Chapter 4

Transforming Holistic Assessment and Grading into a Vehicle for Complex Learning

D. Royce Sadler

Introduction

One of the themes running through my work since 1980 has been that students need to develop the capacity to monitor the quality of their own work during its actual production. For this to occur, students need to appreciate what constitutes work of higher quality; to compare the quality of their emerging work with the higher quality; and to draw on a store of tactics to modify their work as necessary. In this chapter, this theme is extended in two ways. The first is an analysis of the fundamental validity of using preset criteria as a general approach to appraising quality. The second is a teaching design that enables holistic appraisals to align pedagogy with assessment.

For the purposes of this chapter, a *course* refers to a unit of study that forms a relatively self-contained component of a degree program. A student response to an assessment task is referred to as a *work*. The assessed quality of each work is represented by a numerical, literal or verbal mark or *grade*. Detailed feedback from the teacher may accompany the grade. For the types of works of interest in this chapter, grades are mostly produced in one of two ways.

In *analytic* grading, the teacher makes separate qualitative judgments on a limited number of properties or *criteria*. These are usually *preset*, that is, they are nominated in advance. Each criterion is used for appraising each student's work. The teacher may prescribe the criteria, or students and teachers may negotiate them. Alternatively, the teacher may require that students develop their own criteria as a means of deepening their involvement in the assessment process. In this chapter, how the criteria are decided is not important. After the separate judgments on the criteria are made, they are combined using a rule or formula, and converted to a grade. Analytic grading is overtly systematic. By identifying the specific elements that contribute to the final grade, analytic grading provides the student with explicit feedback. The template used in implementing the process may be called a *rubric*, or any one of *scoring*, *marking* or *grading* paired with *scheme*, *guide*, *matrix* or *grid*. As a group, these models are sometimes referred to as *criterion-based* assessment or *primary trait analysis*.

In *holistic* or global grading, the teacher responds to a student's work as a whole, then directly maps its quality to a notional point on the grade scale. Although the teacher may note specific features that stand out while appraising, arriving directly at a global judgment is foremost.

Reflection on that judgment gives rise to an explanation, which necessarily refers to criteria. Holistic grading is sometimes characterised as impressionistic or intuitive.

The relative merits of analytic and holistic grading have been debated for many years, at all levels of education. The most commonly used criterion for comparison has been *scorer reliability*. This statistic measures the degree of consistency with which grades are assigned to the same set of works by different teachers (inter-grader reliability), or by the same teacher on separate occasions (temporal reliability). Scorer reliability is undoubtedly a useful criterion, but is too narrow on its own. It does not take into account other factors such as the skills of the markers in each method, or the extent to which each method is able to capture all the dimensions that matter.

The use of analytic grading schemes and templates is now firmly established in higher education. Internationally, rapid growth in popularity has occurred since about 1995. Nevertheless, the basic ideas are not new. Inductively decomposing holistic appraisals goes back at least to 1759, when Edmund Burke set out to identify the properties that characterise beautiful objects in general. In the forward direction, the theory and practice of assembling overall judgments from discrete appraisals on separate criteria has developed mostly over the last 50 years. It has given rise to an extensive literature touching many fields. Key research areas have been clinical decision making (Meehl, 1954/1996) and human expertise of various types (Chi, Glaser & Farr, 1988; Ericsson & Smith, 1991). The terminology used is diverse, and includes ‘policy capturing’ and ‘actuarial methods’.

Specifically in educational assessment, Braddock et al. (1963) reported early developmental work on analytic approaches to grading English composition, and the rationale for it; Primary Trait scoring is described in Lloyd-Jones (1977). Researchers in higher education assessment have explored in recent years the use of criteria and rubrics, specifically involving students in self- and peer-assessment activities (Bloxham, 2004; Orsmond, 2000; Rust et al., 2003; Woolf, 2004). Many books on assessment in higher education advocate analytic grading, and provide practitioners with detailed operational guidelines. Examples are Freeman & Lewis (1998); Huba & Freed (2000); Morgan et al. (2004); Stevens & Levi (2004); Suskie (2004); and Walvoord & Anderson (1998).

For the most part, both the underlying principles and the various methods of implementation have been accepted uncritically. In this chapter, the sufficiency of analytic grading as a general approach for relevant classes of student works is called into question, on both theoretical and practical grounds. The basic reason is that they set up appraisal frameworks that are, in principle, sub-optimal. Although they work adequately for some grading decisions, they do a disservice to others by unnecessarily constraining the scope of appraisals. The assumption that using preset criteria is unproblematic has had two inhibiting effects. First, teachers typically have not felt free to acknowledge, especially to students, the existence or nature of certain limitations they encounter. Second, there has been little or no imperative to explore and develop alternative ways forward.

The theme of this chapter is developed around five propositions. The first four are dealt with relatively briefly; the fifth is assigned a section of its own. The driving principle is that if students are to achieve consistently high levels of performance, they need to develop a conceptualisation of what constitutes ‘quality’ as a generalised attribute (Sadler, 1983). They also need to be inducted into evaluating quality, without necessarily being bound by tightly specified criteria. This approach mirrors the way multi-criterion judgments are typically made by experienced teachers. It is also an authentic representation of the ways many appraisals are made in a host of everyday contexts by experts and non-experts alike. Equipping students with evaluative insights and skills therefore contributes an important graduate skill. All five propositions are taken into account in the second half of the chapter, which outlines an approach to the assessment of complex student productions.

