Instructional Science 18: 119-144 (1989)

© Kluwer Academic Publishers, Dordrecht – Printed in the Netherlands

Formative assessment and the design of instructional systems

D. ROYCE SADLER
Assessment and Evaluation Research Unit, Department of Education, University of Queensland, St.Lucia, Queensland 4067, Australia

Abstract. The theory of formative assessment outlined in this article is relevant to a broad spectrum of learning outcomes in a wide variety of subjects. Specifically, it applies wherever multiple criteria are used in making judgments about the quality of student responses. The theory has less relevance for outcomes in which student responses may be assessed simply as correct or incorrect. Feedback is defined in a particular way to highlight its function in formative assessment. This definition differs in several significant respects from that traditionally found in educational research. Three conditions for effective feedback are then identified and their implications discussed. A key premise is that for students to be able to improve, they must develop the capacity to monitor the quality of their own work during actual production. This in turn requires that students possess an appreciation of what high quality work is, that they have the evaluative skill necessary for them to compare with some objectivity the quality of what they are producing in relation to the higher standard, and that they develop a store of tactics or moves which can be drawn upon to modify their own work. It is argued that these skills can be developed by providing direct authentic evaluative experience for students. Instructional systems which do not make explicit provision for the acquisition of evaluative expertise are deficient, because they set up artificial but potentially removable performance ceilings for students.

Introduction

This article is about the nature and function of formative assessment in the development of expertise. It is relevant to a wide variety of instructional systems in which student outcomes are appraised qualitatively using multiple criteria. The focus is on judgments about the quality of student work: who makes the judgments, how they are made, how they may be refined, and how they may be put to use in bringing about improvement. The article is prompted by two overlapping concerns. The first is with the lack of a general theory of feedback and formative assessment in complex learning settings. The second concern follows from the common but puzzling observation that even when teachers provide students with valid and reliable judgments about the quality of their work, improvement does not necessarily follow. Students often show little or no growth or development despite regular, accurate feedback. The concern itself is with whether some learners fail to acquire expertise because of specific deficiencies in the instructional system associated with formative assessment.

The discussion begins with definitions of feedback, formative assessment and qualitative judgments. This is followed by an analysis of certain patterns in teacher-student assessment interactions. A number of causal and conditional linkages are then identified. These in turn are shown to have implications for the design of instructional systems which are intended to develop the ability of students to exercise executive control over their own productive activities, and eventually to become independent and fully self-monitoring.

Formative assessment, feedback and self-monitoring

Etymology and common usage associate the adjective *formative* with forming or moulding something, usually to achieve a desired end. In this article, *assessment* denotes any appraisal (or judgment, or evaluation) of a student's work or performance. (In some contexts, assessment is given a narrower and more specialized meaning; some North American readers in particular may prefer to substitute the term *evaluation* for *assessment.*)

Formative assessment is concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning.

Summative contrasts with formative assessment in that it is concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting at the end of a course of study especially for purposes of certification. It is essentially passive and does not normally have immediate impact on learning, although it often influences decisions which may have profound educational and personal consequences for the student. The primary distinction between formative and summative assessment relates to purpose and effect, not to timing. It is argued below that many of the principles appropriate to summative assessment are not necessarily transferable to formative assessment; the latter requires a distinctive conceptualization and technology.

Feedback is a key element in formative assessment, and is usually defined in terms of information about how successfully something has been or is being done. Few physical, intellectual or social skills can be acquired satisfactorily simply through being told about them. Most require practice in a supportive environment which incorporates feedback loops. This usually includes a teacher who knows which skills are to be learned, and who can recognize and describe a fine performance, demonstrate a fine performance, and indicate how a poor performance can be improved. Feedback can also be defined in terms of its effect rather than its informational content: "Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (Ramaprasad, 1983, p. 4). This alternative definition emphasizes the system-control function. Broadly speaking, feedback provides for two main audiences, the teacher and the student. Teachers use feedback to make programmatic decisions with respect to readiness, diagnosis and remediation. Students use it to monitor the strengths and weaknesses of their performances, so that aspects associated with success or high quality can be recognized and reinforced, and unsatisfactory aspects modified or improved.

An important feature of Ramaprasad's definition is that information about the gap between actual and reference levels is considered as feedback *only when it is used to alter the gap.* If the information is simply recorded, passed to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed and "dangling data" substitute for effective feedback. In any area of the curriculum where a grade or score assigned by a teacher constitutes a one-way cipher for students, attention is diverted away from fundamental judgments and the criteria for making them. A grade therefore may actually be counterproductive for formative purposes.

In assessing the quality of a student's work or performance, the teacher must possess a concept of quality appropriate to the task, and be able to judge the student's work in relation to that concept. But although the students may accept a teacher's judgment without demur, they need more than summary grades if they are to develop expertise intelligently. The indispensable conditions for improvement are that the *student* comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced *during the act of production itself*, and has a repertoire of alternative moves or strategies from which to draw at any given point. In other words, students have to be able to judge the quality of what they are producing and be able to regulate what they are doing during the doing of it. As Shenstone (correctly) put it over two centuries ago, "Every good poet includes a critick; the reverse will not hold" (Shenstone, 1768, p. 172).

Stated explicitly, therefore, the learner has to (a) possess a concept of the *standard* (or goal, or reference level) being aimed for, (b) compare the *actual* (or current) *level of performance* with the standard, and (c) engage in appropriate *action* which leads to some closure of the gap. These three conditions form the organizing framework for this article. It will be argued that they are necessary conditions, which must be satisfied simultaneously rather than as sequential steps. It is nevertheless useful to make a conceptual distinction between the conditions. The (macro) process of grading involves the first two in that it is essentially comparing a particular case either with a standard or with one or more other cases. Control during production involves all three conditions and is, by contrast, a (micro) process carried out in real time. Judging from assessment practices common in many subjects, information generated without the participation of the learner but made available to the learner from time to time (as intelligence) is evidently assumed to satisfy these conditions. A detailed examination of the three conditions shows why this assumption falls short of what is actually necessary.

For purposes of discussion, it is convenient to make a distinction between feedback and *self-monitoring* according to the source of the evaluative information. If the learner generates the relevant information, the procedure is part of self-monitoring. If the source of information is external to the learner, it is associated with feedback. In both cases, it is assumed that there has to be some closure of the gap for feedback and self-monitoring to be labelled as such. Formative assessment includes both feedback and self-monitoring. The goal of many instructional systems is to facilitate the transition from feedback to self-monitoring.

Feedback and formative assessment in the literature

Authors of textbooks on measurement and assessment published during the past 25 years have placed great emphasis on achieving high content validity in teacher-made tests, producing reliable scores or grades, and the statistical manipulation or interpretation of scores. Only cursory attention has usually been given to feedback and formative assessment, and then it is mostly hortatory, recipe-like and atheoretic. In many cases feedback and formative assessment (or their equivalents) are not mentioned at all in either the body of the text or the index, although the books by Rowntree (1977), Bloom, Madaus and Hastings (1981), Black and Dockrell (1984) and Chater (1984) are notable exceptions.

