Post-critical Philosophy* (London: Routledge & Kegan Paul, 1962).

47. Najder, *Values,* 149.

48. Pagano and Dolan, "Foundations for a Unified Approach."

49. Michael Quinn Patton, *Utilization-focused Evaluation* (Beverly Hills, Calif.: Sage, 1978).

50. Elliot W. Eisner, *The Educational Imagination: On the Design and Evaluation of School Programs* (New York: Macmillan, 1979); and Eisner, "The Use of Qualitative Forms of Evaluation for Improving Educational Practice," *Educational Evaluation and Policy Analysis* 1, no. 6 (1979): 11-19.

51. George Willis, "Curriculum Criticism and Literary Criticism," in *Qualitative Evaluation,* ed. G. Willis (Berkeley, Calif.: McCutchan, 1978), 93-111.

Kelly,[52] Gray,[53] and others. As yet there is no long tradition of connoisseurship in educational evaluation such as exists with film criticism.

Connoisseurship, or judgment by experts, is also important when the criteria *can* be defined. Many aspects of education, including teaching itself, are rarely so uniformly expressed that one set of criteria is appropriate for each and every case. Although experts may not be in complete agreement with one another, their judgments are often tolerably reliable and can be distinguished from lay judgments because the experts know the rules for applying different criteria to different cases and because they "share with their peers a theory or theories as to why the rules are applicable."[54]

METACRITERIA

Given a fixed set of criteria and a wide variety of phenomena, situations arise in which a particular criterion may become irrelevant because it is subject to another criterion of a different order. There are criteria for using criteria, just as there are rules for breaking rules, and rules for breaking the rules for breaking rules, and so on. Such criteria are latent and not part of the set that is ordinarily appropriate and sufficient. Metacriteria are neither necessary, nor do they become operational, until a concrete situation arises. Thus the criteria that are routinely used in appraising a painting become irrelevant if and when the work is discovered to be a forgery. At that point, authenticity becomes the dominant consideration. Similarly, a morals charge against a teacher or the embezzlement of funds from a project makes the usual criteria of no consequence.

Metacriteria are not simply alternative criteria that one person accepts and another rejects. They are principles (of justice, appropriateness, and so on) which demand or justify the suspension of some criterion (even a well-intentioned or principal one) in favor of another. Typical situations in which metacriteria become important are (a) when an educational program is conceived without adequate knowledge of practical constraints or preexisting conditions; (b) when an educational program, with clear and definite expectations, is translated from one setting to another where it does not sit comfortably; and (c) when shifts in social conditions or cultural values result in a program's becoming seriously out of date. An example of the first situation is to be found in the goal of bicultural education, which, while appropriate for mainland communities with large Latin-American populations, is, according to Thomas,[55] an unreasonable goal for American Samoans because it fails to take account of the talents and cultural backgrounds of the learners and the capability of the school system. The result, he claims, has been the superposition of an educational system which is incongruent with Samoan culture and is therefore doing American Samoans a grave disservice.

CONCLUSION

Specifying criteria is a complex but not intractable problem and has been portrayed in this article as comprising three phases: valuation, criterion identification, and prescription. Valuations arise from the reaction between axiological values such as freedom and individual worth and the existential character of a phenomenon. Because axiological values exist at a high level of abstraction, they are indeterminate to the extent that their content is not made explicit. Criteria are useful for making, and necessary for substantiating, value claims because they determine existentially the content of the axiological values by clarifying and interpreting them.

---

52. Edward F. Kelly, "Curriculum Criticism and Literary Criticism: Comments on the Analogy," in *Qualitative Evaluation,* 114-38.

53. James U. Gray, "Vintage Connoisseurship: A Practitioner's View of Educational Criticism," *Curriculum Inquiry* 11 (1981): 343-58.

54. Harry S. Broudy, *The Continuing Search for Criteria,* Evaluative Criteria Reference Paper no. 3 (Washington: American Association of Colleges for Teacher Education, 1967).

55. R. Murray Thomas, "Evaluation Consequences of Unreasonable Goals: The Plight of Education in American Samoa," *Educational Evaluation and Policy Analysis* 3, no. 2 (1981): 41-49.

Criteria, when they are expressed as objectives, function as ends-in-view and so serve as foci for action. But any evaluation that restricts itself to comparing achievements with goals stems the potential flow of new or alternative criteria that could originate in more holistic evaluations and would be accepted as legitimate if identified. This means that to specify in advance a set of criteria by which something is to be judged is self-limiting. While any set of criteria for educational evaluation should in general be made explicit, complete prior specification is undesirable in principle. The case for this assertion rests on four grounds: (a) decompositions of value principles are always partial; (b) criteria emerge from experience and cannot be deduced from value principles; (c) not all criteria *can* be made explicit; and (d) there are criteria for using criteria.