Applicable Types of Assessment Tasks

The types of tasks to which this chapter applies are those that require divergent or ‘open’ responses from students. Divergent tasks provide opportunities for learners to demonstrate sophisticated cognitive abilities, integration of knowledge, complex problem solving, critical reasoning, original thinking, and innovation. Producing a response requires abilities in both design and production, allowing considerable scope for creativity. There are no formal techniques or recipes which, if followed precisely, would lead to high-quality responses. There is also no single correct or best answer, result or solution.

Common formats for divergent responses include field and project reports, seminar presentations, studio and design productions, specialised artefacts, professional performances, creative works, term papers, essays, and written assignments. In assessing achievement across a broad range of disciplines and professions, divergent responses predominate. Within each genre, student works may take quite different forms, yet be of comparable quality. This characteristic is regarded as highly desirable in many domains of higher education.

Determining the quality of divergent types of works requires skilled, qualitative judgments using multiple criteria. A qualitative judgment is one made directly by a person, the person's brain being both the source and the instrument for appraisal (Sadler, 1989). The judgment cannot be reduced to a set of measurements or formal procedures that lead to the ‘correct’ appraisal. Qualitative judgments are unavoidable in many fields of higher education, and both holistic and analytic grading are based on them. The two approaches differ primarily in their granularity. Holistic grading involves appraising student works as integrated entities; analytic grading requires criterion-by-criterion judgments.

Historically, a steady swing has occurred away from holistic and towards analytic judgments, and then a further trend has occurred within analytic judgments. When scoring guides and marking schemes first became common, the focus tended to be on either the inclusion or omission of specific content, or the structure of the response. For a written piece, this structure could be Introduction, Statement of the problem, Literature review, Development of an argument or position, and Conclusion. The subsequent shift has concentrated on properties or dimensions related to quality. Regardless of focus, all analytic grading schemes introduce formal structure into the grading process, ostensibly to make it more objective and thus reduce the likelihood of favouritism or arbitrariness.

The First Four Propositions

Already referred to briefly above, *Proposition 1* is that students need to develop the capacity to monitor the quality of their work during its actual production (Sadler, 1989). In relation to creating responses to assessment tasks, this capability needs to be acquired as a course proceeds. Teaching therefore needs to be designed so as to make specific provision for its development. As the learning sequence progresses, students’ understanding of quality needs not only to grow but also to become broadly consonant with that held by the teacher. This is partly because the teacher usually has a strong say in the final grade, and partly because the teachers’ feedback does not make much sense otherwise. But there are deeper implications. Ultimately, the concept of quality needs to relate to works that graduates will produce after their formal studies are completed, as they demonstrate professional expertise. This implies, therefore, that the teacher’s frame of reference about quality

should reflect the conventions and expectations evident in other relevant environments such as the arts and professions, industry, and elsewhere in academia.

Self-monitoring means that students make conscious judgments on their own, without help from teachers or peers. It entails being weaned away from ongoing dependence on external feedback, irrespective of its source or character. Among other things, self-monitoring requires an appreciation of what makes a work of high quality. It also requires enough evaluative skill to compare, with considerable detachment, the quality of what the producer is creating with what would be needed for it to be of high quality. For self-monitoring to have any impact on an emerging work, the student also needs a repertoire of alternative moves upon which to draw at any pertinent point or stage in the development. Otherwise the work cannot be improved. This in turn necessitates that the student become sensitive to where those 'pertinent points' are, as they arise, during construction.

Many of the students whose usual levels of performance are mediocre are hampered by not knowing what constitutes work of high quality. This sets an upper bound on their ability to monitor the quality of their own developing work (Sadler, 1983). Raising students' knowledge about high quality and their capability in self-monitoring can lead to a positive chain of events. These are improved grades, increased intrinsic satisfaction, enhanced motivation and, as a consequence, higher levels of achievement. Within the scope of a single course, it is obviously not realistic to expect learners to become full connoisseurs. But as a course proceeds, the learners' judgments about the quality of their own works should show progressively smaller margins of error. Self-monitoring raises self-awareness and increases the learner's metacognition of what is going on. Attaining this goal is not intrinsically difficult, but it does require that a number of specific conditions be met. Not to achieve the goal, however, represents a considerable opportunity loss.

Proposition 2 is that students can develop evaluative expertise in much the same way as they develop other knowledge and skills, including the substantive content of a course. Skilled appraisal is just one of many types of expertise, although it seldom features explicitly among course objectives. Initially, a key tool for developing it is credible feedback, primarily from the teacher and peers. Feedback usually takes the form of descriptions, explanations or advice expressed in words. Preset criteria coupled with verbal feedback stem from a desire to tell or inform students. It might be thought that the act of telling would serve to raise the performance ceiling for learners, but just being told is rarely an adequate remedy for ignorance. The height of the ceiling depends on what students make of what is 'told' to them.

The next step along the path can be taken when relevant examples complement the verbal descriptions (Sadler, 1983, 1987, 1989, 2002). Examples provide students with concrete referents. Without them, explanatory comments remain more or less abstract, and students cannot interpret them with certainty. If the number of examples is small, they need to be chosen as judiciously as examples are for teaching. If examples are plentiful, careful selection is not as critical, provided they cover a considerable range of the quality spectrum. Even more progress can be made if teachers and learners actively discuss the descriptions and exemplars together. Reading verbal explanations, seeing pertinent exemplars, and engaging in discourse provide categorically different cognitive inputs. But the best combination of all three still does not go far enough.