In general, a concern with the aims of summative assessment has dominated the field in terms of both research and the guidance given to teachers (Black, 1986). This dominance is implicit in the treatment given, for instance, to reliability and validity. Textbooks almost invariably describe how the validity (of assessments) is to be distinguished from the reliability (of grades or classifications). Reliability is usually (and correctly) said to be a necessary but not sufficient condition for validity, because measurements or judgments may be reliable in the sense of being consistent over time or over judges and still be off-target (or invalid). Reliability is therefore presented as a precondition for a determination of validity. In discussing formative assessment, however, the relation between reliability and validity is more appropriately stated as follows: validity is a sufficient but not necessary condition for reliability. Attention to the validity of judgments about individual pieces of work should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement. Reliability will follow as a corollary. Acceptance of this principle, which is emphasized by only a few writers (such as Nitko, 1983), has implications for how the process of appraisal is conceptualized, and the mechanisms of improvement understood.

In the literature on learning research, feedback is usually identified with knowledge of results (often abbreviated to KR), a concept which gained considerable currency through Thorndike's (1913) so-called Law of Effect. Reviewing a series of experimental studies on learning from written materials (texts and programmed instruction), Kulhavy (1977, p. 211) defined feedback as "any of the numerous procedures that are used to tell a learner if an instructional response is right or wrong". Kulik and Kulik (1988) adopted a similar definition in their review of research on the timing of feedback. Learning researchers have been particularly interested in the effect of various feedback characteristics (such as immediacy, pertinence, data form and type of reward) on the retention of learned material. The research hypotheses tested have almost invariably been based on stimulus-response learning theories, the aim being to discover the types of stimuli and incentives that promote learning. For the most part, this line of research has been confined to learning outcomes that can be assessed by quizzes and progress tests consisting of problems to be solved or objective items that can be scored correct or incorrect. The learning programs are conceived of as divisible into logically dependent units which can be mastered more or less sequentially, one by one. The resulting technology is associated with test scores, diagnostic items, criterion-referencing and mastery learning.

Other lines of research occur in specific subject areas. Of particular interest is the literature on the assessment of writing, which contains descriptions of a number of different approaches, including assessment by means of general impression, analytic scales, primary traits, syntactic features, relative readability and intellectual strategy (Gere, 1980). These differ not only in procedural detail, but also in their theoretical bases. Much of the discussion about and evaluation of the various possibilities has revolved around which assessment criteria should be used (and how), which of the techniques has the soundest theoretical foundation (such as a theory of composition), or which produces the best agreement among competent judges (reliability considerations). An alternative criterion for adjudicating among assessment approaches is the extent to which students improve either as consumers of assessments arrived at by different methods, or through being trained to use a particular assessment approach themselves. With respect to the teaching of writing, these issues have not been thoroughly explored, although they are touched upon by Cooper (1977), Odell and Cooper (1980) and several others.

While the line of development in this article is different from that in the literature on writing assessment, it shares an interest in learning outcomes which are complex in the sense that qualitative judgments (defined below) are invariably involved in appraising a student's performance. In such learnings, student development is multidimensional rather than sequential, and prerequisite learnings cannot be conceptualized as neatly packaged units of skills or knowledge. Growth takes place on many interrelated fronts at once and is continuous rather than lockstep. The outcomes are not easily characterized as correct or incorrect, and it is more appropriate to think in terms of the quality of a student's response or the degree of expertise than in terms of facts memorized, concepts acquired or content mastered.

Qualitative judgments defined and characterized

A *qualitative judgment* is defined (Sadler, 1987) as one made directly by a person, the person's brain being both the source and the instrument for the appraisal. Such a judgment is not reducible to a formula which can be applied by a non-expert. In general, qualitative judgments have some or all of the following five characteristics:

1. Multiple criteria are used in appraising the quality of performances. As well as the individual dimensions represented by the criteria, the total pattern of relationships among those dimensions is important. In this sense the criteria interlock, so that the overall configuration amounts to more than the sum of its parts. Decomposing a configuration tends to reduce the validity of an appraisal.
2. At least some of the criteria used in appraisal are *fuzzy* rather than *sharp*. A sharp criterion contains an essential discontinuity which is identifiable as an abrupt transition from one state to another, such as from correct to incorrect. There may be two or more well-defined states, but it is always possible in principle to determine which state applies. Sharp criteria are involved in all objective testing (including that in the arts and humanities), and the assessment of many outcomes in mathematics and the sciences which involve problem solving and theorem proving. By contrast, fuzzy criteria are characterized by a continuous gradation from one state to another. *Originality*, as applied to an essay, is an example of a fuzzy criterion because everything between wholly unoriginal and wholly original is possible. A fuzzy criterion is an abstract mental construct denoted by a linguistic term which has no absolute and unambiguous meaning independent of its context. If a student is to be able to consciously use a fuzzy criterion in making a judgment, it is necessary for the student to understand what the fuzzy criterion means, and what it implies for practice. Therefore, learning these contextualized meanings and implications is itself an important task for the student.
3. Of the large pool of potential criteria that could legitimately be brought to bear for a class of assessments, only a relatively small subset are typically used at any one time. The competent judge is able not only to make an appraisal, but also to decide which criteria are relevant, and to substantiate a completed judgment by reference to them. In many cases, the teacher may find it impossible to specify all of the relevant criteria in advance, or may find that a fixed set of criteria is not uniformly applicable to different student responses, even though those responses may ostensibly be to the same task. Professional qualitative judgment consists in knowing the rules for using (or occasionally breaking) the rules. The criteria for using criteria are known as *metacriteria*.

4. In assessing the quality of a student's response, there is often no independent method of confirming, at the time when a judgment is made, whether the decision or conclusion (as distinct from the student's response) is correct. Indeed, it may be meaningless to speak of correctness at all. The final court of appeal is to another qualitative judgment. To give an example of methodological independence, suppose that two essays are to be compared. One approach is to ask a competent person to judge which is of higher quality, with or without specifying the criteria. A different method of judging quality would be to use a computer program to analyse certain textual properties such as the frequency of commas, and the proportions of prepositions, conjunctions and uncommon words. These two methods are independent because they use essentially different means for arriving at a conclusion. But having two persons instead of just one would not constitute independent methods, even if both persons were to make the judgments without reference to each other, and in that sense work independently.
5. If numbers (or marks, or scores) are used, they are assigned after the judgment has been made, not the reverse. In making qualitative judgments, the final decision is never arrived at by counting things, making physical measurements, or compounding numbers and looking at the sheer magnitude of the result.

Complex learning outcomes of the type that are assessed by making direct qualitative judgments are common in a wide variety of subjects in secondary, vocational, further and higher education. These subjects include English, foreign languages, humanities, manual and practical arts, social sciences, and the visual and performing arts. They are also important in industrial training and in many areas of science and mathematics, particularly where students are required to devise experiments, formulate hypotheses or explanations, carry out open-ended field or laboratory investigations, or engage in creative problem solving. Assignments and tasks set in all of these areas involve students in actively synthesizing and integrating ideas, concepts, movements or skills to produce extended responses in some form. In all assessments of such extended responses, qualitative judgments are of fundamental importance.

