Assessment & Evaluation in Higher Education
Vol. 30, No. 2, April 2005, pp. 175–194

Interpretations of criteria-based assessment and grading in higher education

D. Royce Sadler*
Griffith University, Australia

The increasing use of criteria-based approaches to assessment and grading in higher education is a consequence of its sound theoretical rationale and its educational effectiveness. This article is based on a review of the most common grading policies that purport to be criteria-based. The analysis shows that there is no common understanding of what criteria-based means or what it implies for practice. This has inhibited high-quality discourse, research and development among scholars and practitioners. Additionally, the concepts of 'criteria' and 'standards' are often confused and, despite the use of criteria, the fundamental judgments teachers make about the quality of student work remain subjective and substantially hidden from the students' view. As they stand, none of the approaches identified in the survey is fully capable of delivering on the aspirations of criteria-based grading. Shifting the primary focus to standards and making criteria secondary could, however, lead to substantial progress.

Introduction

Universities usually have some sort of official examination, assessment or grading policy. Along with a range of administrative matters, the typical policy sets out two obligations for university teachers. The first is to tell students clearly, when they enrol in a unit, about the proposed assessment program and the relative weightings of the various components. The second is to provide timely and helpful feedback after each assessment episode. Specifically on the grading issue, the policy may or may not contain a section with detailed guidance on how, and on what bases, judgments about the quality of student performance should be made and appropriate grades assigned. If the policy contains this type of guidance, it is nevertheless rare to find an explicit, coherent and well-argued account of the supporting philosophical or educational

*GIHE, Mt Gravatt Campus, Griffith University, Brisbane, QLD 4111, Australia. Email: r.sadler@griffith.edu.au

ISSN 0260-2938 (print)/ISSN 1469-297X (online)/05/020175–20

DOI: 10.1080/0260293042000264262

principles. In some cases, however, a number of such principles may be deduced from either the policy document itself or from accepted practice. It is immaterial for the current purpose whether the policies are institution-wide or limited to a single department or school. In a relatively small proportion of universities, the stated policy is to devolve all responsibility for grading decisions to individual university teachers, on the grounds that any external prescription or guidance would cut across the teacher's role as an academic professional.

The primary interest in this article is in grading policies that claim to be criteria-based. It is taken as axiomatic that if a policy is so designated, it is reasonable to ask what the 'criteria' are. Four models are described and discussed below. All are claimed by at least some of their institutions or authors to denote criteria-based assessment or grading, for which there is strong in-principle support in the educational literature. The models were identified through a scan of three sources, the dominant one being policy documents from various universities in English-speaking countries accessed by mail request or through Internet searches. The original mail survey resulted in policies being received from 65 universities in Australia, Canada, New Zealand, South Africa, Sweden, United Kingdom, and the USA. Many of these made no reference to the philosophy of their grading policies. This survey was complemented by Internet searches using each of the terms 'criteria-based' and 'criterion-referenced' crossed with each of 'university' and 'higher education'. The other two sources were books on assessment, testing, examining and grading in higher education (extended to sections that dealt with these topics in books on teaching and learning generally) and, finally, journal articles.

Variants and hybrids of the four models abound, so for the sake of brevity, each description is intended to represent a class of similar models. Core properties are emphasised but fine-grained differentiating details are left out. No attempt is made to report on the incidence of the models because the purpose of the study was to map the territory and review how 'criteria-based' is being interpreted in various higher education contexts. However, many institutions employ identical or related models without necessarily calling them criteria-based.

Terminology

Finding an appropriate terminology to use in an analysis of criteria-based assessment and grading involves some essentially arbitrary decisions, because terms are used differently in different countries, and even within a single country, in different education sectors. For example, 'assessment' in some contexts in the USA refers to the evaluation of a wide range of characteristics and processes relating to higher education institutions, including entry levels, attrition rates, student services, physical learning environments and student achievements. In the UK, assessment can mean what students submit by way of project reports, written papers and the like as distinct from what they produce under examination conditions. Similarly, a 'grade' may refer to the classification of the level of a student's performance in an entire degree, the summary of achievement in a single degree component or the quality of a single piece of work

a student submits in response to a specified task. Textbooks and research articles naturally reflect the same level of diversity. To take the discussion forward, decisions about terminology had to be made, the cost being that readers may have to translate the terms used into their own contexts.

Criteria-based. For the purposes of this article, all combinations of 'criteria' or 'criterion' coupled with '-based' or '-referenced' are taken as equivalent. It is possible to draw some distinctions between, say, criteria-based and criterion-referenced grading on historical, conceptual or technical grounds. In practice, however, these terms and the other two combinations are used fairly loosely in the higher education environment, although any one of them may be used with a consistent meaning within a particular institution, department or publication. As with 'criteria', the term 'norm' may also be used with either '-based' or '-referenced'.