The remaining element is that students engage in making evaluative decisions themselves, and justifying those decisions. No amount of telling, showing or discussing is a substitute for one's own experience (Sadler, 1980, 1989). The student must learn how to perceive works essentially through the eyes of an informed critic, eventually becoming 'calibrated'. Learning environments are self-limiting to the extent that they fail to make appropriate provision for students to make, and be accountable for, serious appraisals.

Proposition 3 is that students' direct evaluative experience should be relevant to their current context, not translated from another. The focus for their experience must therefore be works of a genre that is substantially similar to the one in which they are producing. Apart from learning about quality, closely and critically examining what others have produced in addressing assessment tasks expands the student's inventory of possible moves. These then become available for drawing upon to improve students' own work. This is one of the reasons peer assessment is so important.

However, merely having students engage in peer appraisal in order to make assessment more participatory or democratic is not enough. Neither is treating students as if they were already competent assessors whose appraisals deserve equal standing with those of the teacher, and should therefore contribute to their peers' grades. The way peer assessment is implemented should reflect the reasons for doing it. Learners need to become reasonably competent not only at assessing other students' works but also at applying that knowledge to their own works.

Proposition 4 is that the pedagogical design must function not only effectively but also efficiently for both teachers and students. There is obviously little point in advocating changes to assessment practices that are more labour intensive than prevailing procedures. Providing students with sufficient direct evaluative experience can be time consuming unless compensating changes are made in other aspects of the teaching. This aspect is taken up later in the chapter.

The Fifth Proposition

Ideally, students should learn how to appraise complex works using approaches that possess high scholarly integrity, are true to the ways in which high-quality judgments are made professionally, and have considerable practical potential for improving their own learning. *Proposition 5* is that students in many higher education contexts should learn how to make judgments about the quality of emerging and finished works holistically rather than using analytic schemes. The case for this proposition is developed by first analysing the rationale for analytic judgments, and then mounting a critique of the method generally.

In recent years, analytic grading schemes using preset criteria have been advocated as superior to holistic appraisals. The rationale for this is more often implied than stated. Basically, it is that such systems:

- a) Improve consistency and objectivity in grading, because the appraisal process is broken down into smaller-scale judgments;
- b) Make transparent to students, as an ethical obligation, the key qualities that will be taken into account;
- c) Encourage students to attend to the assessment criteria during development of their work, so the criteria can play a product-design role which complements the assessment task specifications;
- d) Enable grading decisions to be made by comparing the quality of a student's work with fixed criteria and 'standards' rather than to the learner's previous level of achievement, the performance of others in the class, or the teacher's personal tastes or preferences; and
- e) Provide accurate feedback more efficiently, with less need for the teacher to write extensive comments.

These arguments appear sound and fair. Who would argue against an assessment system that provides more and better feedback, increases transparency, improves accountability, and achieves

greater objectivity, all with no increase in workload? On the other hand, the use of preset criteria accompanied by a rule for combining judgments is not the only way to go. In the critique below, a number of issues that form the basis of Proposition 5 are set out. The claim is that no matter how comprehensive and precise the procedures are, or how meticulously they are followed, they can, and for some student works do, lead to deficient or distorted grading decisions. This is patently unfair to those students.

In the rest of this section, the case for holistic judgments is presented in considerable detail. Although any proposal to advocate holistic rather than analytic assessments might be viewed initially as taking a backward step, this chapter is underwritten by a strong commitment to students and their learning. Specifically, the aim is to equip students to work routinely with holistic appraisals; to appreciate their validity; and to use them in improving their own work. The five clauses in the rationale above are framed specifically in terms of criteria and standards. This wording therefore presupposes an analytic model for grading. By the end of this chapter, it should become apparent how the ethical principles behind this rationale can be honoured in full through alternative means. The wording of the rationale that corresponds to this alternative would then be different.

Beginning of the Case

Whenever a specific practice becomes widespread in a field of human activity, each implementation of it contributes to its normalisation. The message that this practice is the only or preferred approach does not have to be communicated explicitly. Consistent uncritical use sends its own strong signals. In this section, two particular analytic assessment schemes, *analytic rating scales* and *analytic rubrics*, are singled out for specific attention. Both are common in higher education, and both are simple for students to understand and apply.

With analytic rating scales, multiple criteria (up to 10 or more for complex works) are first specified. Each criterion has an associated scale line defined. In use, the appraiser makes a qualitative judgment about the ‘strength’ or level of the work on each criterion, and marks the corresponding point on the scale line. If a total score is required, the relative importance of each criterion is typically given a numerical weighting. The sub-scores on the scales are multiplied by their respective weightings. Then all the weighted scores are summed, and the aggregate either reported directly or turned into a grade using a conversion table.

An analytic rubric has a different format. Using the terminology in Sadler (1987, 2005), a rubric is essentially a matrix of cross-tabulated criteria and ‘standards’ or levels. (This nomenclature is not uniform. Some rubrics use ‘qualities/criteria’ instead of ‘criteria/standards’ as the headings.) Each ‘standard’ represents a particular level on one criterion. Common practice is for the number of standards to be the same for all criteria, but this is not strictly necessary. A rubric with five criteria, each of which has four standards, has a total of 20 cells. Each cell contains a short verbal description that sets out a particular ‘strength’ of the work on the corresponding criterion. This is usually expressed either as a ‘verbal quantifier’ (how much), or in terms of sub-attributes of the main criterion. For each student work, the assessor identifies the single cell for each criterion that best seems to characterise the work. The rubric may also include provision for the various cell ‘standards’ to carry nominated ranges of numerical values that reflect weightings. A total score can then be calculated and converted to a grade.

