Evaluation and the Improvement of Academic Learning

D. Royce Sadler

The Journal of Higher Education, Vol. 54, No. 1 (Jan. - Feb., 1983), 60-79.

Stable URL:
http://links.jstor.org/sici?sici=0022-1546%28198301%2F02%2954%3A1%3C60%3AEATIOA%3E2.0.CO%3B2-L

The Journal of Higher Education is currently published by Ohio State University Press.

D. Royce Sadler

Evaluation and the Improvement of Academic Learning

Introduction

Evaluation is a dominating aspect of educational practice. It strongly influences what students attend to, how hard they work, how they allocate their study time, and what they can afford to get interested in [8]. To some students, evaluation is puzzling and contradictory; to some faculty, it is difficult and burdensome. In this article the role of evaluation in *improving* academic performance is explored. The focus is on formative rather than summative evaluation, on growth rather than grading. This is an issue that is currently receiving some attention in the literature [4, 5].

The evidence upon which this essay is based comes from informal investigations with my own classes into self-evaluation as a workable concept and into the utility of providing students with criteria, exemplars, and opportunities for reworking. I have also drawn on interviews with students and discussions with faculty. The evidence is fairly extensive, but would not pass the tests of experimental rigor. In arguing for a different theoretical perspective from which to view academic learning, I intend the treatment to be tentative, but not speculative.

For the purposes of this article, the broad goal of academic learning is not assumed to be the mere accumulation of factual knowledge, passing grades, or course credits, but the pursuit of excellence in a discipline. In order to make the discussion manageable, it is necessary to restrict the scope of the term "academic learning." I shall use it to refer to those types of scholarship commonly found in the arts, humanities, and social sciences in which the aim is to develop higher cognitive skills (such as

D. Royce Sadler is lecturer in education, University of Queensland, Australia.

0022-1546/83/0183-0060$01.00/0
Journal of Higher Education, 1983, Vol. 54, No. 1

synthesis and critical thinking) through the manipulation of theories, ideas, and facts. The complex patterns of thinking and doing are assumed to lead to a tangible product (such as a written paper) which is open to inspection and for which multiple criteria are used in assessing quality. Although there is some overlap with the development of scientific knowledge and problem-solving skills, and with certain aspects of training for the professions, these are not the primary targets.

The article itself is divided into two parts. In the first part, a theoretical development of formative evaluation in higher education is presented. The second is a critique of certain aspects of higher education in the light of that theory. It is shorter than the first part because the grounds for criticism which are implicit in the theory are not restated.

Part 1: An Approach to Formative Evaluation

Few would want to argue with the proposition that academic learning requires complex strategies to achieve complex ends. In order to analyze the role of evaluation in the process, I shall first try to clarify what I mean by complex means and ends, and to show that one does not necessarily imply the other. The distinction made here between means and ends is convenient in the analysis which follows, but such a classification is admittedly arbitrary: a logical argument (end) is presumably the result of logical and not random or muddled thinking (means), but may itself be instrumental (means) in the development of a larger thesis (end). The distinction between simple and complex is also one of convenience, and is not meant to disparage any forms of learning, to suggest that some forms are superior to others, or to equate *simple* with easy or trivial.

An end or goal is defined as *simple* if the criteria for success, quality, or excellence are few in number and easy to state and understand. A *complex* goal has multiple interlocking criteria, including some that are highly abstract in nature; this may make them difficult, or even impossible, to specify. A considerable proportion of part 1 consists of an elaboration on the nature of complex goals.

A strategy (or sequence of moves) is defined as simple if it is composed of a number of straightforward steps; in principle, it would not be impossible to construct a robot or program a computer to carry out the operations. Complex strategies involve multiple decision points because alternative courses of action are possible. This may be because the context is dynamic (and the strategies therefore need to be adaptive) or because there are many courses of action that lead to acceptable if not identical outcomes.

Simple means and simple ends often go together; so do complex means and complex ends. However, exceptions are possible. Riding a bicycle is an example of a simple end (moving forward on the straight, round corners, up hill, down hill, and on the level, all the while maintaining balance) that requires complex means. This is so in spite of the fact that most children learn the delicate art of balancing fairly easily through experience, without much conscious thought. Ironically, adults sometimes find it more difficult because there are one or two "recovery" actions (knowing which way to turn the wheel to regain balance or to minimize a skid) that seem to go against adult intuitions.

It is also possible to achieve a complex end through simple strategies. To the extent that the steps in making wine are standardized (not varying much from year to year), winemaking is an example of a simple process designed to attain a complex end (a superior vintage, as judged by an experienced palate).

If this so far seems to be only laboring the obvious, two points should be noted. First, while many teachers acknowledge the complexities of production, they frequently underestimate the complexities of assessment. That is, they assess as though the ends were simple, by supplying only grades or brief comments to their students. To expect improvement as a result is, I shall argue, to make a fundamental miscalculation.