Assessment in this article refers to the process of forming a judgment about the quality and extent of student achievement or performance, and therefore by inference a judgment about the learning that has taken place. Such judgments are mostly based on information obtained by requiring students to attempt specified tasks and submit their work to a university teacher or tutor for an appraisal of its quality. Generally, students know which of their works are to be assessed. All types of tasks, and the conditions under which they are to be completed, are included: tests, term papers, laboratory projects, field reports, oral seminar presentations, studio productions, professional placements, assignments and examinations.

Scoring and *marking* are used interchangeably in this article to refer to the processes of representing student achievements by numbers or symbols. Scoring includes counting the number of items correct on a multiple-choice test and assigning a number to reflect the quality of a student's response to an examination item. In most cases, scoring and marking apply to items and tasks rather than to overall achievement in a whole course.

Grading refers to the evaluation of student achievement on a larger scale, either for a single major piece of work or for an entire course, subject, unit or module within a degree program. Scores or marks often serve as the raw material for grade determinations, especially when they are aggregated and the result converted into a different symbolic representation of overall achievement. Grading symbols may be letters (A, B, C, D, etc.) descriptive terms (such as Distinction, Honours, Credit, Pass, etc.), or numerals (such as 7, 6, … , 1). Numerals are usually deemed to represent measurements, and this provides a straightforward route to the calculation of Grade Point Averages (GPAs). The other symbols need a table of numerical equivalents. In this article, the grade scale A, B, C, D, etc., is used throughout.

Once appraisals of student works are made and encoded as scores, marks or grades, the connection between these symbols and both the course objectives and the student productions is essentially broken. Furthermore, marks and grades do not in themselves have absolute meaning in the sense that a single isolated result can stand alone as an achievement measurement or indicator that has a universal interpretation. Assessment and grading do not take place in a vacuum. Professional judgments about the quality of student work together with interpretations of such

judgments are always made against some background framework or information. Two of these frameworks have been mentioned above—criteria-based and norm-based. In recent years, more and more universities have made explicit overtures towards criteria-based grading and reporting. In the process, they typically express reservations about so-called 'grading on the curve', which is one of several forms of norm-referencing. A further framework is self-referenced assessment and grading, in which the reference point for judging the achievement of a given student is that student's previous performance level or levels. What counts then is the amount of improvement each student makes. Self-referenced grading is not normally advocated in higher education but undoubtedly finds its way informally into practice nevertheless. Other frameworks exist besides these three, but they tend to have a lower profile.

General rationale for criteria-based grading

The arguments for criteria-based grading as given in the journal articles, textbooks and grading policy statements consulted for this review could, in essence, be expressed by two ideals that have strong educational and ethical underpinnings:

- Students deserve to be graded on the basis of the quality of their work alone, uncontaminated by reference to how other students in the course perform on the same or equivalent tasks, and without regard to each student's previous level of performance. These two conditions set criteria-based grading apart from all forms of norm-referencing and self-referencing, but they do not specify how it should be done.
- At the point of beginning a course of study, students deserve to know the criteria by which judgments will be made about the quality of their work. This has a primarily prospective purpose, which is to enable learners to use the information to shape their work intelligently and appropriately while it is being developed. However, specifying the bases for grading also serves retrospectively in that stated criteria help to provide a rationale for grading judgments after they have been made and the results given back to the students.

Despite the broad desirability of criteria-based grading, higher education institutions have different conceptions of what it means in theory and in practice. Even the basic interpretation of what constitute the criteria is often unclear. Before analysing grading models that claim to be criteria-based, it is useful to have a general working definition of what a criterion is. The most comprehensive dictionaries list over 20 meanings for 'criterion' (plural 'criteria'). Many of these meanings overlap. Here is a working dictionary-style definition, verbatim from Sadler (1987), which is appropriate to this discussion and broadly consistent with ordinary usage:

> ***criterion*** n. A distinguishing property or characteristic of any thing, by which its quality can be judged or estimated, or by which a decision or classification may be made. (Etymology: from Greek *kriterion*: a means for judging.)

Criteria are attributes or rules that are useful as levers for making judgments. Although judgments can be made either analytically (that is, built up progressively using criteria) or holistically (without using explicit criteria), it is practically impossible to explain a particular judgment, once it has been made, without referring to criteria.

Grading models and their connection with criteria

Grading models may be designed so as to apply to whole courses, or alternatively to student responses on specific assessment tasks. Some can be appropriate for both. Whatever the intended scope, it should be obvious what the 'criteria' are. For the four grading models outlined below, the interpretation of criteria falls within the general definition given above. As will become evident, however, they represent quite different things in practice. A university or department that consistently uses a particular one of these interpretations is likely to assume that this interpretation is more or less constitutive of criteria-based grading. It may not be aware that in another department or university, or even further up the corridor, others have quite different perspectives and understandings.