Sometimes the student response or end product has a permanent form, an existence separate from the learner. That is, it is an artefact which is open to leisurely inspection. Examples include essays, musical compositions, welding jobs, and articles of pottery. If the scaffolding used in the construction of the work is carefully dismantled, the final product may retain no evidence of false starts, unfruitful paths followed in its production, or (if it has not been produced under time-constrained test conditions), the time taken to produce it. The product is, in fact, infinitely malleable prior to its release, and the author can modify it by any desired amount. A contrasting type of end "product" is when the learner's work is transient, such as a live production performed by the learner in real time.

Examples are a dramatic performance, a speech, an interview with a patient or client, a classroom lesson, or a game of tennis. Note that making a recording of a live performance produces only a secondary artefact which, while useful in analysis and review, is distinctively different in character from the performance itself, and from, say, a carefully edited movie or record album produced over several months. Artefactual and transient end products make different demands on the instructional system in terms of evaluative feedback.

It is also useful to make a distinction among end products according to the degree of design expected. In some fields of learning, the desired end product is tightly specified (for example, by technical drawings) to the extent that if the constructive abilities of all producers were perfect, the outcomes would be more or less identical. What is assessed in these situations is essentially the learner's productive skill. Assessing such outcomes may or may not involve making qualitative judgments, depending on the number and nature of the criteria. In other fields (such as writing), design itself is an integral component of the learning task, although it may be so closely linked with production that is does not appear as a distinct phase. In yet other fields (such as fashion and architecture), design itself may be the primary consideration. Wherever the design aspect is present, qualitative judgments are necessary and quite divergent student responses could, in principle and without compromise, be judged to be of *equivalent* quality.

Communicating standards to students

Earlier in this article, it was argued that the transition from feedback to self-monitoring can occur only when three conditions are satisfied. The first of these is that the student comes to know what constitutes quality. In a teaching setting, this presupposes that the teacher already possesses this knowledge, and that it must somehow be shared with the student. In a particular context, however, it is often difficult for teachers to describe exactly what they are looking (or hoping) for, although they may have little difficulty in recognizing a fine performance when it occurs among student responses. Teachers' conceptions of quality are typically held, largely in unarticulated form, inside their heads as tacit knowledge. By definition, experienced teachers carry with them a history of previous qualitative judgments, and where teachers exchange student work among themselves or collaborate in making assessments, the ability to make sound qualitative judgments constitutes a form of guild knowledge.

While such in-the-head standards exhibit a degree of stability, they are not immutable but can be shown to adapt to the circumstances. In particular, teachers are often strongly influenced by the range of quality which exists among a set of things to be appraised, and typically find it difficult to make an isolated judgment of quality (that is, without reference to other students' work). Teachers tacitly acknowledge the difficulty of relying on memory alone when they make a survey of pieces of student work before assigning grades to them. This survey generates a loosely quantitative baseline or frame of reference for what is to be regarded as barely satisfactory and what is to count as excellent in the context. Even after a survey has been made, however, smaller scale *order* effects (especially severity, leniency, and carryover) almost invariably occur. This is a subject of continuing research (see, for example, the work of Hales and Tokar, 1975, and Daly and Dickson-Markman, 1982) and can be interpreted in terms of Helson's (1959) adaptation level theory. It therefore appears that teachers' conceptions of quality and standards exist in some quiescent and pliable form until they are reconstituted by fresh evaluative activity.

In an instructional system, an exclusive reliance on teachers' guild knowledge works against the interests of the learner in two important ways. In the first place, although the practice of surveying a sample of performances is common (and advisable where the aim is fair ranking of one student's work against that of other students), it is inappropriate for formative assessment because it legitimates the notion of a standards baseline which is subject to existential determination. Strictly speaking, all methods of grading which emphasize rankings or comparisons among students are irrelevant for formative purposes. Assuming that sorting and stratifying learners is not the main purpose of education and training, the objective for each student is acquire expertise in some absolute sense, not merely to surpass other students. Secondly, guild knowledge keeps the concept of the standard relatively inaccessible to the learner, and tends to maintain the learner's dependence on the teacher for judgments about the quality of performance. How to draw the concept of excellence out of the heads of teachers, give it some external formulation, and make it available to the learner, is a nontrivial problem. It is dealt with at some length elsewhere under the rubric of *standards-referenced* assessment (Sadler, 1987). Some of that material is summarized below.

Two approaches to specifying standards are through descriptive statements and exemplars. While neither of these is sufficient in itself, a combination of verbal descriptions and associated exemplars provides a practical and efficient means of externalizing a reference level. Descriptive statements set out the characteristic properties of a performance at a designated level of quality. The following generic description of high quality in a particular writing task is an example of a descriptive statement:

> There is a logical progression of ideas from an original hypothesis to a final conclusion. Facts are reported accurately, and the inferences drawn are plausible. The author maintains some "distance" from the content, thereby achieving a degree of objectivity. The whole piece hangs together well, the wording is appropriate, and the mechanical aspects of writing are flawless.

Descriptive statements may be used to specify anchor points on a quality continuum, and may include specifics that are present/absent (such as a statement of the hypothesis) or correct/incorrect (such as spelling and punctuation), along with other features which are present to a greater or lesser degree (such as "hanging together well"). They go part way towards externalizing standards, and may be derived inductively by first classifying or grading student achievements holistically, and then abstracting and codifying the distinguishing features of the different classes.

Levels of quality or performance can also be conveyed in part by means of a set of key examples or *exemplars*, chosen so as to illustrate what distinguishes high quality from low. The advantage of exemplars for both teacher and learner is that they are concrete. The minimum number necessary to convey a particular reference level exclusively by exemplars can be shown theoretically to depend upon the number of criteria to be used. The more criteria there are, the greater the number of ways in which work of a given quality may be constructed.

Some teachers may be concerned that the use of exemplars as indicators of standards would encourage students to slavishly copy the exemplars themselves, and so stimulate convergent or stereotyped rather than original responses from students. Students could become blinkered and have their creativity stifled. The first counterargument to this view is that a *single* exemplar is inadequate to convey a standard anyway. Students need, in many educational contexts, to be presented with several exemplars (for a single standard) precisely to learn that there are different ways in which work of a particular quality can find expression. There is often a wide variety of objects within the same genre which are regarded as excellent. Unless students come to this understanding, and learn how to abstract the qualities which run across cases with different surface features but which are judged equivalent, they can hardly be said to appreciate the concept of quality at all.

The second consideration is that originality and creativity are not usually, contrary to some opinion, best developed in a completely freewheeling environment. Bailin (1987) pointed out that there is no essential conflict between creative processes and the production of something which is generally accepted as of high quality. Creative productions are mostly highly disciplined, and are almost invariably produced not by accident or through random risk taking but when the producer, by being thoroughly conversant with the characteristics of the discipline or genre, understands when and how to transcend the normal boundaries. Knowing the metacriteria, that is, knowing when the suspension of some criterion, even on occasion a principal one, can be justified in favour of another, is an important element in creativity. But to return to the issue of exemplars, it is the experience of many teachers that even if some students do in fact copy, they may learn something valuable in the process. Emulation is an ancient and still almost universal learning method. When students have gained whatever they can from, in the worst case, slavish copying, there is time for the teacher to wean them away from it.