Second, different disciplines have different mixes of the simple and the complex (as I have defined the terms), and these differences are not adequately reflected in course structures, teaching techniques, timetabling, or assessment practices. Let us take two examples. Suppose that a simple regression equation has to be fitted to data using least squares. The strategy is simple because the steps involved are clear and plain; so is the end. The whole process can be readily routinized. Even though the manner in which the result is interpreted involves considerable statistical understanding if it is to be done properly, it is fairly standardized. Assessment takes place in a context of objectivity because there *are* correct and incorrect procedures and interpretations.

For the second example, let us suppose that in optimizing a certain industrial process, high levels of insight and highly sophisticated analytical tools are required. Let us also suppose that there are well-codified procedures for testing whether the optimum has been attained, and for investigating the sensitivity of the optimum to changes in the input parameters. The teaming of either a simple strategy (regression equation calculation) or a complex strategy (search for an optimum) with a simple end implies that much of the educational effort should, in these cases, be directed towards the development of skills and strategies.

By contrast, academic learning in the arts and humanities is directed towards complex ends and needs to face more squarely the key role of evaluation in the development of expertise. Good evaluation is not an adjunct to good teaching: it *is* good teaching.

Academic Learning as a Process

"Intelligent" academic learning occurs when a person knows what is to be achieved, works towards ways of doing it, and can tell when progress is being made. There are therefore three components: the "directional" (attending to goals), the "algorithmic" (devising strategies and making moves), and the "evaluative" (monitoring the discrepancy between current status and the desired end). It is taken as axiomatic that conscious attention to goals regulates strategies and therefore achievement. There are no formulas or algorithms that work equally well for all students and there are no fixed criteria that can be used for objective assessments of quality.

Even though the goals in academic learning may initially be vague, they are still capable of providing sufficient motivation and direction for the first few steps. Under ideal conditions, a "knowledge of excellence" develops progressively. It comes about through a series of cognitive tacking maneuvers among productive acts, outcomes, criteria, standards, and external judgments of quality. If conditions are poor, apprehension of the goal does not unfold but remains static. This occurs when there is too much reliance on external judgments from the teacher, and too little emphasis on conscious reflection by the students themselves on what has been done and where they are heading.

If it is true that a person can concentrate on only one thing at a time, it must be possible for attention to switch freely among the directional, algorithmic, and evaluative aspects. The proportions of time devoted to each depend on the nature of the task, certain personal characteristics of the learner, and the stage of development. Once expertise is acquired, the need for conscious attention to goals and processes never entirely disappears, because the types of skills being considered here cannot be maintained merely through mechanical practice.

Assuming that students are able to adopt the necessary critical detachment from their own work, self-monitoring is possible only when students know (1) what a good performance is, (2) their own status or level of performance, and (3) how to compare the two. Current practice too often focuses on condition (2); while necessary, it is not sufficient. Students need to know not only *that* they have achieved, but *how* and *why* as

well. If intelligent learning is to take place, this state of affairs implies a dual role for the teacher: helping the student develop a concept of excellence, and helping the student develop skills and strategies to achieve it.

An Improvement Model

The dynamic connection proposed in this article between the developing concept of excellence and actual performance can be represented diagrammatically. The principal elements of the model are two interdependent curves, both of which steadily increase. The first, which I shall call the *unfolding curve,* represents the development of the goal over time, where development is a broad term taken to include (1) clarity or improved personal knowledge on the part of the learner as to what constitutes quality (i.e., knowledge of criteria and standards), and (2) aspiration for that ideal. (In practice, of course, we know that many students set their sights considerably lower). The second or *performance* curve represents level of performance, achievement, proficiency, or expertise and can be roughly identified with classical learning curves. The unfolding curve represents the end-in-view; the second, competence.

The first few steps are illustrated in Figure 1(a); *P* represents the learner's initial concept of what is good. Although aspiring to *P,* something short of it, *Q,* is actually achieved. Assuming an ideal evaluative environment, the concept of excellence becomes clearer (*R* in the fig.), aspiration is redefined, and the next performance results in an improvement (*S* in the fig.). As further attempts are made, the gap between the desired end and the quality of performance decreases, but may not disappear altogether. At the same time, knowledge about the nature of the desired end improves, and with it the ability to make better evaluative judgments. Figure 1(b) shows a smooth hypothetical long-term progression.