Grading Model 1: Achievement of course objectives

Under this model, grades are required to represent how well students achieve the course objectives. The objectives model may be expressed in various forms, with different levels of specificity. Three of these are illustrated in Figure 1.

Whichever way this grading policy is expressed, each form makes clear connections between the achievement of course objectives and the grades awarded, without reference to the achievements of other students in the course. In that sense, they all conform, at least in part, to the spirit of criteria-based assessment. The objectives are assumed to provide the basis for the criteria, but exactly what the criteria are is in essence left undefined. Furthermore, students cannot necessarily see a close connection between the course objectives and particular assessment items, and are not in a strong position to be able to judge the extent to which they demonstrate achievement of the course objectives. As it stands, therefore, this type of model has little prospective value for students, and so does not achieve the second desirable goal of criteria-based grading. This is because statements of objectives are framed for a different purpose and have a different structure from statements of criteria.

Consider, for example, this objective for a course in assessment: 'By the end of this course, it is expected that students will be able to demonstrate skill in developing or adapting assessment programs, items and instruments.' This is just one of a complementary set of objectives for this particular course, but is not untypical of how objectives are generally stated. It outlines some of the knowledge and skill students ideally should be able to exhibit by the end of the course, but does not state the process to be used for determining the nature of each student's achievement, or what will constitute 'enough' for it to be taken as 'demonstrated'. There are also no indications of

Form (a)—General statement of principle:

> *Grading is to be based on the principles of* criteria-based assessment*, which is that the desired learning outcomes for a course of study are clearly specified; assessment tasks are designed to indicate progress towards the desired learning outcomes; and the assessment grade is a measure of the extent to which the learning outcomes have been achieved.*

Form (b)—Broad verbal statements about course objectives with generic qualitative descriptions for each grade:

Grade	*Interpretation*
A	Clear attainment of all course objectives, showing complete and comprehensive understanding of the course content, with development of relevant skills and intellectual initiative to an extremely high level.
B	Substantial attainment of most course objectives, showing a high level of understanding of the course content, with development of relevant analytical and interpretive skills to a high level.
C	Sound attainment of some major course objectives, with understanding of most of the basic course content and development of relevant skills to a satisfactory level.
D	Some attainment of a range of course objectives, showing a basic understanding of course content with development of relevant skills.

Form (c)—Achievement of objectives expressed as a tabulation:

Grade	*Major objectives achieved*	*Minor objectives achieved*
A	All	All
B	All	Most
C	Most	Some
D	Some	Some
E	Few or none	Few or none

Figure 1. Three forms of 'objectives-based' grading (Model 1)

whether attainment on this objective will be assessed by itself or in combination with attainments on the other objectives.

Form (a) of this grading model is the most general, and its wording admits the possibility of a continuous range rather than discrete levels of performance outcomes.

Form (b) also implies a continuum. On the other hand, Form (c) implies that the course objectives can be partitioned into major and minor and that the achievement of each objective can be determined on a yes/no basis. In principle, this would enable the respective achievement or non-achievement of each objective to be tallied and a grade assigned accordingly. However, many—perhaps most—educational outcomes cannot be assessed as dichotomous states, although the competency assessment movement predominantly adopts that perspective. More usually in higher education, students perform across a continuum, although that continuum can obviously be divided into two segments, say 'satisfactory' and 'unsatisfactory'. Regardless of whether the scale is considered continuous or discrete, an underlying difficulty is that the quality of performance in a course, judged holistically on the basis of the quality of work submitted, may not be determinable by, or correlate well with, the attainment of course objectives.

The specification of objectives in terms of observable student outcomes is a non-trivial task and requires considerable skill. Where grading against explicit objectives has been attempted, the tendency has often been to become more and more specific, with objectives that are finer and finer grained, partly in a belief that this increases the objectivity of the process or how clearly it can be communicated. When this is taken too far, however, the sets of objectives tend to become atomistic and unmanageable. In addition, the more the objectives are expressed as distinct outcomes, the more likely it is that each objective will become operationally isolated from the others, and the overall configuration that constitutes what the unit is about and what the students are supposed to acquire by way of integrated knowledge and skills recedes into the background.

There are, nevertheless, some exceptions to these generalisations about objectives and criteria-based grading, as for instance, in the special case when the objectives and assessment tasks are formulated jointly so as to be in one-to-one correspondence. Completion of each task (even after multiple attempts, if necessary) is taken as evidence of achievement of the relevant objective. A more sophisticated approach is described in Biggs (1999). He advocates a clear 'alignment' between objectives, teaching and learning activities and assessment tasks, with a high level of coordination and consistency between them. By this means, the criteria to be used in assessment and grading can be linked directly to the way the objectives are expressed. This approach has some conceptual parallels with the behavioural objectives movement. According to Mager (1962), for example, a behavioural objective is not properly formulated unless it includes a statement of intent, descriptions of the terminal behaviour desired, the conditions under which this behaviour is to be demonstrated and the minimum acceptable level of performance (which Mager called the 'criterion') that signifies attainment of that objective.