Students develop a concept of a reference level more readily in some learning contexts than in others. In the manual, visual and performing arts, for example, students are usually able to observe, as a matter of course, the results of other students' efforts together with the teachers' appraisals of those efforts, simply because the work is produced in workshops, studios, theatres and other open environments. The best examples, or perhaps exemplary material developed outside the classroom, serve naturally and unobtrusively as reference points. In the liberal arts and humanities, however, students often work privately, and do not get to see or read what other students have produced. What constitutes work of high quality then remains to some extent unknown. Exceptional cases aside, it is ironic that the prototypes of competency levels which Myers (1980) recommended as necessary for assessors using holistic methods for the evaluation of writing are not similarly considered a general requirement for students learning to write or to master other complex skills.

Standards as goals or aspirations

In its simplest form, a standard or reference level is a designated degree of performance or excellence. It becomes a *goal* when it is desired, aimed for, or aspired to. Some goals are external (assigned by a teacher) while others are developed or adapted by the learners themselves. A learner may decide to ignore or reject an external goal, in which case it is likely to have little if any effect on achievement except in a coercive situation. Only when a learner assumes ownership of a goal can it play a significant part in the voluntary regulation of performance.

The effect of goals on performance has been the subject of a great deal of research over recent decades. For a review of some of it, see Locke, Shaw, Saari, and Latham (1981). In a wide variety of field and laboratory settings, it has been found that what are called *hard goals* have the greatest impact on performance. Hard goals are defined as being specific and clear rather than general or vague, harder and challenging rather than simple or easy, and closer to the upper limit of an individual's capacity to perform than to the current level of performance. Hard goals act to focus attention, mobilize effort, and increase persistence at a task. By contrast, do-one's-best goals often turn out to be not much more effective than no goals at all.

The discussion above has more or less implied that a single standard operates for a particular student at a particular stage of development. In general, of course, the quality of work expected of a student rises steadily as the student progresses through various years of schooling or the stages of a training program. If the rate at which expectations are raised is consistently greater than the rate of improvement, the inability of the student to keep pace results in little or no sense of accomplishment even though improvement may actually be occurring. This in turn may lead to a situation where successive attempts are taken less and less seriously, the performance gap widens progressively and becomes self-reinforcing, and the student loses heart and effectively drops out. In some subjects, the rungs of the ladder of achievement take the form of a gradation in both scope and complexity; in others, they reflect different standards on a well-defined quality dimension. In classroom settings, students may need access to a range of standards (not just the top rung) to cater for different abilities. (Whether this range corresponds to the grade designations on an educational certificate is irrelevant).

It would be useful to research the optimum gap between an individual learner's current status and the aspiration. If the learner perceives the gap as too large, the goal may be regarded as unattainable. The same gap (in absolute terms) may, however, provide a powerful stimulus for another highly motivated and confident student, who would not be put off by a sequence of initial failures. Conversely, if the gap is perceived as too small, closing it might be considered not worth any additional effort. Initially, the teacher may find it useful to negotiate the aspiration level with the student, or at least to take individual student characteristics into account. The ultimate aim should be to have the student set, internalize and adopt the goal, so that there is some determination to reach it.

Making multicriterion judgments

In addition to knowing about appropriate standards, students have to be able to compare their actual levels of performance with these standards. This requires that they be capable not only of making multicriterion judgments about their own work but also of making them with a proper degree of objectivity and detachment. To provide a background for the discussion in this section, consider the special case of the assessment of written composition. This choice has been made because of the substantial body of literature on the topic and because written work is required in a wide variety of subjects.

At least 50 criteria have been identified for assessing the quality of written composition. All of the criteria in the list below have been extracted from published sources, although an examination of teachers' written comments indicates that even this list is not exhaustive. The criteria themselves are italicized, with apparent synonyms placed together.

> *accuracy* (of facts, evidence, explanations); *audience* (sense of); *authenticity; clarity; coherence; cohesion; completeness; compliance* (with conventions of the genre); *comprehensiveness; conciseness (succinctness); consistency* (inter-

nal); *content* (*substance*); *craftsmanship; depth* (of analysis, treatment); *elaboration; engagement; exemplification* (use of examples or illustrations); *expression; figures of speech; flair; flavour; flexibility; fluency* (or *smoothness*); *focus; global* (or overall) *development; grammar; handwriting* (*legibility*); *ideas; logical* (or chronological) *ordering* (or *control of ideas*); *mechanics; novelty; objectivity* (or *subjectivity*, as appropriate); *organization; originality* (*creativity, imaginativeness*); *paragraphing; persuasiveness; presentation* (including *layout*); *punctuation* (including *capitalization*); *readability; referencing; register; relevance* (to task or topic); *rhetoric* (or *rhetorical effectiveness*); *sentence structure*; *spelling; style; support for assertions; syntax; tone; transition; usage; vocabulary; voice; wording.*

Several of these appear in a number of the most popular listings, of which Diederich's (1974) is one of the best known. However, most of the others (even those not commonly used by teachers in general) would be acknowledged as relevant (at least for some genres of writing) by teachers of English. Some of the criteria are fairly subtle. (What exactly is meant by flair?) Some are likely to be used so infrequently that detailed explication is hardly justified. Some apply to particulars (accuracy, support for assertions); others apply only to a work taken as a whole (coherence, comprehensiveness). Some are sharp (certain aspects of punctuation, for example); most are fuzzy. Some overlap conceptually with others (rhetoric, style, persuasiveness); some apply to particular genres of writing but not to others (referencing); and some logically subsume others (mechanics subsumes spelling). Many are operationally correlated together, so that whenever an attempt is made to change a piece of writing according to one dimension, other properties are inevitably affected at the same time. For example, it may be impossible to change the vocabulary of a piece of writing without simultaneously affecting the tone. In short, this set of criteria is large and includes subsets which overlap and interlock. It is therefore obvious that behind the customary published lists (usually consisting of from seven to ten criteria) there lies a much larger set of potential criteria that could be brought into play if and when the need arises. Given this fact, and the complex interrelations which exist among the criteria, it is clear that to use the whole set for a particular assessment would be unmanageable. How judges cope with the situation therefore requires some investigation.

The literature on research into human judgmental processes in a variety of settings is both instructive and extensive, and cannot be adequately summarized here. But of particular concern to researchers have been the inefficiency of intuitive judgmental processes, and the limitations in human information processing capacities which result in biased or defective decisions (Sadler, 1981). In broad terms, the many techniques proposed for making complex judgments fall more or less into two camps, each of which has its research tradition, its advocates and its detractors. Fortunately, it not necessary to make a firm decision on one or the other for purposes of formative assessment. Both can be drawn upon because evaluative input can take any appropriate form, and in any case is always open to discussion, clarification, and revision if necessary.