Holistic rubrics form a less commonly used category than the two above. They associate each grade level with a reasonably full verbal description, which is intended as indicative rather than

definitive or prescriptive. These descriptions do not necessarily refer to the same criteria for all grade levels. Holistic rubrics are essentially different from the other two and have a different set of limitations, particularly in relation to feedback. They are not considered further here.

In most analytic schemes, there is no role for distinctly global primary appraisals. At most, there may be a criterion labelled 'overall assessment' that enjoys essentially the same status as all the other criteria. Apart from the technical redundancy this often involves, such a concession to valuing overall quality fails to address what is required. Students therefore learn that global judgments are normally compounded from smaller-scale judgments. The learning environment offers regular reinforcement, and the stakes for the student are high.

Grading methods that use preset criteria with mechanical combination produce anomalies. The two particular anomalies outlined below represent recurring patterns, and form part of a larger set that is the subject of ongoing research. These anomalies are detected by a wide range of university teachers, in a wide range of disciplines and fields, for a wide range of assessment task types and student works. The same anomalies are, however, invisible to learners, and the design of the appraisal framework keeps them that way. Therein lies the problem.

Anomaly 1

Teachers routinely discover some works for which global impressions of their quality are categorically at odds with the outcomes produced by conscientious implementation of the analytic grading scheme. Furthermore, the teacher is at a loss to explain why. A work which the teacher would rate as 'brilliant' overall may not be outstanding on all the preset criteria. The whole actually amounts to more than the sum of its parts. Conversely, a work the teacher would rate as mediocre may come out extremely well on the separate criteria. For it, the whole is less than the sum of its parts. This type of mismatch is not confined to educational contexts.

Whenever a discrepancy of this type is detected, teachers who accept as authoritative the formula-based grade simply ignore their informal holistic 'appraisal' without further ado. Other teachers react differently. They question why the analytic grade, which was painstakingly 'built' from component judgments, fails to deliver the 'true' assessment, which they regard as the holistic judgment. In so doing, they implicitly express more confidence in the holistic grade than in the analytic. To reconcile the two, they may adjust the reported levels on the criteria until the analytic judgment tells the story it 'should'. At the same time, these teachers remain perplexed about what causes such anomalies. They feel especially troubled about the validity and ethics of what could be interpreted by others as fudging. For these reasons, they are generally reluctant to talk about the occurrence of these anomalies to teaching colleagues or to their students. However, many are prepared to discuss them in secure research environments.

What accounts for this anomaly? There are a number of possible contributing factors. One is that not all the knowledge a person has, or that a group of people share, can necessarily be expressed in words (Polanyi, 1962). There exists no theorem to the effect that the domain of experiential or tacit knowledge, which includes various forms of expertise, is co-extensive and isomorphic with the domain of propositional knowledge (Sadler, 1980). Another possible factor is the manner in which experts intuitively process information from multiple sources to arrive at complex decisions, as summarised in Sadler (1981). Certain holistic appraisals do not necessarily map neatly on to explicit sets of specified criteria, or simple rules for combination. Further exploration of these aspects lies outside the scope of this chapter. It would require tapping into the extensive literature on the processes of human judgment, and into the philosophy of so-called ineffable knowledge.

Anomaly 2

This anomaly is similar to the first in that a discrepancy occurs between the analytically derived grade and the assessor's informal holistic appraisal. In this case, however, the teacher knows that the source of the problem is with a particular criterion that is missing from the preset list. That criterion may be important enough to set the work apart, almost in a class of its own. To simplify the analysis, assume there is just one such criterion. Also assume that the teacher has checked that this criterion is not included on the specified list in some disguised form, such as an extreme level on a criterion that typically goes under another name, or some blend of several criteria. Strict adherence to the analytic grading rule would disallow this criterion from entering, formally or informally, into either the process of grading or any subsequent explanation. To admit it would breach the implicit contract between teacher and student that only specified criteria will be used.

Why do sets of criteria seem comprehensive enough for adequately appraising some works but not others? Part of the explanation is that, for complex works, a specified set of criteria is almost always a selection from a larger pool (or 'population') of criteria. By definition, a sample does not fully represent a population. Therefore, applying a fixed sample of criteria to all student works leaves open the possibility that some works may stand out as exceptional on the basis of unspecified criteria. This applies particularly to highly divergent, creative or innovative responses. Arbitrarily restricting the set of criteria for these works introduces distortion into the grading process, lowering its validity.

To illustrate this sampling phenomenon, consider a piece of written work such as an essay or term paper, for which rubrics containing at least four criteria are readily available. Suppose the teacher's rubric prescribes the following criteria:

- relevance
- coherence
- presentation
- support for assertions.

Behind these sits a much larger pool of potentially valid criteria. An idea of the size of this pool can be obtained by analysing a large number of published sets. One such (incomplete) collection was published in Sadler (1989). In alphabetical order, the criteria were:

accuracy (of facts, evidence, explanations); audience (sense of); authenticity; clarity; coherence; cohesion; completeness; compliance (with conventions of the genre); comprehensiveness; conciseness (succinctness); consistency (internal); 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; transitions; usage; vocabulary; voice; wording.