Grading Model 2: Overall achievement as measured by score totals

This model has the longest tradition in higher education, and remains widely used internationally. It predates by decades the concept and terminology of criteria-based

Table 1. Grading according to aggregate scores (Model 2)

Grade	Mark range	
A	90 – 100	
B	80 – 89	
C	65 – 79	
D	50 – 64	(Passing grade)
E	45 – 49	(Conditional pass)
F	<45	

assessment and grading. Basically, scores on different assessment tasks are added together and then projected—usually linearly—on to a 100-point scale, so it is sometimes called 'percentage grading'. Component scores may be weighted before being added so as to reflect their relative importance in the assessment program. The 100-point scale is then divided into segments according to the number of grades desired. The segments do not have to be of equal size. The schedule in Table 1 illustrates this grading model.

Using numerical ranges gives the impression of definiteness and precision and the system is easy to make operational, but the grade cut-off scores are not usually linked directly to mastery of specific subject matter or skills. It is left to the faculty member to work this out for each unit taught. The obvious measurement issue raised by this policy is how the marks are generated in the first place. Validity, sampling adequacy, item quality, marking standards, marking reliability and measurement error generally are all significant variables that produce an underlying softness in the basic data that typically goes unrecognised.

A number of universities refer to this model as criteria-based, simply because all students who reach the respective threshold scores receive the grades that go with them. This approach does not depend in principle on sorting or ranking student work and does not allocate grades according to relative position within the group. Because no explicit reference to other students' performances occurs—at least at the aggregated stage—it is clearly not norm-based and therefore could be construed as criteria-based. Of course, this claim holds true only if all the scores for all components are also derived by making non-relative judgments. It would be difficult to control for that condition unless the institution has a complementary policy and commitment to having all contributing scores reflect achievement on an absolute scale so that the procedures are not norm-based neither are they, for that matter, self-(that is, student-) referenced.

Where contributing assessment components are each given numerical scores, they are usually but not universally combined by simple addition. Official institutional score-entry spreadsheets often assume that this will be done, supply the algorithm as part of the standard software and use predetermined cut-off scores for automatic grade allocation. The criterion or characteristic by which grades are decided is clearly 'overall achievement as measured by the aggregate score', and this is in broad accord with the general definition of a criterion given above.

Some universities claim that adherence to uniform mark ranges throughout the institution produces a high degree of comparability of standards across schools, departments or disciplines. In some countries, this concept has been extended and embodied as national policy, presumably on the assumption that if all institutions comply with the same grade allocation rule, national consistency in academic standards is achieved.

Grading Model 3: Grades reflecting patterns of achievement

This model differs from the previous one in that, under Grading Model 2, the sub-scores are almost invariably aggregated by simple addition. There are many situations, however, in which this rule for combining sub-scores produces apparent anomalies in the form of grades that seem to be at odds with the university teacher's best judgment about a student's overall performance. The reason is that simple addition is intrinsically fully compensatory. This means that weak performance in some areas can be compensated for by superior performance elsewhere. Once the aggregate is calculated, the actual pattern of strengths and weaknesses is lost entirely.

Grading Model 3 covers the situation where university teachers need a more sophisticated representation of a given student's overall work quality than can be achieved using simple addition for combining the student's sub-scores. Model 3 can take into account such aspects as (1) consistency in performance, (2) a proper balance between work that is submitted earlier in a course (possibly with some formative emphasis) and later (with a summative emphasis), (3) required minimum performance levels on certain critical components (such as professional competence) and (4) developmental tasks that are substantially subsumed by more sophisticated tasks attempted towards the end of the course. Customised rules for combining sets of sub-scores from different sources can be devised to produce course grades that are in substantial accord with the teacher's global judgments, and in ways that cannot be achieved by any simple weight-and-sum method.

To achieve this end, a non-additive non-compensatory rule for combining scores is required. For example, to be awarded an **A** in a course, the combination rule may state that a student needs to perform 'at **A** level on at least two thirds of the assessment tasks and at least at **B** level on the remainder'. Minimum levels on particular components may also be specified. In the literature on human judgmental processes, these types of combinations are known as conjunctive/disjunctive decision rules (Coombs, 1964). Without going into detail, the first example above is said to be conjunctive because there is some scope to trade off performance levels across components in the assessment program, provided the specified pattern is satisfied. When separate minimum levels of performance are required for each component, the rule is said to be disjunctive. Conjunctive/disjunctive decision rules are more likely to be used within a department or academic program than across an institution as a whole. Because students ordinarily expect that their component scores will simply be added together, a faculty member who intends to use a different rule is obliged to spell out exactly what it is at the beginning of the course. To achieve a particular grade

level, a student must satisfy whatever 'criteria' are implied by the composition rule. In this broad sense, the decision rule itself specifies the grading criteria.