The first general line of attack is to devise and implement a procedure which begins with identifying a number of relevant criteria, then measures the amount present on each criterion and combines the various levels or estimates into an overall measure of merit by means of a formula. The criteria are treated separately, so that the order in which characteristics are considered is arbitrary and has no effect on the final result. The combining formula may be simple, and require only the addition of weighted or unweighted component scores or ratings. On the other hand, the formula may be complicated (taking, for example, conjunctive or disjunctive form). This so-called *analytic* approach is common in evaluating consumer products. The global judgment is made by breaking down the multicriterion judgment using separate criteria and then following explicit rules. If necessary, the judgment can be justified by retracing and checking for integrity all the steps that led to it. In assessing student work, the analytic approach typically settles on the set of criteria considered to be most relevant to the work of most students at a particular stage of development. The criteria may be simply selected by a teacher on the basis of their logical relevance to the task, or may result from empirical studies (using factor or regression analyses) of the judgmental behaviours of competent assessors. Diederich (1974) followed the latter approach. This component-wise attack on the problem of making multicriterion judgments is often advocated as the ideal towards which impressionistic, holistic or informal systems should be made to move. It assumes, however, that the set of criteria nominated is sufficient for all cases, that the criteria do not overlap, and that use of the combining formula leads to judgments which would not conflict (except perhaps rarely) with more holistic approaches. A substantial argument has been mounted elsewhere (Sadler, 1985) that for complex phenomena, use of a fixed set of criteria (and therefore the analytic approach) is potentially limiting.

The second approach to making complex judgments is for the evaluator to react to the work as a whole, making an entire, or what Kaplan called a *configurational* (1964, p. 211), assessment first and then to substantiate it (to whatever extent is necessary) by referring to separate criteria, which may or may not be drawn from a prespecified set. In this approach, imperfectly differentiated criteria are compounded as a kind of gestalt and projected onto a single scale of quality, not by means of a formal rule but through the integrative powers of the assessor's brain. To produce a rationale for such a holistic or global judgment, the assessor unpacks some of the conceptual unidimensionality. Configurational assessments do not require the specification of all criteria in advance, neither do they assume operational independence among the criteria.

In making configurational judgments, competent judges select, from the large pool of possible criteria, those which are salient to a particular appraisal. All of the properties of a piece of student work which the teacher regards as normal, ordinary, or expected (and which therefore do not call for either positive or negative comment) naturally have low salience. Wittgenstein (1967, 1974) pointed out something which is obvious once it is stated: what is ordinary does not call attention to itself. "Does everything that we do not find conspicuous make an impression of inconspicuousness? Does what is ordinary always make the impression of ordinariness?" (Article 600). Something ordinary, therefore, is not "remark"-able. Something out of the ordinary invites attention. High salience implies that the amount of the property the object or performance possesses is different from what is considered normal, and that an evaluation of the object would typically mention this characteristic in its rationale.

Once a criterion has been identified in one or more evaluations, the judge's sensitivity to that criterion is temporarily increased and it is more likely to be attended to in subsequent evaluations. That is, the potential salience increases. In the sense described above, the salience of a particular criterion is related to the way the work being appraised is perceived. It is, therefore, a function of both the condition of mind of the perceiver and the properties of the object being assessed. Which of the potential criteria are singled out for mention has less to do with what is detectable through the senses than with what is deemed to be *worth noticing.* Consider, for example, the comments a teacher may make on a student's written work, particularly those which are made progressively as the teacher (more or less instantaneously) senses positive and negative points worthy of note. Some comments (such as "Yes", or "I agree!") are non-specific, or are not related directly to the quality of the written piece. Other comments are evaluative, and clearly imply criteria. It can be demonstrated that when a teacher, on two or more separate occasions, makes running evaluative comments together with an overall assessment of quality on a piece of student work, the overall judgments may be identical but the running comments may differ from occasion to occasion. The comments may be made at different places in the writing, or if at the same point, may differ in content. It also can be demonstrated that several assessors may agree on an overall judgment, but for different reasons.

This phenomenon has implications for formative assessment, because it raises the question of whether students can be expected to make systematic progress when teachers appear to operate probabilistically. The obvious solution is to revert to the analytical approach and make it clear to students that certain nominated criteria are what will be used in appraisal. Many teachers follow this practice, distributing *criteria sheets* to their students either as part of the task specifications or (less usefully if the criteria change from task to task) when returning assessed papers. Teachers who use criterion sheets regularly, however, find that while such sheets are helpful, they may lead to frustration because of their inflexibility. The qualities of a piece of work cannot necessarily be dealt with adequately using a fixed criterion set, and teachers often feel the need to call upon nonstandard criteria.

A more satisfactory (and less mechanistic) solution to the problem is to consider the universe of criteria as notionally partitioned into two subsets called for convenience *manifest* and *latent* criteria (Sadler, 1983). Manifest criteria are those which are consciously attended to either while a work is being produced or while it is being assessed. Latent criteria are those in the background, triggered or activated as occasion demands by some (existential) property of the work that deviates from expectation. Whenever there is a serious violation of a latent criterion, the teacher invokes it, and it is added (at least temporarily) to the working set of manifest criteria. This is possible because competent teachers have a thorough knowledge of the full set of criteria, and the (unwritten) rules for using them. But it is precisely this type of knowledge which must be developed within the students if they are to be able to monitor their own performances with a reasonable degree of sophistication. The translation of a criterion from latent to manifest should therefore not be interpreted by either the student or the teacher as unfair or as some sort of aberration. Because of the practical impossibility of employing all criteria at once, it is inevitable and perfectly normal. Marshall (1958, 1968) referred to this as the *flotation principle*, and advocated its use in evaluation. In an interesting shift of metaphor, it also formed the basis for Elbow's (1973) so-called *center of gravity* approach to appraising student writing for formative purposes.

The art of formative assessment is to generate an efficient and partly reversible progression in which criteria are translated for the student's benefit from latent to manifest and back to latent again. The aim is to work towards ultimate submergence of many of the routine criteria once they are so obviously taken for granted that they need no longer be stated explicitly. The necessity to recycle work through the teacher (for appraisal) can be reduced or eliminated only to the extent that students develop a concept of quality, and the facility for making multicriterion judgments. This in turn requires that they be given adequate evaluative experience themselves.

Direct evaluative experience

When students have to rely solely on, say, teachers' written comments, not only is the feedback conveyed in propositional form, but the number of comments and their content depends upon the willingness of the teacher (and the time available) to actually make the comments, the ability of the teacher to express the feedback in words, and the ability of the student to interpret the comments. The student may not, for instance, know what is implied by references to particular evaluative criteria. For example, suppose a teacher points out to a student that something produced is not as *coherent* as it should be. As a criterion, coherence implies that how something hangs together is important in appraising it. Coherence is clearly relevant to evaluating a variety of things: a painting, an essay, a dramatic segment, and so on. The nature of the elements that have to cohere (visual elements, concepts and ideas, physical movements), the serial and lateral connections between these elements, and the relation of each part to the whole, may not necessarily be clear to the student unless the contextual meaning of coherence is explained. Exactly what coherence implies in one context does not transfer directly to another context, although the basic idea is the same. Because much of the evaluative knowledge underlying teachers' comments is tacit, the learner also has a need to develop an appropriate body of tacit knowledge to be able to interpret formal statements.