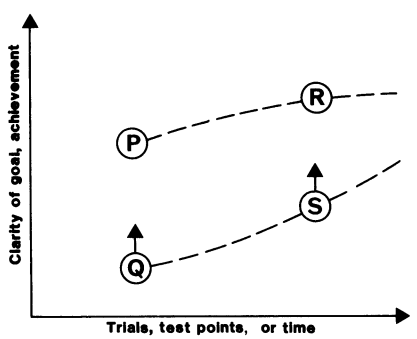

FIG. 1(a) Improvement model: first steps

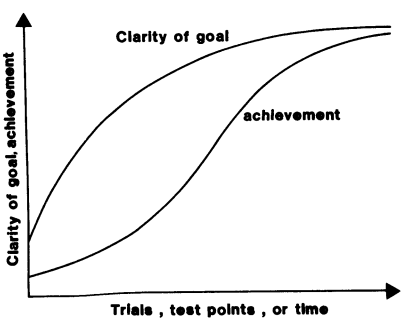

FIG. 1(b) Improvement model: idealized development

FIG. 1. Model of the Relationship between Academic Aspiration and Achievement

Although this model attempts to (1) relate achievement directly to both previous performance and knowledge of the goal, and (2) represent refinement, resetting, or clarification of the end-in-view as a result of partial achievement of the previous goal, it is obviously an idealization. In the normal course of events, deviations from the model must be expected to occur. Except for tasks where skill develops through practice alone, no learning proceeds at a uniform or precisely specifiable rate. Creative experimentation with strategies results in the occasional accidental success, the occasional unexpected failure. Sometimes there are sudden insights into the nature of a complex end and abrupt leaps in performance. There are accommodations, compromises, extensions, and extrapolations, some conscious, some not, all along the way. Some lead forward, others backward, but provided there remains determination to pursue, useful knowledge can result even from backward steps.

Evaluative Criteria

Judgments about quality are defended by referring characteristics of the object to criteria and standards external to the object. Some of the criteria used in various forms of academic learning are:

coherence	support for assertions
length	logical reasoning
organization	persuasiveness
integration	objectivity
originality	comprehensiveness
wording	referencing

There are, of course, many others, but this list serves for purposes of discussion. Some of these criteria overlap conceptually (support for assertions, persuasiveness); others are essentially independent (logical reasoning, correct referencing). In order to understand how criteria are used in assessing quality, it is necessary to explore their origins, nature, and relationships.

Identification of Criteria

The fundamental question which needs to be considered at this point is: what are the origins of criteria? Do they flow from evaluations already made? Or are evaluations made by rigidly applying criteria? If valuations are possible only when the criteria to be used are explicit, how do criteria arise in the first instance?

Criteria have their roots in experience, that is, in previous valuations. Valuation, or the recognition of merit (beauty, justice, and so on) is itself a primary act. Even where "excellence," "quality," and "success"

may be difficult to define, it is common knowledge that they may often be recognized. After global valuations have been made, it is often possible to reflect on them in such a way as to identify the operative principles or common elements (the criteria) that seem to "explain" the valuation. (Of course, one tries to be sure that the apparent criteria would stand up in future judgments.) This "emergence" principle seems to apply regardless of whether the evaluations are essentially aesthetic, logical, or technical. Once criteria are identified, they serve to (1) provide a rationale for the current judgment so that others may understand the reasons, (2) foreshadow future valuations, making them more, but not absolutely, predictable, and (3) invite others to judge similarly. However, when validity is established by common acceptance, criteria are often used quite routinely.

Identification of criteria is not always easy, and there are situations when it is extraordinarily difficult to identify why something is judged excellent. For example, students and faculty may be unanimous in their selection of the outstanding teacher of the year, but be unable to explain exactly what it is that makes the difference between a good and an outstanding teacher. Pressed, they may even advance quite different reasons, or fall back on something like "charisma," a comfortably indefinite characteristic.

The emergence principle suggests that, rather than prespecifying criteria for all tasks for all subjects using a standard list such as the one above (and then perhaps not actually using them because they turn out not to be all relevant), a teacher might be better identifying and recording the criteria that *have* been used immediately *after* evaluating performances of a given type, while the experience is still fresh. This would make it possible for those criteria whose relevance and salience are obvious to be made explicit before students begin work on another task of the same type.

Is it possible to specify in advance all of the criteria that should be used in assessing the quality of academic work? If the emergence principle is accepted, the answer has to be negative. No listing of criteria can be exhaustive, because there is always the possibility that more could, or should, be invoked or that new ones will emerge. However, lack of comprehensive specification need not inhibit action. People frequently engage in activity with only vague or general expectations of whether the outcome is likely to be good or bad, or what shall make it so. After the event, they may be able to judge its worth and give fairly precise reasons.

In academic learning, it is not absolutely essential that teacher and student share identical conceptions of quality, but it is reasonable to

expect general conformity, otherwise the roles of teacher and learner make no sense. However, specific variation is not only inevitable, it is desirable. It authorizes originality and creativity, and leaves open the possibility that the performance of the learner may be better than the teacher's.