Once a grade is assigned according to one of these non-additive rules, the grade awarded will be known to represent a minimum agreed pattern of performance over all components, although individual components still cannot in general be deduced just from the grade. More complex hybrid models can be devised that are partly compensatory and partly conjunctive/disjunctive, but further discussion of these possibilities lies outside the scope of this article.

Grading Model 4: Specified qualitative criteria or attributes

Qualitative criteria can cover a wide range. At the broadest level within an institution, criteria that are intended to characterise graduates from all academic programs may be developed and given official status. These types of generic criteria may be stated in a university's mission statement, the expectation being that all courses studied will contribute in some way to their overall development in students. Performance on generic criteria is not usually graded. At a level much closer to actual teaching and learning, and therefore to assessment and grading, a university may develop a policy that obliges all university teachers to specify the criteria or qualitative properties that the assessor intends to take into account in making a judgment about the quality of student responses to each assessment task. Examples of such criteria (relevant, in this particular example, to written work) are as follows:

Relevance (to task as set);

Validity (say, of an argument, including logical development);

Organization (of the response, including clarity of expression); and

Presentation (the technical aspects of production).

The potential number of criteria relevant to written work is quite large (at least 60), but these four are enough to illustrate. Different sets of criteria are likely to be appropriate, respectively, to other response formats such as seminar presentations, experimental projects, artistic works and professional practice. For each piece of work that students perform or submit as part of the assessment program, the teacher specifies (or in some cases negotiates with students) the criteria that will be used for making the judgment about quality. This type of qualitative teacher judgment is generally inappropriate in so-called objective tests, including multiple-choice, in which case this grading model has virtually no applicability.

Each list of criteria can be elaborated into a marking grid, of which there are several forms. The simplest is a numerical rating scale for each criterion, in which case the ratings could be added to arrive at an overall mark or grade for the piece of work. Alternatively, a simple verbal scale could be used for each criterion (such as *poor*, *acceptable*, *good, excellent*), or expanded into verbal statements that indicate different degrees on each criterion. Constructing this verbal-scale format is sometimes referred to as 'writing the standards'. Not all criteria are of the same type (Sadler, 1983), and

not all require the same sophistication or fineness of judgment, so there is no necessity for the number of standards to be the same for all criteria. In condensing the non-numerical information that refers to a particular student's work, the teacher may simply eyeball the matrix to assign an overall grade. The matrix itself has considerable diagnostic value for the learner.

Characteristics that apply explicitly and directly to student responses clearly lie within the scope of the definition of criteria above. Such response-property criteria, however, typically do not map in any simple way onto course objectives. They are often distributed to students on so-called 'grading criteria sheets' or as 'scoring rubrics' at the time the assessment task is set (Brooker *et al.*, 1998; Montgomery, 2002). The criteria may apply to all tasks of a particular type in a given course, or they may be customised for particular assessment tasks with different sets of criteria applying to different types of responses.

Specifying qualitative criteria for student responses has become a significant movement in higher education during the past two decades and now has an extensive literature. The practice has been championed and refined by scholars in teaching and learning in higher education (Alverno College Faculty, 1994; Brown & Knight, 1994; Knight, 1995). It is foundational to grading using Primary Trait Analysis, originally developed by Lloyd-Jones (1977) for written composition but now having broad applicability across a range of disciplines in higher education (Walvoord & Anderson, 1998; Palomba & Banta, 1999). This burgeoning interest has been driven largely by a concern to inject more transparency into the learning context. Students themselves are inducted directly into the processes of making academic judgments so as to help them make more sense of, and assume greater control over, their own learning and therefore become more self-monitoring. Peer assessment, self-assessment, authentic assessment and a heightened awareness of how high-quality feedback is critical to the development of higher-order learning outcomes have been significant elements in this relatively recent phenomenon. All of these, in one way or another, provide a key instrumental resource for helping students acquire realistic evaluative experience in appraising works of the same kind that they themselves are producing, an approach advocated by Sadler (1989), Rust *et al.* (2003) and many others.

A somewhat less explicit and less systematic format for specifying the grading criteria is to create verbal grade descriptions that apply to a given assessment task, with a separate description for each grade level. An example relevant to an extended written assignment is given in Table 2.