Criteria often seem elusive partly because what a criterion means and what it implies for appraisal cannot necessarily be defined in isolation from concrete examples of things which possess the property in question, which in any case is usually only one of many properties. Coming to an understanding of the property is therefore as much an epistemological as it is a technical matter. To clarify the meaning and implications of a particular criterion, it would be useful to have a set of graded examples exhibiting more or less of that property. But for works of art or pieces of literature, the various properties are inevitably compounded together, so that one cannot create or collect examples for which all properties other than the one in question are held constant. This is in contrast with a dichotomous criterion such as correctness, for which positive and negative instances may usually be produced on demand.

A novice is, by definition, unable to invoke the implicit criteria for making refined judgments about quality. Knowledge of the criteria is "caught" through experience, not defined. It is developed through an inductive process which involves prolonged engagement in evaluative activity shared with and under the tutelage of a person who is already something of a connoisseur. By so doing "the apprentice unconsciously picks up the rules of the art, including those which are not explicitly known to the master... Connoisseurship... can be communicated only by example, not by precept" (Polanyi, 1962, p. 53-54). In other words, providing guided but direct and authentic evaluative experience for students enables them to develop *their* evaluative knowledge, thereby bringing them within the guild of people who are able to determine quality using multiple criteria. It also enables transfer of some of the responsibility for making evaluative decisions from teacher to learner. In this way, students are gradually exposed to the full set of criteria and the rules for using them, and so build up a body of evaluative knowledge. It also makes them aware of the difficulties which even teachers face of making such assessments; they become insiders rather than consumers.

For some types of learning, there is a further fundamental reason for deliberately developing tacit (as distinct from explicit or propositional) evaluative knowledge through experience. Consider the case when the learner's work consists of a live production such as a musical performance. If the performer focuses too consciously on either the mechanics of production or the control of production during the performance itself, the quality of the performance frequently suffers. Occasionally the loss of quality is catastrophic. The performer needs to control the performance using what Polanyi calls *subsidiary awareness* (1962, p. 55) of the state of play at any instant. Subsidiary awareness draws subconsciously on a body of tacit evaluative knowledge. By contrast, a *focal awareness* may interfere with and be detrimental to the performance. Fortunately, learning contexts in which live performances are common also provide, in most cases, an abundance of illustrative performances and opportunities for appraisal.

Most of the discussion above is valid regardless of whether the criteria are viewed as discrete or interlocked. If the criteria are considered separately, appraisals are concerned more with individual properties or *qualities* than with *quality* in the broader sense. There are, however, two reasons for encouraging students to make configurational judgments of overall quality as well, making use of a number of criteria simultaneously. Firstly, students need to be able to appraise a work as a whole in order to appreciate how different varieties within the one class or genre (such as the short story) can be of comparable quality even though the basic design or structural features are different. Separate consideration of the criteria does not necessarily create the experience of how they may all be put together. Part of the acquisition of creative expertise lies in learning about the permissible limits to variation within the one class, and different classes are often distinguished less by individual criteria than by characteristic configurations. The same list of criteria may be used for assessing several classes, but the criteria may require different interpretations, or differ in relative significance, from class to class. The ability to make global appraisals is therefore fundamental to understanding the nature of different classes, and hence to producing something within a particular class.

Secondly, something may apparently meet requirements on all appropriate criteria taken individually yet be unsatisfactory overall. It may be difficult to explain this anomaly to students, unless students themselves are confronted with the same evaluative problem. In a different context, Tversky (1969) suggested a line of argument which is perhaps helpful here. Suppose there exists some maximum deficit that could be tolerated on a single criterion before it would be noticed that the expectation had not been met. If on each one of a set of criteria the deficit is less than the tolerable limit, and if there are a number such criteria, the global assessment actually fails the minimum-quality test by an amount equal to the sum of the individual deficits. The global shortfall may be noticeable but not the individual shortfalls. The disqualification is then due less to a single identifiable cause than to the combined effects of marginal deficits.

Evaluative experience and task specifications

The concept of guild knowledge can be extended beyond the confines of evaluating a piece of work in isolation, to evaluating a piece of work in relation to the task specifications. In situations where students construct assignments or term papers according to specifications laid down by the teacher, it is common (and frustrating for the teacher) for a proportion of students not to address themselves to the task set. The student, for example, may do a creditable job of recounting the story of a novel instead of identifying the theme. Some teachers adopt a policy of accepting and giving partial credit (deliberately or by default) for a response which is well put together but is off-target. On the surface, this practice appears to make a reasonable concession to the hardworking student for the time and effort put in. In the long run, however, it undermines the learning which is supposed to take place, and reduces the student's incentive to tackle tasks of the type actually set. If learning how to address a set task or how to produce something within an established genre is an important instructional outcome, sticking to the task has to be a pre-emptive criterion. Meeting the generic requirement is a logical precondition for an appraisal to be made within a particular genre, but the significance of this fact may be brought home to the students only when they themselves are faced with deciding whether or not several pieces of work meet the original task specifications. In addition, it may demonstrate to them how common it is for students not to respond to the task that is actually set.

Some of the variation in quality of different students' responses to a set task may also be due to deficiencies in task definition. An appraisal of quality is then confounded by a factor which has nothing to do with the student. The specifications may be vague, incomplete or ambiguous. Alternatively, they may be technically adequate for the expert, but contain terms whose meanings and implications are not understood by the student. A common task in teaching English literature, for instance, requires the student to identify and describe the theme of a novel. Any student who does not know what is meant by the *theme* of a novel, and how the theme is distinguished from the story or the plot, cannot address the task as it is set. If the theme of a novel had been included as part of the syllabus for a previous year of schooling, the teacher might mistakenly assume that all students know what a theme is, and that the matter does not require explicit attention. Joint teacher-learner assessment is therefore useful in testing the adequacy of task specifications and modifying them if necessary for future use.

Evaluative expertise as curriculum content

In the above discussion of evaluative experience and knowledge, evaluation as curriculum content should be clearly distinguished from evaluation as an agent in learning. Evaluation and critical thinking are important aspects of many subjects and courses. It is common to find references to evaluation in syllabus statements, lists of objectives, and course outlines (relating, for example, to literary or artistic works, the significance of historical events or economic policies, or the impact of pollutants on the environment). In such cases, the student is cast in the role of assessor but the subject of the evaluation is external to both learner and teacher. This contrasts with the instrumental use of evaluative knowledge discussed above, in which the subject of the evaluation is work of the type or genre being produced or performed by the students, (but is not, of course, limited to the student's own work). Evaluative activity in the latter situation is inextricably connected with constructive activity, and is primarily enabling and facilitative rather than an end in itself.

Strategies for gap closure

In many contexts, students traditionally have more or less relied on their teachers to tell them how to effect improvement. This aspect is not dealt with in detail here, except to observe that if the teacher is to be in a position to suggest remedial moves, the teacher should ideally possess current productive expertise of the kind to be developed by the student. Apart from the issue of credibility with students, a teacher should not be purely a connoisseur who never engages in any disciplined way in productive activity. Many teachers of writing, for example, do not voluntarily write prose or poetry for either pleasure or profit apart from personal letters and other necessities. Their writing experience is vicarious and limited to the classroom setting. It consists of launching students into writing tasks of various kinds, and later helping them to improve their work. This anomalous situation parallels the experience of many students, whose only exposure to evaluative and editorial activity is as it is received from the teacher. It therefore is also vicarious.