The majority of these would be familiar to experienced teachers, but dealing with them as separate properties is not at all straightforward. In the abstract, the listed criteria may appear to represent distinct qualities. When they come to be applied, however, some seem to merge into others. The reasons are manifold. First, they are uneven in their scope, some being broad, others narrow. Second, even common criteria often lack sharp boundaries and standardised interpretations, being defined differently by different assessors. When their meanings are probed, vigorous debate usually

ensues. Some interpretations are contextually dependent, being defined differently by the same teacher for different assessment tasks. Third, some are subtle or specialised. An example is ‘flair’, which is relevant to writing and many other creative arts, and appears to capture a special and valued characteristic that is hard to describe. Fourth, some criteria are effectively alternatives to, or nested within, others. In addition, one cluster of criteria may have the same ‘coverage’ as another cluster, but without one-to-one correspondence. Finally, suppose it were possible to assemble and clarify the whole population of relevant criteria. Any attempt to use them all would be unworkable for assessors and students alike. The obvious way out of this bind is to restrict the list of criteria to a manageable number of, say, the most important, or the most commonly used. In contexts similar to that of written works, such restriction necessarily leaves out the majority.

A Way Forward: Design for Teaching and Learning

Should the teacher disclose the existence of anomalies to students and, in so doing, expose the weaknesses of analytic grading? A substantial part of the rationale for prior specification is that students are entitled to have advance knowledge of the basis for their teachers’ appraisals. Traditionally, holistic grading allowed the teacher to keep the reasons more or less private. Given the anomalies above, it is ironic that strictly honouring a commitment to preset criteria can be achieved only at the expense of non-disclosure of anomalies and other limitations once the grading is completed.

The two anomalies above, along with other deficiencies not covered in this chapter, raise doubts about the sufficiency of relying on analytic frameworks that use prespecified criteria. These difficulties are structural, and cannot be dealt with by making templates more elaborate. If there is to be a way forward, it has to approach the problem from a different direction. The rest of this chapter is devoted to outlining a design that has dual characteristics. It seeks to reinstate holistic appraisals for grading a large and significant class of complex student works. It also allows for at least some criteria to be specified in advance – without grade determination being dependent on following an inflexible algorithm. Although this is a step back from grading all responses to an assessment task by fixing both the criteria and the combination formula, the commitment to validity, openness and disclosure can remain undiminished. This goal can be achieved by shifting the agenda beyond telling, showing and discussing how appraisals of students’ works are to be made. Instead, the process is designed to induct students into the art of making appraisals in a substantive and comprehensive way. Such a pedagogical approach is integral to the development of a valid alternative not only to analytic grading but also to holistic grading as done traditionally.

This induction process can function as a fundamental principle for learning-oriented assessment. Properly implemented, it recognises fully the responsibility of the teacher to bring students into a deep knowledge of how criteria actually function in making a complex appraisal, and of the need for the assessor to supply adequate grounds for the judgment. In the process, it sets learners up for developing the capability to monitor the quality of their own work during its development. In particular, it provides for an explicit focus on the quality of how the student work is coming together – as a whole – at any stage of development.

The approach outlined below draws upon strategies that have been trialled successfully in higher education classes along with others that are worthy of further exploration and experimentation. Major reform of teaching and learning environments requires significant changes to approaches that have become deeply embedded in practice over many years. Because the use of preset criteria has gained considerable momentum in many academic settings, transition problems are to be expected. Some ideas for identifying and managing these are also included.

Developing Expertise

The general approach draws from the work of Polanyi (1962). It could be described as starting learners on the path towards becoming connoisseurs. There is a common saying that goes, 'I do not know how to define quality, but I know it when I see it'. In this statement, to 'know' quality is to recognise, seize on, or apprehend it. To recognise quality 'when I see it' means that I can recognise it – in particular cases. The concrete instance first gives rise to perception, then to recognition. Formally defining quality is a different matter altogether. Whether it is possible for a particular person to construct a definition depends on a number of factors, such as their powers of abstraction and articulation. There is a more general requirement, not limited to a particular person. Are there enough similarly classified cases that share enough key characteristics to allow for identification, using inductive inference, of what they have in common? The third factor is touched on in the account of the first anomaly described above. It is that some concepts appear to be, in principle, beyond the reach of formal definition. Many such concepts form essential elements of our everyday language and communication, and are by no means esoteric. Given these three factors, there is no logical reason to assume that creating a formal definition is always either worth the effort, or even possible.

On the other hand, it is known empirically that experts can recognise quality (or some other complex characteristic) independently of knowing any definition. This is evidence that recognition can function as a fundamentally valid, primary act in its own right (Dewey, 1939). It is also what makes connoisseurship a valuable phenomenon, rather than one that is just intriguing. Basically, connoisseurship is highly developed form of competence in qualitative appraisal. In many situations, the expert is able to give a comprehensive, valid and carefully reasoned explanation for a particular appraisal, yet is unable to do so for the general case. In other situations, an explanation for even a particular case is at best partial. To accept recognition as a primary evaluative act opens the door to the development of appraisal explanations that are specifically crafted for particular cases without being constrained by predetermined criteria that apply to all cases. Holistic recognition means that the appraiser reacts or responds (Sadler, 1985), whereas building a judgment up from discrete decisions on the criteria is rational and stepwise. That is the key distinction in intellectual processing. The idea of recognising quality when it is observed immediately strikes a familiar chord with many people who work in both educational and non-educational contexts where multi-criterion judgments are required. Giving primacy to an overall assessment has its counterparts in many other fields and professions.

This does not imply that grading judgments should be entirely holistic or entirely analytic, as if these were mutually exclusive categories. The two approaches are by no means incompatible, and how teachers use them is a matter of choice. To advocate that a teacher should grade solely by making global judgments without reference to any criteria is as inappropriate as requiring all grades to be compiled from components according to set rules. Experienced assessors routinely alternate between the two approaches in order to produce what they consider to be the most valid grade. This is how they detect the anomalies above, even without consciously trying. In doing this, they tend to focus initially on the overall quality of a work, rather than on its separate qualities. Among other things, these assessors switch focus between global and specific characteristics, just as the eye switches effortlessly between foreground (which is more localised and criterion-bound) and background (which is more holistic and open). Broader views allow things to be seen in

perspective, often with greater realism. They also counter the atomism that arises from breaking judgments up into progressively smaller elements in a bid to attain greater precision.