Issues in the Use of Criteria

Using criteria and standards is complicated for several reasons. First, there is the sheer number of them which could be employed. It is clear that even the number of criteria in the list given earlier is more than one can keep in mind at once. They are intuitively difficult to apply for this reason alone, during either production or assessment phases. The more criteria there are which have to be attended to simultaneously, the more difficult evaluation is, especially if the criteria cannot be readily reduced to a simple checklist.

Second, judges differ in the ways they employ criteria. Even if it were desirable to use a fixed set of criteria, different assessors have different "evaluation policies." Some teachers use them disjunctively: to be acceptable, a performance must achieve minimum qualifying levels on a number of essential criteria. Others use them conjunctively: a paper that is outstanding in one characteristic has its failings in all other areas overlooked. Yet others use a simple compensatory rule: poor showings on some characteristics are balanced by high levels on some others. So far, the evaluation policies of academics seem to have been little researched.

It might seem possible, in principle, to define a "weighting function" to indicate how much each characteristic counts and the nature of the compensations possible. This is certainly the intent of analytical scoring schemes. However, this philosophy assumes that all the important criteria can be specified in advance. It also assumes that the weighting coefficients remain fixed over *all* levels of the criteria, a possibly naïve point of view. The weighting associated with a particular criterion may need to depend on the level of the characteristic present, or on the combined effects of the levels on two or more specific characteristics. A further serious objection to the "weighting function" idea is that it implicitly supports the notion that evaluation is primarily concerned with getting an overall assessment of worth, that is, with grading.

Third, characteristics may be conceptually distinct, but always occur in combination. That is, they might be "structurally" dependent. "Support for assertions" and "persuasiveness" might be distinguishable ideas; one could claim that an argument may have all of its assertions supported and still not be persuasive. However if, *in practice,* the two are nearly always

associated, it may be unnecessary to treat them as distinct, especially if some weighting function is used or implied. Here again is an area in which research could provide some answers.

Fourth, there are situations when rules, even the best ones, are better broken. Metacriteria (for using criteria) are situationally dependent, being neither necessary nor operational until a concrete situation arises. For example, suppose there is a certain logical order in which an argument might be developed. Depending on the purpose in making the argument, it is conceivable that the most *persuasive* development might not follow the most logical order. That is, in some circumstances persuasiveness might take precedence and justify a temporary suspension of step-by-step logic. The circumstances under which it is necessary to invoke an overriding criterion are part of an irreducible residue of judgment that attends all valuations.

Finally, some criteria simply defy expression, even though they are part of the tacit understandings shared by experts. (One could argue that criteria are not criteria at all until they are articulated. I shall ignore this quibble.) It is possible for different people to be in possession of the same knowledge but to be unable to express that knowledge in words [6]. It can, however, be shared at least in part through common experiences. The connoisseur, for example, makes valuations using a complex mix of the specifiable and the unspecifiable. The unspecificable components are developed through prolonged evaluative experience under the guidance of experts. Where there are subtle, indefinable criteria in the assessment of academic performance, it is all the more necessary for students to engage in evaluative activity themselves. Unless learners are insiders who share the evaluative schemata of the connoisseur-teacher, they are mere consumers of evaluations and have no alternative but to rely on the authority and competence of the judge.

Classes of Criteria

Leaving to one side the possibility that some criteria may be unspecifiable (the point touched upon above), several classes of criteria may be usefully distinguished: regulative, logical, prescriptive, and constitutive. In suggesting this classification, I have drawn on some of the ideas in [7].

Regulative criteria. In order to achieve a degree of uniformity in presentation, rules are made to govern such aspects as length, layout, structure, and conventions for language and spelling. Some rules are simply accepted practice (e.g., in punctuation) expressed in codified form. Others cover points of presentation where wide variations are possible (as in

footnoting or citing published works), and are based on essentially arbitrary decisions. Publishers' manuals of style are compilations of regulative criteria.

Because the rules themselves are quite straightforward, it is generally easy to tell when a particular rule has been broken or a requirement not met.

Logical criteria. Logical criteria have to do with valid chains of reasoning. They cut across most of our disciplinary boundaries, and while they may not be strictly universals, they appear quite fundamental to the way we think. The broad cultural or research framework within which they exist (in our case, Western science and philosophy) automatically accords value to certain types of outcomes which, when the logic is followed, we label "successes": equations are *solved,* theorems are *proved,* compounds are *synthesized,* poisons are *identified,* and conclusions are *reached.* In this context, it makes sense to speak of valid deductions, wrong assumptions, correct inferences, and logical consequences. Faulty reasoning is no less faulty even when it happens to lead to a correct conclusion.

Embedded as they are in our thinking patterns and our notions of what constitutes knowledge, logical criteria often lead to statements that appear to be purely descriptive, rather than evaluative, because value inheres in the unambiguously true or correct.