The third condition for self-monitoring to occur is that students themselves be able to select from a pool of appropriate moves or strategies to bring their own performances closer to the goal. This requirement warrants separate consideration because the ability to evaluate others' or one's own work is not necessarily matched by the ability to produce. It is also consistent with the thesis that the possession of evaluative expertise is a necessary (but not sufficient) condition for improvement. A student in English, for example, may be able to recognize the theme in a novel once it has been identified by another person, or be able to distinguish between the theme and other nominated characteristics of a novel but be unable to engage in the abstract thought which is necessary to identify from scratch the theme or themes in an unseen novel, or to structure a written response appropriately. This ability to recognize and evaluate but not construct is not an isolated phenomenon, nor is it limited to education. There are many domains of human activity where people are expert at appraising existing objects, sometimes in a highly sophisticated way, but are themselves incapable of producing objects of the type in question. Art criticism is an example, as is anything involving connoisseurship as such. An important task of teaching, of course, is to help students develop various kinds of expertise, including those of production.

In many complex artificial systems, control is achieved by having a large number of feedback loops consisting of sensors, comparators and effectors. Typically, each corrective action is singular and tied deterministically to a particular deficiency. This also occurs with particular aspects of creative activity, such as spelling, punctuation, and the accuracy of facts in producing a composition. But the more complex a task is, and the greater the divergence in outcomes which can be regarded as acceptable, the more likely it is that a variety of ways can be devised to alter the gap between actual and reference levels, and therefore the less likely it is that information about the gap will by itself suggest a remedial action. Moves have to be imported from outside, and choices made from a range of options or possibilities available to the learner. Provided the learner appreciates the nature of the task, experience in production, evaluation and remediation provides a means of developing and maintaining a resource pool.

The complexity of multicriterion learning tasks suggests that if the student is prepared to act upon a set of identified deficiencies with a view to improvement, one list of weaknesses may be as formatively effective as another if the criteria are highly intercorrelated. On the other hand, improvements made in some directions may expose residual (or even precipitate new) shortcomings in other directions. For these reasons it would be difficult if not impossible in the situations described above to automate or develop a computer-based system for feedback or formative assessment, or for generating remedial moves and appropriate corrective procedures. Any attempt to mechanize such educational activities and creative efforts is unlikely to be successful because of the large number of variables involved, the intense relations often existing among them, and their essential fuzziness (Sadler, 1982). But the inability to mechanize a system that ordinarily depends heavily on qualitative judgments does not, of course, mean that such a system cannot be made to work. People frequently not only make, share, and broadly agree on qualitative judgments, but also use them as the basis for their own improvement. By definition, something which can be shown to occur is more than just a theoretical possibility, and it is common knowledge that a complex activity can be subject to a high degree of control even when the individual processes have not been comprehensively analyzed and are not fully understood.

The most readily available material for students to work on for evaluative and remedial experience is that of fellow students. Apart from availability, and provided steps are taken to ensure that mutual exchange does not cause friction or resentment or make weaker students feel threatened or humiliated, engaging in evaluative and corrective activity on other students' work has the advantages that (a) the work is of the same type and addressed to the same task as their own, (b) students are brought face to face with a wide range of moves or solutions to creative, design, and procedural problems, and exposure to these incidentally expands their own repertoire of moves, (c) other students' attempts normally cover a wide spectrum of imperfections, including global and particular inadequacies, and (d) the use of other students' work in a cooperative environment assists in achieving some objectivity in that students are less defensive of, and committed emotionally to, other students' work than to their own. A practical spin-off of the use of peer-appraisal is that it reduces the assessment workload for teachers. That traditional approaches to formative assessment are typically labour intensive partly explains teachers' reluctance to do much of it.

Constructively appraising the work of fellow learners is already established as part of normal teaching in some subjects and fields. Many teachers, for example, encourage their students to exchange work with one another in class. In particular, these principles are foundational to certain approaches to the teaching of writing, specifically reader-writer conferencing, peer review, and process writing. Students develop their pool of strategies by learning to revise and refine their own work in cooperation with the teacher, and by editing and helping other students to improve theirs (Beaven, 1977; Pianko and Radzik, 1980; Thompson, 1981; Chater, 1984). "Students who become conscious of what they're doing by explaining their decisions to other students also learn new strategies for solving writing problems. And because students should become progressively more independent and self-confident as writers, they need to evaluate each other's work and their own frequently, a practice which teaches constructive criticism, close reading, and rewriting" (Lindemann, 1982, p. 234). Boud (1986) reported similar findings in higher education when self-assessment and peer-assessment were built into instructional procedures for law, engineering and architecture students. It is clear that to build explicit provision for evaluative experience into an instructional system enables learners to develop self-assessment skills and gap-closing strategies simultaneously, and therefore to move towards self-monitoring. Some resistance to this proposition can, however, be expected.

Factors militating against self-monitoring

The lack of opportunities typically given to students to make appropriate qualitative judgements suggests an underlying assumption that only teachers have the skill and expertise to evaluate student work, and that this skill is not transferable to the students. Bloom's (1956) influential taxonomy places evaluation at the top of the hierarchy of cognitive skills, and some learning theorists hold that learners typically do not (and perhaps cannot) engage in high-level abstract thought while young. Although the exact position of evaluation in the Bloom hierarchy is debatable, it almost certainly requires abstract thinking and is situated above knowledge, comprehension, and application. This may give the impression that evaluation is some kind of esoteric activity engaged in only by adults or experts. If so, it ignores the fact that even children (certainly in their hours out of school) continually engage in evaluative activity and, if asked, can often produce rudimentary but reasonably sound rationales for their judgments.

Some teachers feel threatened by the idea that students should engage openly and cooperatively in making evaluative judgments. An assessment which results in a grade is used by many teachers as a tool for the control or modification of behaviour, for rewards and punishments. To remove some of the responsibility for assessment from teachers and place it in the hands of students may be considered to have the potential for undermining the teacher's authority. A less pathological concern is that many teachers perceive evaluation as the responsibility primarily of teachers because it constitutes part of the specialized knowledge and expertise that they have acquired as professionals. Assessment is regarded as strictly the teachers' prerogative: it sets them apart from their students and to some extent from parents and the rest of society. Part of the teacher's responsibility is surely, however, to download that evaluative knowledge so that students eventually become independent of the teacher and intelligently engage in and monitor their own development. If anything, the guild knowledge of teachers should consist less in knowing how to evaluate student work and more in *knowing ways to download evaluative knowledge* to students.