Inducting students into the processes of making multi-criterion judgments holistically, and only afterwards formulating valid reasons for them, requires a distinctive pedagogical environment. The end goal is to bring learners at least partly into the guild of professionals who are able to make valid and reliable appraisals of complex works using all the tools at their disposal. Among other things, students need to learn to run with dual evaluative agendas. The first involves scoping the work as a whole to get a feel for its overall quality; the second is to pay attention to its particular qualities.

In some learning contexts, students typically see only their own works. Such limited samples cannot provide a sufficient basis for generalisations about quality. For students to develop evaluative expertise requires that three interconnected conditions be satisfied. First, students need exposure to a wide variety of authentic works that are notionally within the same genre. This is specifically the same genre students are working with at the time. The most readily available source of suitable works consists of responses from other students attempting the same assessment task. As students examine other students' works through the lens of critique, they find that peer responses can be constructed quite differently from one another, even those that the students themselves would judge to be worth the same grade. They also discover, often to their surprise, works that do not address the assessment task as it was specified. They see how some of these works could hardly be classified as valid 'responses' to the assessment task at all, if the task specifications were to be taken literally. This phenomenon is common knowledge among higher education teachers, and is a source of considerable frustration to them.

The second condition is that students need access to works that range across the full spectrum of quality. Otherwise learners have difficulty developing the concept of quality at all. To achieve sufficient variety, the teacher may need to supplement student works from the class with others from outside. These may come from different times or classes, or be specially created by the teacher or teaching colleagues. The third condition is that students need exposure to responses from a variety of assessment tasks. In a fundamental sense, a developing concept of 'quality' cannot be entirely specific to a particular assessment task.

In the process of making evaluations of successive works and constructing justifications for qualitative judgments, fresh criteria emerge naturally. The subtlety and sophistication of these criteria typically increase as evaluative expertise develops. A newly relevant criterion is drawn – on demand – from the larger pool of background or 'latent' criteria. It is triggered or activated by some property of a particular work that is noteworthy, and then added temporarily to the working set of 'manifest' criteria (Sadler, 1983). The work may exhibit this characteristic to a marked degree, or to a negligible degree. On either count, its relevance signals that it needs to be brought into the working set. As latent criteria come to the fore, they are used in providing feedback on the grade awarded.

Latent criteria may also be shared with other students in the context of the particular work involved. Starting with a small initial pool of criteria, extending it, and becoming familiar with needs-based tapping into the growing pool, becomes a key part of the pedagogical design. This practice expands students' personal repertoires of available criteria. It also reinforces the way criteria are translated from latent to manifest, through appraising specific works. As additional criteria need to be brought into play, a class record of them may be kept for interest, but not with a view to assembling a master list. This is because the intention is to provide students with experience in the latent-to-manifest translation process, and the limitations inherent in using fixed sets of criteria. Students should then come to understand why this apparently fluid approach to

judging quality is not unfair or some sort of aberration. It is, in a profound sense, rational, normal and professional.

Although it is important that students appreciate why it is not always possible to specify all the criteria in advance, certain criteria may always be relevant. In the context of written works, for example, some criteria are nearly always important, even when not stated explicitly. Examples are grammar, punctuation, referencing style, paragraphing, and logical development, sometimes grouped as ‘mechanics’. These relate to basic communicative and structural features. They facilitate the reader’s access to the creative or substantive aspects, and form part of the ‘craft’ or technique side of creative work. Properly implemented, they are enablers for an appraisal of the substance of the work. A written work is difficult to appraise if the vocabulary or textual structure is seriously deficient.

In stark contrast to the student’s situation, teachers are typically equipped with personal appraisal resources that extend across all of the aspects above. Their experience is constantly refreshed by renewed exposure to a wide range of student works, in various forms and at different levels of quality (Sadler, 1998). This is so naturally accepted as a normal part of what being a teacher involves that it hardly ever calls for comment. It is nevertheless easy to overlook the importance of comparable experience for students grappling with the concept of quality.

Providing direct evaluative experience efficiently is a major design element. It is labour intensive, but offsets can be made in the way teaching time is deployed. Evaluative activity can be configured to be the primary pedagogical vehicle for teaching a considerable proportion of the substantive content of a course. For example, in teaching that is structured around a lecture-tutorial format, students may create, in their own time, responses to a task that requires them to employ specific high-order intellectual skills such as extrapolating, making structural comparisons, identifying underlying assumptions, mounting counter-arguments, or integrating elements. These assessment tasks should be strictly formative, and designed so that students can respond to them successfully only as they master the basic content. Tutorial time is then spent having students make appraisals about the quality of, and providing informed feedback on, multiple works of their peers, and entering into discussions about the process. In this way, student engagement with the substance of the course takes place through a sequence of ‘produce and appraise’ rather than ‘study and learn’ activities. Adaptations are possible for other modes of teaching, such as studio and online.

Challenges of Transition

Complex learning, regardless of the field, requires multiple attempts (practice), in a supportive, low-stakes environment, with good feedback. To make progress on the road to connoisseurship is to replace initially inconsistent degrees of success in appraisal, and therefore self-monitoring, by progressively higher levels of expertise. The pedagogical environment in many higher education institutions is not set up in ways that facilitate this type of learning. The necessary changes cut across a number of well-established policies and practices, and therefore require careful management. Six obstacles or potential sources of resistance are identified below. The first three are outlined only; the second three are discussed in more detail.