Furthermore, the conclusions which are arrived at are replicable; they can be confirmed (or disconfirmed) at any time, at any place, by any competent person capable of understanding the meanings of terms in the statements describing them. Where the "success" is stated in the form of a generalization, particular instances may be generated at will by following well-established rules. Judgments as to adequacy, and therefore to quality, are made by appeal to precision and the rules of logic, never to conventions, popular opinion, or personal taste. Hence, logical criteria do not readily admit matters of degree: it is all or nothing.

Prescriptive criteria. Many of the criteria used in assessing quality in the arts and humanities are *prescriptions,* normative statements that something is to be valued when the something is neither an empirical (wholly describable, replicable) nor an institutional *fact.* The feature or characteristic is existential, embodied in a particular instance. If an abstract principle can be found that links features of like nature together, the principle can be given a label such as "coherence," "originality," or "readability." This principle or *criterion* can be identified only through experience with particular cases, that is, with existential facts because

when attempts are made to define prescriptive criteria, the definitions tend to be circular. A prescriptive criterion may be useful in recognizing instances, but it cannot be used to generate them. Standards are defined in a similar way, that is, by rank ordering performances or instances according to the degree of the characteristic present.

Constitutive criteria. These serve to define, and are characteristic of, a discipline. They have their basis in consensus among members of a guild on standard categories, concepts, and methodologies. To take history as an example, constitutive rules distinguish what is, for example, "history proper" from what is straight narrative or a contemporary record of "facts." They specify historical meanings for evidence, explanation, objectivity, verification, and interpretation. They govern how sources of data should be used, and how disagreements are to be assessed. In spite of the fact that there are differences of opinion among philosophers of history, and granted that the constitutive rules may evolve over time, the fact remains that history is identifiable as history (at any given time), and is distinguishable from philosophy, political science, or literary criticism.

Regulative and logical criteria have this in common: they both refer to empirical facts of the performance. Errors, mistakes, or noncompliance are perfectly detectable, at least in principle. Standards can be defined in terms of well-defined outcomes: it is clear to a mathematics instructor, for example, what solutions to second-order differential equations with constant coefficients look like.

On the other hand, prescriptive and constitutive criteria are matters of degree. The judgments are subjective, and rely greatly upon whether the assessor is persuaded or convinced. We try to answer questions like these: Is this analysis penetrating enough? Is that position adequately defended? Is appropriate support given for this statement? Does that treatment constitute a philosophical approach?

With respect to standards, given a collection of prescriptive and constitutive criteria, it would be difficult or impossible to guess the educational level at which they are applicable (freshman or graduate, for instance) because the standards depend on existential facts, that is, on concrete examples. Quality is assessed along a continuum, not in terms of the number of correct performances of a given type. It follows (1) that it is impossible to engage in evaluation using these classes of criteria without experience of particulars, and (2) that one may understand which criteria (in the abstract) are appropriate for an appraisal without knowing much about levels of competence or standards of performance.

Three Principles for Effective Formative Evaluation

Students often perceive evaluation as coercive and threatening. It ought to be the opposite. For evaluation to be congenial, the assessment of merit ought to take place in an open environment where the teacher understands the difficulty the novice faces in becoming expert, and where the student appreciates the undesirability or even the impossibility of complete prior specification. Good evaluation weakens the image of the teacher as the unquestionable authority, and stands in direct contrast to conditioning. It recognizes the importance of negotiated criteria, of autonomous and creative discourse among teachers and learners. I propose three commonsense principles for better formative evaluation.

The Communication Principle

There are both moral and pedagogical obligations to state beforehand the criteria by which a performance is going to be assessed. Some criteria are a direct outgrowth of the task definition, and how well the task is specified is obviously of vital importance. Because principles of item construction receive thorough treatment in the literature on evaluating student achievement, this aspect is only briefly mentioned here. However, the fact that many teachers complain that students do not address themselves to the task *as it is set* points to a breakdown in communication. Either the task specifications do not adequately convey the teacher's expectations, or they do and the students do not know what to do with them.

Explicit criteria fulfill a product-design function. They are critically important when there is only limited sharing of expectations between learner and teacher. This is typically earlier rather than later in a course of study. Constitutive criteria, where they apply, should receive high priority. There is no case for delaying their introduction until an advanced-level course in the philosophy of the discipline.

Ideally, there should be considerable overlap in the criteria across courses in the same discipline at the same level, even with different teachers. Furthermore, the criteria should be nested, with complete upward compatibility. As a student proceeds from introductory through advanced courses in the same discipline, no criterion becomes irrelevant, but there is instead a progression towards more sophisticated criteria and higher standards on existing criteria.

If teachers are to accept the tasks of fostering expertise and promoting a concept of excellence (without which self-evaluation is impossible), it is usually necessary to work with a set of explicit criteria. But it is not

sufficient: evaluative experiences shared between novice and expert are essential if a repertoire of existential facts is to be built up. Many of our abstract concepts are developed through exposure to positive and negative instances. Similarly, experience in criticism is fertile ground for identifying both a new criterion and compliance with or violation of an existing criterion.