Apart from personal factors, formative assessment can be inhibited by certain circumstances outside the control of the teacher. School-based or internal examination systems often make use of so-called *continuous* (or *progressive*, or *periodic*) assessment. One of the arguments in favour of continuous assessment is that a series of assessments made over an extended period of time tends to reduce the high levels of anxiety experienced by some students under formal make-or-break examinations at the end of a course. (It may, of course, create a different form of stress.) Another argument is that continuous assessment permits wider and more varied sampling of a student's knowledge and skills. A third argument is that continuous assessment provides frequent feedback on progress. Continuous assessment cannot, however, function formatively when it is *cumulative*, that is, when each attempt or piece of work submitted by a student is scored and the scores are added together at the end of the course. This practice tends to produce in students the mindset that if a piece of work does not contribute towards the total, it is not worth doing. The longer-term goal of excellence may therefore be forfeited because of the drive to accumulate credit. Optional recycling of work for purposes of improvement becomes an unattractive proposition, and also raises the question of fairness to other students if a teacher works with some of the students (but perhaps not others) in helping to raise the standard of performance. Any work which is to form the basis for a course grade is normally expected, of course, to be produced by the student without aid from the teacher.

A further factor follows from the widespread policy of allocating course grades according to some predetermined statistical distribution. This is often considered to be the best or only practical method of maintaining standards. Such grading-on-the-curve, however, does not allow for the recognition of improvement in performance in absolute terms; it creates a zero-sum game, encourages competitiveness among students, and is inimical to the goal of genuine improvement for all students.

A final factor is associated with curriculum structure. There has been a trend over recent decades towards breaking up long courses into units or modules in the name of providing increased curriculum flexibility for students. Each unit is designed so that it can to a substantial extent stand alone, and each is taught over a single term or semester, or even a few weeks. Students compile a customized curriculum by putting together a collection of units. For purposes of formative assessment, the length of each unit is often not long enough for students to submit work, have it assessed, rework it in an effort to become proficient, and finally submit a different but well-produced piece for a grade. There is simply not the time to do it.

Conclusion

To improve their performance, students need to know how they are progressing. Feedback is commonly defined in terms of information given to the student about the quality of performance (knowledge of results). But in many educational and training contexts, students produce work which cannot be assessed simply as correct or incorrect. The quality of the work is determined by direct qualitative human judgment. The traditional definition of feedback is then too narrow to be of much use, and in this article a more appropriate conception is presented. It requires knowledge of the standard or goal, skills in making multicriterion comparisons, and the development of ways and means for reducing the discrepancy between what is produced and what is aimed for.

Improvement can, of course, occur if the teacher provides detailed remedial advice and the student follows it through. This, however, maintains the learner's dependence on the teacher. The alternative approach which is described and advocated in this article is for students to develop skills in evaluating the quality of their own work, especially during the process of production. The transition from teacher-supplied feedback to learner self-monitoring is not something that comes about automatically. For an important class of learning outcomes, the instructional system must make explicit provision for students themselves to acquire evaluative expertise. It is argued that providing direct and authentic evaluative experience is a necessary (instrumental) condition for the development of evaluative expertise and therefore for intelligent self-monitoring. It is insufficient for students to rely upon evaluative judgments made by the teacher.

The practices recommended are not radically new, and are already employed in some instructional systems. Empirically, they are known to produce results. What this article provides is a theoretical perspective on these practices, and an argument for their generalization to any instructional system designed to produce learner outcomes which are judged qualitatively using multiple criteria. The corollary is that not to design authentic evaluative experience into the instructional system either places an artificial performance ceiling on many students or limits their rate of learning.

References

Bailin, S. (1987). Creativity or quality: a deceptive choice. *Journal of Educational Thought, 21,* 33–39.

Beaven, M. H. (1977). Individualized goal-setting, self-evaluation, and peer evaluation. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: describing, measuring, judging.* Urbana, IL.: National Council of Teachers of English.

Black, H. D. (1986). Assessment for learning. In D. L. Nuttall (Ed.), *Assessing educational achievement* .London: Falmer.

Black, H. D. and Dockrell, W. B. (1984). *Criterion-referenced assessment in the classroom.* Edinburgh: Scottish Council for Research in Education.

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: Handbook 1, Cognitive domain. New York: David McKay.

Bloom, B. S., Madaus, G. F. and Hastings, J. T. (1981). *Evaluation to improve learning.* New York: McGraw-Hill.

Boud, D. (1986). Implementing student self-assessment. HERDSA Green Guide No.5. Kensington N.S.W.: Higher Education Research and Development Society of Australasia.

Chater, P. (1984). *Marking and assessment in English.* London: Methuen.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Daly, J. A. and Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement, 19,* 309–316.

Diederich, P. B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Elbow, P. (1973). *Writing without teachers.* New York: Oxford University Press.

Gere, A. R. (1980). Written composition: toward a theory of evaluation. *College English, 42*(1), 44–58.

Hales, L. W. and Tokar, E. (1975). The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement, 12,* 115–117.

Helson, H. (1959). Adaptation level theory. In S. Koch (Ed.), *Psychology: a study of a science. Volume 1: Sensory, perceptual and physiological formulations.* New York: McGraw-Hill.

Kaplan, A. (1964). *The conduct of inquiry: methodology for behavioral science*. San Francisco: Chandler.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.

Kulik, J. A. and Kulik, C-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*, 79–97.

Lindemann, E. (1982). *A rhetoric for writing teachers*. New York: Oxford University Press.

Locke, E. A., Shaw, K. N., Saari, L. M. and Latham, G. P. (1981). Goal setting and task performance: 1969–1980. *Psychological Bulletin, 90*, 125–152.

Marshall, M. S. (1958). This thing called evaluation. *Educational Forum, 23*, 41–53.

Marshall, M. S. (1968). *Teaching without grades*. Corvallis, Oregon: Oregon State University Press.

Myers, M. (1980). A procedure for writing assessment and holistic scoring. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills, National Institute of Education, and National Council of Teachers of English.

Nitko, A. J. (1983). *Educational tests and measurement: an introduction*. New York: Harcourt Brace Jovanovich.

Odell, L. and Cooper, C. R. (1980). Procedures for evaluating writing: assumptions and needed research. *College English, 42*(1), 35–43.

Pianko, S. and Radzik, A. (1980). The student editing method. *Theory into Practice, 19*, 220–224.

Polanyi, M. (1962). *Personal knowledge: towards a post-critical philosophy*. London: Routledge and Kegan Paul.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4–13.

Rowntree, D. (1977). *Assessing students: how shall we know them?* London: Harper and Row.

Sadler, D. R. (1981). Intuitive data processing as a potential source of bias in naturalistic evaluations. *Educational Evaluation and Policy Analysis, 3*(4), 25–31.

Sadler, D. R. (1982). Evaluation criteria as control variables in the design of instructional systems. *Instructional Science, 11*, 265–271.

Sadler, D. R. (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education, 54*, 60–79.

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educational Theory, 35*, 285–297.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 191–209.

Shenstone, W. (1768). On writing and books, LXXIX. In *Works: In verse and prose Vol. 2, (3rd ed.)*. London: Dodsley.

Thompson, R. F. (1981). Peer grading: some promising advantages for composition research and the classroom. *Research in the Teaching of English, 15*, 172–174.

Thorndike, E. L. (1913). *Educational Psychology, Vol.1: The original nature of man*. New York: Teachers College, Columbia University.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31–48.

Wittgenstein, L. (1974). *Philosophical investigations*. (G.E.M. Anscombe, Trans.). Oxford: Basil Blackwell. (Original work: 3rd ed. published 1967).