Obstacle 1 is a view held by many students and teachers, overtly or by default, that virtually every course exercise or requirement should contribute towards the course grade. Unless something counts, so the thinking goes, it is not worth doing. Once this climate is established, teachers know that students will put little effort into any exercise that does not carry credit towards the course grade. Student and teacher positions reinforce each other, setting up a ‘credit accumulation economy’. The

development of evaluative expertise requires production, peer appraisal and peer feedback in a context where there is neither credit nor penalty for trial and error, experimentation, or risk taking. If these are to become normalised as legitimate processes in learning, ways have to be found to subvert the credit accumulation economy.

Obstacle 2 arises with teachers who, on occasion, use grading for purposes other than reporting exclusively on each student's achievement. Teachers may inflate a grade to reward a student who has made an outstanding effort, or who has shown a marked improvement. Such rewards compromise the meaning of grades and retard the development of evaluative expertise. Students and faculty have to be hard-nosed in their focus on quality alone, including the match between the assessment task specifications and the nature of the student responses. They should therefore rigorously exclude all non-achievement influences from the assessment environment.

Obstacle 3 arises through institutional policy, grading practices in other courses, or both. If all or most other courses use rubrics, for example, students may be wary of any grading method that appears to be unsystematic, subjective or unfair. Students may also be reluctant to engage in peer appraisal unless they have a rubric in front of them. Many students have internalised the principle that using preset criteria is automatically the only or the best way to appraise complex outcomes. Such conditioning has to be explicitly unlearned (Sadler, 1998). Some ground may be made by explaining the role of appraisal in learning to produce high-quality works. To develop expertise in this domain follows the same basic principles that are used in developing other complex skills or forms of expertise. These involve gathering or receiving information (being told), processing that information in the light of actual examples (being shown), and applying the principles for oneself (doing).

Obstacles 4 - 6 refer specifically to the appraisal processes described in this chapter, and are conceptually connected. They are expressed particularly strongly in some cultures. Obstacle 4 is the belief that appraisal or grading is a teacher role, not a student role. The teacher is perceived not only as the authoritative figure for course content but also as the only person who has the knowledge and experience to assign grades. Student peers are simply not qualified; they may judge too harshly or too leniently, or give feedback that is superficial and lacking in credibility.

Obstacle 5 has to do with the students' perceptions of themselves. They may feel ill-equipped to grade the works of peers. This is true initially, of course, but changing both the actuality and the self-perception are two of the course goals. The rationale for using other students' works is not just one of convenience. It ensures that the student responses, which are used as raw material, are as authentic as it is possible to get. This case may be easier to establish with students if the Course Outline includes an objective specifically to that effect, and a complementary statement along the following lines:

The learning activities in this course involve self- and peer-assessment. This is an important part of how the course is taught, and how one of its key objectives will be attained. Other students will routinely be making judgments about the quality of your work, and you will make judgments about theirs. The students whose work you appraise will not necessarily be those who appraise yours. If this aspect of the teaching is likely, in principle, to cause you personal difficulty, please discuss your concern with the person in charge of the course within the first two weeks of term.

Obstacle 6 is students' fear of exposure, loss of face or impending sense of humiliation among their peers. This may be because they lack experience, status or skill. These feelings are personal, about themselves and about how confident they are. Unless students are already familiar with engaging in peer assessment, they may appear to accept the logic behind a transition to a different pedagogy, but retain their reservations and reluctance. Such students need reassurance that just starting on this

path is likely to be the hardest part. Once they become accustomed to it, they typically find it highly rewarding, and their learning improves.

Learning the skills of appraisal and self-monitoring can be compared with the ways in which many other skills are learned. They are not easy for the uninitiated, and learners may feel embarrassed at their early attempts if they know others will become aware of their efforts. By contrast, when young children are learning to speak, their first bumbling attempts to say a few words are not treated with scorn. The children are encouraged and cheered on whenever they get it right. They then repeat the performance, maybe again and again. Translate this into the higher education context. As soon as students realise that they are making progress, their confidence grows. They accept new challenges and become motivated to try for more. Furthermore, they derive joy and satisfaction from the process. If that could be made a more widespread phenomenon, it is surely a goal worth striving for.

Conclusion

The practice of providing students with the assessment criteria and their weightings before they respond to an assessment task is now entrenched in higher education. The rationale contains both ethical and practical elements. Most analytic rubrics and similar templates fix the criteria and the rule for combining separate judgments on those criteria to produce the grade. This practice is rarely examined closely, but in this chapter it is shown to be problematic, in principle and in practice.

To address this problem, the unique value of holistic judgments needs to be appreciated, with openness to incorporating criteria that are not on a fixed list. To maintain technical and personal integrity, a number of significant shifts in the assessment environment are necessary. The commitment to mechanistic use of preset criteria needs to be abandoned. Teachers and students need to be inducted into more open ways of making grading decisions and justifying them. Students need to be provided with extensive guided experience in making global judgments of works of the same types they produce themselves. Ultimately, the aim is for learners to become better able to engage in self-monitoring the development of their own works.

Shifting practice in this direction requires a substantially different alignment of pedagogical priorities and processes. It also requires specific strategies to overcome traditions and sources of potential resistance. The aim is to turn the processes of making and explaining holistic judgments into positive enablers for student learning.