There are at least two ways to facilitate this. First, students should have access to *exemplars,* complete with criticism. Both good and bad examples should be available. Second, opportunity should be given for students to engage in evaluation of works other than their own. If the improvement model presented earlier in this article has any validity, it should be possible to detect whether the concept of excellence is developing in the learners. Students will be able to make sound appraisals of works of the same genre they themselves are trying to produce, and there will be fewer genuine surprises when students' performances are independently assessed.

The Progression Principle

At this point, we face something of a dilemma. On the one hand, there is the need to prespecify criteria so that students can work towards reasonably clear goals. On the other hand, as connoisseur-experts we want to retain the right to perhaps use evaluative criteria that are not part of the standard list, should occasion demand. A working solution can be arrived at by defining a system of rolling sets of criteria, an open and beneficial approach. First, a distinction is made between latent and manifest criteria.

Criteria that are operational either before a work is produced or while it is being assessed I shall call *manifest* criteria; those criteria waiting in the wings I shall call *latent.* The need for manifest criteria is obvious, but to try to set out all the criteria that could possibly be used would be absurd. The number made explicit at any one time simply has to be limited to be manageable.

Latent criteria are either understood and taken for granted by both teachers and learners, or are part of the teacher's repertoire alone. They emerge as and when the occasion demands because breaches of unwritten rules need to be corrected and fortuitously successful trials need to be capitalized on. When an unexpected feature stands out and sets a particular performance apart, this datum should not be ignored in the assessment simply because it was not anticipated. The translation of a criterion from latent to manifest should not be viewed as unfair, but as inevitable and perfectly normal. (Additionally, there is the remote possibility that a

dimension could arise that is wholly new to both teacher and learner. I shall ignore this in subsequent discussion.)

Latent criteria become operational by passing over some cognitive detection threshold. In other words, some feature has to be exceptional or unexpected for it to be consciously noticed by the assessor, and the deviation may be either on the side of too much (overqualification) or not enough (disqualification). Overqualification and disqualification are often used intuitively when the manifest criteria are few in number. This situation may legitimately occur when the performance being assessed is only tenuously related to other works, either because the genre to which it belongs is not itself well defined, or because its membership in a well-defined genre is in doubt. Criteria and standards from other genres are then not directly transferable. When an assessment *has* to be made (as in examining a highly original thesis or refereeing an unorthodox journal article), previous experience may be an inadequate guide. A positive appraisal may result by default because no disqualifying evidence can be found. Reliance on latent criteria should not be abused. It is no substitute for hard thinking, or a working set of explicit criteria.

The art of evaluation in teaching for improvement is to generate an efficient and partly reversible progression in which criteria are translated from latent to manifest to latent. The aim is to work towards ultimate submergence of criteria once they are so obviously taken for granted that they need no longer be stated explicitly.

The Iteration Principle

The requirements of formative and summative evaluation often conflict. In particular, while the practice of "counting" everything a student submits towards a course grade improves the sampling, it is too often at the cost of improvement.

One of the puzzling aspects about learning in higher education is the apparent unwillingness or inability of students to take account of criticism in subsequent work. When feedback is given, it is often ineffective as an agent for improvement. Students seem to show the same weaknesses again and again. Feedback does not apparently transfer from one task to the next. There are many possible explanations. The truth is that we do not yet know enough about the ways learners come to understand connections between criteria and the objects being evaluated, and how they can best use criteria in improving their own work. But here is one hypothesis.

Suppose that a student has been given written criticism about a performance. Even though both the existential fact (i.e., what the student wrote) and the criticism may be laid before the student together, the

meaning and significance of the feedback may not become apparent until the student attempts to repair the defect. (It is assumed here that the flaw is due to the student's lack of appreciation of what the criterion implies, and not to a simple oversight.) Without this, the student is in possession of a negative instance and a negative appraisal, but has no corresponding positive instance. The connection is made when the student successfully constructs a positive instance. A simple analogy might help: to be in possession of a red object to which is attached the label "not green" supports the idea of greenness but cannot define it; one needs a green object as well.

An obvious implication is that for improvement to occur, students should be given opportunity and incentive to rework and resubmit papers, with continuous rather than single-shot access to evaluative feedback during the reworking. Clarification of criteria and standards through recycling enables but does not, of course, guarantee transfer to the next task. However, without recycling, prospects for clarification are slim.