Each grade description has embedded within it a number of criteria, which are of much the same scope and style as those set out in list form. However, in the verbal description format, whereas some or all of the criteria may feature in all of the descriptions, others may occur in as few as one. It is not uncommon, as in the example shown here, for the descriptions to mention slightly different characteristics at the different grade levels. Presumably, this is because the writers of the grade descriptions try to identify particular aspects that could serve as distinctive properties for each grade level. In reality, a student's work is typically never perfectly characterised by any one of the grade descriptions, so the assessor makes a judgment as to which verbal

Table 2. Sample grade descriptions for extended written responses (Model 4)

Grade	Interpretation
A	The work is of very high quality throughout; there is clear evidence of mastery over the subject matter; the material is organised logically; the articulation between various parts is smooth and mutually supportive and the whole work is presented nearly faultlessly.
B	The work addresses the specified issue; it shows a sound level of critical thinking and discrimination; the development provides adequate supporting arguments with reasons and uses factual data correctly; the work is focussed but lacks the originality and creativity that would lift it to **A** level; and the work is substantially free of errors in grammar, punctuation and spelling.
C	The work contains mostly relevant material with some that is marginal; the level of organisation could be improved, with many potential connections between content parts not made; the general approach is reproductive with not a great deal of evidence of creative thought or synthesis; the technical production is reasonably competent, but a lack of clarity in expression sometimes produces ambiguity.

description has the best overall fit with the characteristics of the work submitted. The assessor does not need to make separate decisions on a number of discrete criteria, as is usual when they are given in list form.

By contrast, norm-based grading

This section is included only for reference. When grades are assigned using purely norm-based principles, the proportions of students receiving the various grades are specified in advance as a matter of policy, and then applied to each class of students. Also known as 'grading on the curve', the distribution of grades is usually not rigidly fixed but includes some tolerance to allow for the characteristics of different student groups or other factors. A typical tabulation for this grading model is provided in Table 3.

Because grades are assigned by determining where each student stands in relation to the others, this grading model is norm-based by definition. The shape of the distribution is irrelevant—it could be bell-shaped, rectangular or something else altogether. The proportions are determined essentially arbitrarily (as, incidentally, are the

Table 3. Norm-based grading according to prescribed proportions

Grade	Proportion of students
A	4 – 6%
B	8 – 12%
C	20 – 30%
D	45 – 55%
E	5 – 15%

cut-off scores in Grading Model 2). In the distribution above, the pattern is roughly bell-shaped but deliberately skewed positively. This is because there is more interest in discriminating among students who perform above the minimum passing level, and also because failure rates are typically much lower than passing rates. Under such a grading model, there obviously still exists a 'criterion' for assigning grades to students, namely, achievement ranking relative to other students.

When the proportion of high grades is tightly controlled so as to keep them in relatively short supply, the worth of these grades is also high. In universities where grading on the curve is official policy, it is usually advocated on the grounds that it protects the values of grades throughout the institution and over time. The intention is to 'maintain standards' by reducing any tendency towards grade inflation. Strictly norm-based grading is, however, essentially self-correcting with respect to achievement measured on an absolute scale, because it takes no account of poor course design, poor teaching or poor assessment processes or tasks. Conversely, excellence in course design, teaching and assessment equally go unrecognised.

How can criteria-based grading deliver on its aspirations?

The various interpretations of criteria-based grading paint a somewhat confused and confusing picture for several reasons. First, the level of applicability of the grading models is uneven. As indicated earlier in this article, one model may apply most appropriately to achievement in a whole course, another may be geared to particular assessment items and tasks and another may be useful for both. Second, the meaning of the 'criteria' that underpin each conception differs widely in character and scope from model to model. Which, if any, is correct or at least preferable? Should one interpretation be adopted as definitive so that a broadly consistent system based on it can be developed, refined and promoted as a way of improving assessment and grading generally? On what grounds should a particular one be selected? Choosing one would obviously be possible, at least in theory, although it would attract substantial opposition from dedicated users of other models. Living with multiple interpretations makes it difficult if not impossible to develop a coherent body of discourse, theorising and scholarship about criteria-based assessment and grading. But suppose that over time a high degree of convergence were to evolve. Would that solve the problem?

The considerations set out below show that there would remain a more fundamental and serious issue. It occurs because of the dominating focus on criteria rather than standards. The analysis of grading models given above was cast in terms of criteria (rather than standards) because the main terminology and language forms have historically used 'criteria-based' and 'criteria', even though academic standards remain a key background concern. Instead of stipulating a particular interpretation for criteria as definitive within the grading context, making a clear and consistent distinction between criteria and standards would be preferable. The term 'criteria'—with the meaning set out above—would remain important to enable reference to properties that are instrumental in decision-making, even though criteria by themselves cannot constitute standards.

From criteria-based to standards-referenced grading

In speech and writing generally, and especially for certain classes of statements, the terms 'criterion' and 'standard' can often be used interchangeably, even though they are not fully equivalent semantically. No confusion results when the context makes the meaning clear. For example, something may be said to 'meet a (particular) criterion' when what is meant is that the object possesses enough of the relevant characteristic to meet the minimum required for a given purpose, that is, it meets a particular 'standard'. In other types of statements, neither criterion nor standard can be replaced by the other. For example, something is not normally said to be 'of a high criterion', meaning that it is of high quality, although it may be said to be 'of a high standard'.