Acknowledgement I am grateful to Gordon Joughin for his constant support and critical readings of different versions of this chapter during its development. His many suggestions for improvement have been invaluable.

References

- Bloxham, S. & West, A. (2004). Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment. *Assessment & Evaluation in Higher Education*, **29**, 712 - 733.
- Braddock, R., Lloyd-Jones, R. & Schoer, L. (1963). *Research in written composition*. Urbana, Illinois: National Council of Teachers of English.

- Burke, E. (1759). *A philosophical enquiry into the origin of our ideas of the sublime and beautiful*. 2nd ed. London: Dodsley. (Facsimile edition 1971. New York: Garland.)
- Chi, M. T. H., Glaser, R. & Farr, M. J. (Eds). (1988). *The nature of expertise*. Hillsdale, New Jersey: Lawrence Erlbaum.
- Dewey, J. (1939). *Theory of valuation*. International Encyclopedia of Unified Science, Vol. 2, No. 4. Chicago: University of Chicago Press.
- Ericsson, K. A. & Smith, J. (Eds). (1991). *Toward a general theory of expertise: prospects and limits*. New York: Cambridge University Press.
- Freeman, R. & Lewis, R. (1998). *Planning and implementing assessment*. London: Kogan Page.
- Huba, M. E. & Freed, J. E. (2000). *Learner-centered assessment on college campuses: shifting the focus from teaching to learning*. Needham Heights, Mass.: Allyn & Bacon.
- Lloyd-Jones, R. (1977). Primary trait scoring, in: C.R. Cooper & L. Odell. (Eds). *Evaluating writing: describing, measuring, judging*. Urbana, Illinois: National Council of Teachers of English.
- Meehl, P.E. (1996). *Clinical versus statistical prediction: a theoretical analysis and a review of the evidence* (New Preface). Lanham, Maryland: Rowan & Littlefield/Jason Aronson. (Original work published 1954).
- Morgan, C., Dunn, L., Parry S. & O'Reilly, M. (2004). *The student assessment handbook: new directions in traditional and online assessment*. London: RoutledgeFalmer.
- Orsmond, P., Merry, S. & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, **25**, 23 - 38.
- Polanyi, M. (1962). *Personal knowledge*. London: Routledge and Kegan Paul.
- Rust, C., Price, M. & O'Donovan, R. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education*, **28**, 147 - 164.
- Sadler, D. R. (1980). Conveying the findings of evaluative inquiry. *Educational Evaluation and Policy Analysis*, **2**(2), 53 - 57.
- 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. (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.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, **18**, 119 - 144.
- Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, **5**, 77 - 84.
- Sadler, D. R. (2002). Ah! ... So that's 'Quality'. In P. Schwartz & G. Webb (Eds). *Assessment: case studies, experience and practice from higher education*. (Chap. 16, pp. 130 - 136). London: Kogan Page.
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, **30**, 175 - 194.
- Stevens, D. D. & Levi, A. J. (2004). *Introduction to rubrics: an assessment tool to save grading time, convey effective feedback and promote student learning*. Sterling, Virginia: Stylus Publishing.
- Suskie, L. (2004). *Assessing student learning: a common sense approach*. Boston, Mass.: Anker Publishing.
- Walvoord, B. E. & Anderson, V. J. (1998). *Effective grading: A tool for learning and assessment*. Etobicoke, Ontario: John Wiley.
- Woolf, H. (2004). Assessment criteria: reflections of current practices. *Assessment & Evaluation in Higher Education*, **29**, 451 - 493.

**Assessment, Learning and Judgement
in Higher Education**

Joughin, Gordon (Ed.)
2009, Approx. 460 p., Hardcover
ISBN: 978-1-4020-8904-6

About this book

|

[Table of contents](#)

About this book

There has been a remarkable growth of interest in the assessment of student learning and its relation to learning processes in higher education over the past ten years. Much of the work has sought to explicate the relationship between assessment and learning, and to provide academics with examples of assessment practices that have been considered as successful in promoting learning as well as judging students' achievements. Other work has provided descriptions and frameworks for assessing what have come to be termed 'learning outcomes'. It is now timely to take stock of some of the critical concepts that underpin our understanding of the relationships between assessment and learning. This book makes a significant contribution to conceptualizing key aspects of assessment that are critical to learning, building on research conducted in the UK and Australia. It focuses on the role of assessment in directing students' study, provides feedback that is acted on by students to improve their work, and helps students to develop the capacity to evaluate the quality of their own work and improve it during its production. The book consists of chapters written by leading assessment scholars. Each chapter provides an in-depth examination of an aspect of the conceptual framework.

Written for:

Academics involved in researching assessment in higher education

Keywords:

Higher Education
educational evaluation
educational measurement

[About this book](#)

|

Table of contents

Preface. **1.** Introduction: Refocusing assessment; G. JOUGHIN. **2.** Assessment, learning and judgment in higher education: A critical review; G. JOUGHIN. **3.** How can practice reshape assessment? D. BOUD. **4.** Transforming holistic assessment and grading into a vehicle for complex learning; D. R. SADLER. **5.** Faulty signals? Inadequacies of grading systems and a possible response; M. YORKE. **6.** The edumetric quality of new modes of assessment: Some issues and prospects; F. DOCHY. **7.** Plagiarism as a threat to learning: An educational response; J. CARROLL. **8.** Using assessment results to inform teaching practice and promote lasting learning; L. SUSKIE. **9.** Instrumental or sustainable learning? The impact of learning cultures on formative assessment in vocational education; K. ECCLESTONE. Assessment, Learning and Judgement: Emerging Directions; G. JOUGHIN. Index.