There is nothing new or radical in a suggestion to set up learning "loops." In many trades, professions, and crafts, the master-apprentice model of learning is still important. It occurs in higher education when graduate students reach the dissertation stage. Performances or trials, feedback, reworking, and a knowledge of excellence are inseparably bound and proceed quite naturally and unselfconsciously. The catch in undergraduate learning is that there is often some understandable reluctance on the part of students to rework something over which they have labored hard and long, and in which they have considerable emotional investment. However, because learning loops provide a retrieval situation, it is not unreasonable in the long term to expect less confusion, improved teacher-learner relations, greater intrinsic motivation, better achievement, and a heightened sense of personal satisfaction.

Part 2: A Critique of Some Current Practices

The dominant paradigm for teaching and learning, and for research into these processes, is based on the sequence: test, response, and feedback, where feedback is narrowly interpreted as knowledge of results or outcome. While this is adequate in situations where the response itself is what has to be learned (such as in learning the multiplication table by rote), it is deficient as a theory of evaluation for complex learning. Ignoring other modes of learning (such as learning by being told), it overemphasizes the importance of feedback and disregards the complex

interactions between learner expectations, the activity itself, and the criteria which determine quality. As Bjorkman put it, "The classical paradigm . . . has been ingenuously transferred to the study of cognitive learning where it is less adequate and too narrow" [1].

Grading

Evaluation and grading have often been referred to in the literature as if they were synonymous, and this confusion persists even in recent writings [3]. While normative grading has a role to play in certification, accountability, and prediction, grades are action-neutral for the purposes of improvement. They do nothing to help students understand the connections between actions and the criteria used to appraise the results, or to participate in intelligent control over the learning. These latter form the essence of formative evaluation.

In fact, some recommended grading techniques depend upon assumptions which are quite inappropriate for formative evaluation. As an example, consider the global scoring procedure often used in assessing writing [2]. It is common knowledge that grading is notoriously unreliable in higher education. That is, different assessors assign (sometimes wildly) different grades to the same piece of work. Since it is obviously desirable to achieve some inter-assessor consistency, special procedures have to be employed to "calibrate" the judges. In global scoring, the student work is sorted into an order of merit by general impression, without appeal to explicit criteria at all. Reliability is obtained at the expense of anything useful for improvement. Furthermore, the almost exclusive reliance on *relative* standards means that absolute improvements go undetected and unrewarded. There are, of course, other scoring schemes recommended in the literature [5], but none holds much promise for formative evaluation.

Grades can also be abused. Unwarranted generous assessments are sometimes given for the purposes of reducing the number of students who object to their grades or to ensure enrollments in subsequent courses. False praise is also given with the sincere but mistaken intention of preventing injury to a student's self-image. This is, in the long run, counterproductive. The desire to label everything as good rests on two false assumptions, namely (1) that any negative reaction is bound to stifle personal development and creativity, and (2) that evaluating a performance as a performance is equivalent to judging a person as a person. Not everything produced by human beings, even honest and diligent ones, is good, and students are not so naïve.

Feedforward

Prior specification, when it is done at all, is often done badly. Typically it focuses on content rather than on what to do with the content or on the evaluative criteria to be used in assessing the work. It is sometimes deliberately left vague so that responses will not be too stereotyped or convergent. This fear is unjustified. Specification of criteria and standards, although difficult to make operational, actually implies little such risk. If anything, the emphasis ought to be reversed: choice of content left open, and criteria and treatment fairly tightly specified.

Teachers often shelter behind undefined criteria until students submit their work, and then provide rationalizations of evaluations and grades after papers are returned. In other words, there is often the temptation to see what the students have done first. It is then irresponsible to say to students: "What I was really looking for was. . . ." The student has no recourse for this, because the teacher can claim to have been "looking for" any number of things, at least some of which could conceivably be invented on the spot. From the student's point of view, such rationalizations are indistinguishable from preexisting sets of criteria that were simply not made public. They cannot, therefore, be easily challenged (although one might challenge the lack of openness and justice).

Looking at what the students have done is also a means of getting around sloppiness in the task specification. It provides a baseline that adjusts readily not only to what the students have done, but also to what the question literally demanded. In a more open environment, any deficiencies in either task specification or performance could be recognized and corrected.

Feedback

Benign comments (such as "a good point" written in the margin) often substitute for sound constructive criticism. The measure of the quality of an evaluation is the extent to which the assessment focuses on significant, difficult, or sophisticated criteria rather than pedantic, trivial, obvious, or unimportant features. Deficient evaluations concentrate on typographical errors, sentence constructions, paragraphing, spelling, and sticking to the rules. Not that these are of no consequence; indeed, any proper evaluation should take them into account. The weak evaluation, however, takes *only* these into account. A severe test of whether a teacher's evaluation of student work is adequate is to ask if, when the shortcomings which have been noted are eliminated, the work would be judged "excellent."