Whether what is called a criterion refers, on the one hand, to a property or characteristic or, on the other hand, to a minimum qualifying level can sometimes be determined from the form of words. The noun 'consistency', for example, clearly refers to a property, but when the adjective 'consistent' is applied to a particular object or set of outcomes, it also signifies the existence of a (notional) qualifying threshold that authorizes the application of that descriptor. This implies some sort of standard. Even though the level or nature of that standard may not be made explicit, if the term is reasonably well understood by both sender and receiver of a message, it achieves its purpose. A great deal of descriptive language assumes such (implicit) thresholds: a 'beautiful' sky, a 'very plain' meal, a 'reasonable' price, a 'fair' deal and 'brilliant' acting in a movie. Characteristics that are potentially useful for evaluating all members of a given class are generally criteria, while those relating to appraisals that have already been made usually imply existing standards.

Obviously, there is overlap between the two terms. But the situation is complicated by the fact that the overlap is not symmetrical. 'Criterion' can cover for 'standard' in more contexts and sentence constructions than can 'standard' cover for 'criterion'. This common usage is implicitly recognised by many ordinary (that is, non-comprehensive desk-type) dictionaries. The entries for 'criterion' typically include 'standard' as a meaning or synonym; the entries for 'standard' generally make no reference to 'criterion'. Of the two terms, criterion is broader in scope and therefore more inclusive. This may explain why so many conversations about criteria and criteria-based assessment and grading are able to cover both 'true' criteria and what are really standards without the participants feeling uncomfortable. In the process, however, the dual meaning use of 'criterion' clouds the discussion. Within the context of assessment and grading in higher education, both criteria and standards as distinct but related concepts have a crucial role to play, and it is necessary to be clear on which is intended at each point in a dialogue. Otherwise, meanings can slide almost imperceptibly from one underlying concept to the other even within a single discourse.

With those comments in mind, here is a general definition of 'standard' (Sadler, 1987) that is broadly consistent with the range of the meanings in comprehensive dictionaries and also appropriate to the business of grading academic achievement:

> ***standard*** n. A definite level of excellence or attainment, or a definite degree of any quality viewed as a prescribed object of endeavour or as the recognized measure of what is adequate for some purpose, so established by authority, custom, or consensus. (Etymology: from Roman *estendre*, to extend).

This definition emphasises the twin ideas of a qualifying threshold for each standard, and of agreed-upon standards that are shared across a relevant community. In the outline of the four grading models presented above, the focus has been on criteria rather than on standards, yet standards are crucial to any academic grading scheme. Behind each of the models lies a highly significant operational issue that affects the likelihood of achieving the aims of criteria-based grading. It is this: the implementation of any one of the models necessarily depends on a range of subjective decisions made by university teachers. With the notable exceptions of the formative uses of criteria discussed in relation to Grading Model 4, these decisions are made in contexts where students typically do not gain access to (1) the judgmental processes of the assessor, or (2) the standards that are applied. For example, all of the following reflect subjective decisions:

- The extent to which course objectives are 'met'.
- In what ways an assignment is judged to be 'relevant', 'logically coherent' or have its assertions 'supported'.
- How a student's performance across a course can show 'mastery' of the content and 'high-level' critical thinking.
- Why a score of '17' or a grade of **B** should be assigned to a particular student production.
- Why the choice of a certain set of problems is 'appropriate' for inclusion on an examination paper.
- The bases on which 'part marks' are awarded for partial solutions to problems set as assessment tasks.
- The extent to which a selection of multiple-choice items in a test reflects outcomes that cover and are consistent with the course objectives.
- The reasonableness of insisting that, say, at least two thirds of a set of assessment components should be at **A** level for a course grade of **A** to be awarded.

At the very heart of all grading processes, criteria-based included, lie the professional judgments of university teachers as to the standards that are employed. This is both normal and inescapable, but by no means poses an intractable problem. The situation needs to be understood and managed rather than deplored. The qualities of a student production are rarely exhibited as either unambiguously present on the one hand, or completely absent on the other. They are almost always matters of degree. Furthermore, in many instances they interact with one another in the sense that changing one property may inevitably cause changes in other properties. In determining the degree, the university teacher has the unquestioned advantage over the students because of superior knowledge and extensive experience. Yet a primary purpose for criteria-based assessment and grading is to communicate to students in advance about how judgments of the quality of their performances will be made, and to assure them that

these judgments will be made solely with respect to the quality of their work, without influence by the extraneous factors of how other students perform or their own previous achievement history.