Course Length

The introduction of semester courses and credit points has certainly increased the range of student choice, but at some cost to other aspects of the learning environment. Short courses (some only eight to ten weeks long) mean frequent transitions and a consequent lack of continuity in criterion usage. In a course as short as one semester, it is not uncommon for students to submit only a few pieces of work. Even if it were possible to have instant, comprehensive feedback, the student would have too few cues to develop a concept of excellence in the time available. Discovery learning of the relationships between a performance and the criteria used to appraise it is inefficient, uncertain, and unjust. These relationships are too important to be left to chance.

Furthermore, there is no universally agreed-upon set of rules, even within a particular discipline, and teachers' expectations differ from course to course. Because a teacher's appraisal often depends in part on how congruent a student's notions are with those of the teacher (whether they be ideas, opinions, or criteria), a student may produce a work that conforms to one teacher's criteria but have it judged inferior by another teacher using a different set. The more transitions there are, the more difficult it is for students to improve academically, even though they may progress through a sequence of courses.

Cumulative Assessment

Although intended to reduce stress and to achieve a better sampling of student performance, the benefits of cumulative assessment have also to be weighed against the costs. When students take seriously only those tasks which count towards a grade or course credit, the notion of external evaluation is consolidated, and there is little incentive for recycling. Because of the emphasis on grades rather than on formative evaluation, faculty feel defensive and students feel vulnerable. The mystique of eval uation replaces openness and negotiated discussion.

A further consequence is that the student who is slow to develop an adequate concept of quality is penalized. Consider a hypothetical example. Two students, *A* and *B*, begin a new course. *A* has completed some cognate courses, and already has a reasonable knowledge of the criteria and standards likely to be used in evaluating performance. On the other hand, *B* is new to the discipline, and has only hazy ideas of what is required. Assume that the students are of equal ability and reach the same level of performance at the end of semester. If work is submitted at

several points during the period and a total score obtained for each by weighting the sub-scores equally, student *A* achieves a much higher points total than *B*, although their ultimate performances are equivalent.

It is clear that cumulative assessment, in the context of students who improve as a direct result of growth in knowledge of the goal, is a mapping of the *route* taken to achievement, rather than a measure of the achievement itself. (Remember that for the complex types of expertise referred to in this article, there is no "simple practice effect" such as occurs in learning to use a typewriter.) This differentiation is quite noticeable among students who begin studies in a new school (for example, education) having previously completed first degrees in arts (verbal emphasis) and science (problem-solving emphasis).

Conclusion

While the general theory outlined in part 1 needs further development and refinement, the intention has been threefold: (1) to show that evaluation for improvement has so far been insufficiently analyzed as a process, (2) to show that in spite of the complexity of the problem, some systematic attack is possible, and (3) to indicate some of the implications for practice which may flow from such analysis.

I have taken an idealistic line, and assumed that students should find academic learning an enjoyable experience. The reality is that many have already been driven underground, developed a hostility towards learning, and use a variety of subterfuges for coping with assessment. How to undo the effects of years of conditioning is a problem of some magnitude. Furthermore, the evaluation principles I have proposed are more demanding of faculty time, and this cannot be easily obtained except at the expense of reduced class contact and fewer, longer courses. Personally, I believe that to be a desirable tradeoff.

Current evaluation technology naturally reflects the structure, organization, and overt requirements of higher education. The emphasis on certification, accountability, and the assessment of faculty are reflected in concern with test and examination procedures, with the reliability and predictive validity of grades, and with methods of enhancing the effectiveness of instruction. There is less concern with open criteria and standards, with ways to make them operational, with absolute standards, and with the justification of evaluations. The requirements for effective formative evaluation to improve academic learning are unique, and constitute an important but neglected area of evaluation research.

References

1. Bjorkman, M. "Feedforward and Feedback as Determiners of Knowledge and Policy: Notes on a Neglected Issue." *Scandinavian Journal of Psychology,* 13 (September 1972), 152–58.
2. Conlan, G. *How the Essay in the CEEB English Test Is Scored.* Urbana, Ill.: National Council of Teachers of English, 1976.
3. Ebel, R. L. "Evaluation of Students: Implications for Effective Teaching." *Educational Evaluation and Policy Analysis,* 2 (January-February 1980), 47–51.
4. Nimmo, D. B. "The Undergraduate Essay: A Case of Neglect?" *Studies in Higher Education,* 2 (October 1977), 183–89.
5. Odell, L., and C. R. Cooper. "Procedures for Evaluating Writing: Assumptions and Needed Research." *College English,* 42 (September 1980), 35–43.
6. Polanyi, M. *Personal Knowledge.* London: Routledge and Kegan Paul, 1962.
7. Smith, P. L. "On Creating and Using Standards." *Educational Theory,* 28 (Winter 1978), 44–53.
8. Stallings, W. M., and E. K. Leslie. "Student Attitudes Toward Grades and Grading." *Improving College and University Teaching,* 18 (Winter 1970), 66–68.