Even when there is careful moderation (also called cross marking) to achieve inter-examiner reliability within a course, the visibility of the standards to students is still often minimal. How, in these conditions, can students use the criteria prospectively? How can they ever know whether their work was judged on an absolute rather than a relative scale? What exactly is that scale? For as long as the standards by which teachers make grading judgments remain implicit, and for as long as teacher–teacher and teacher–learner discussions focus on criteria rather than standards, the critical issue of standards remains submerged.

In practice, the dominant approach to judging the quality of student responses is a hybrid of two factors: the teachers' personal expectations (which cannot be fully communicated using criteria), and how other students have performed (despite protestations to the contrary). The first of these is 'teacher-centred' when there is a single teacher or, when a teaching team engages in cross marking, 'guild-centred'. In neither case does the set of standards have external anchorage. That is, the standards are not conceptualised as having an existence or relevance separately from the context of the teaching team, the course as it was taught and its current students. The second factor arises whenever a teacher peruses a sample of student responses to an assessment task 'to get a feel for the level of performance' or does some trial scoring of them before assigning the final marks. These approaches use the performances of the class members as part of the reference framework for grading.

To realise on the aspirations for criteria-based grading, a major shift in orientation is required towards 'standards-referenced' grading. Criteria-based grading begins with a focus on the criteria, leaving the standards to be implied or experienced incidentally. Criteria form an essential element of the evaluation and communication process, but ultimately it is the students' appreciation of quality, set against a background of external standards, that is of significance. A more appropriate approach would be to focus on the standards as the primary reference points against which student submissions are judged. Whatever criteria are relevant to the decision-making (whether these criteria are pre-specified or emergent) should be invoked—to whatever extent is necessary—to support or explain the standards and the decisions. Whether work can be appropriately referred to as 'good', 'mediocre' or 'poor' depends on a shared interpretation of the underlying standards and, at least implicitly, the corresponding grade boundaries.

For standards to function properly for both formative and summative purposes, they need to be established, and made accessible (1) to students, before and during the course, (2) to the university teacher or assessor, so that the students' work can be appraised within that framework, and (3) to the panels that review grade distributions. Only then can proper discussions take place about what standards mean. Only then can the appropriateness of the standards employed be subjected to scrutiny. Only then can proper judgments be made about the quality of student work submitted for marking. And only then can the legitimacy of those judgments be vouched for.

The four fundamental challenges facing university educators who wish to grade students according to true standards are these:

- Coming to grips with the *concept of a standard*;
- Working out how to *set* standards;
- Devising ways to *communicate* standards to students and colleagues; and
- Becoming proficient in the *use* of standards.

In this article, the scope of this developmental agenda can be sketched only in the broadest terms. An appropriate starting point is a consideration of what the typical university teacher brings to the task of grading student work. Assuming that the teacher has an ideological preparedness to engage in making judgments about actual quality (rather than, say, wanting to reward effort or improvement), the teacher usually has a reasonable idea of, or a feel for, the 'standards' they intend to apply, even though they may never have attempted to formulate them explicitly. The typical teacher also has a desire to be consistent and fair in applying their 'standards'. Even though there may be some conscious or subconscious adjustment to expectations in the light of the work that actually comes in, the existence and use of these underlying 'standards' means that the grading activity is not, in a wide variety of fields and disciplines, about making judgments that are wholly relative. Personal 'standards' have their origins in accumulated discipline knowledge and skill, the teachers' recollections of having had their own work graded during their studies, and their previous personal experience in grading student work. (The use of quotation marks around 'standards' here signifies that a teacher's personal 'standards' are not true standards in the sense of the definition given earlier in this article because they are not necessarily accepted and shared within a relevant community, which in this context is a competent group of academic peers.)

An additional factor that supports the use of 'standards' is essentially practical. It follows from the fact that, although arranging a set of student works in order of (relative) quality may appear to be a fairly straightforward process, it is extremely laborious when the number of works to be graded is large. Making pair-wise comparisons among a small set of student submissions so that they can be sorted according to quality is generally not a difficult cognitive task. Strict norm-referencing becomes increasingly unmanageable, however, as attempts are made to scale up the process to larger and larger numbers of submissions. Resorting to the lecturer's own framework of 'standards', which is often assumed to have some stability about it, makes large-scale grading tasks feasible. To some extent, these two factors—personal 'standards' and practical necessity—intuitively predispose university teachers to incorporate elements of 'absolute' grading into their appraisal practice. This phenomenon can be capitalised upon to establish the groundwork for a standards-referenced system. The next step in the process is to identify ways to conceptualise and formulate standards so that they can be made accessible to both teachers and learners. Finally, it is necessary for appropriate levels for the standards to be negotiated and fixed.

There are four basic approaches to setting and promulgating standards generally: numerical cut-offs on a proper measurement scale, tacit knowledge, verbal

descriptions, and the use of exemplars. The outline below draws from the analysis in Sadler (1987), but is necessarily indicative rather than comprehensive.

Numerical cut-offs in the context of grading student work are not normally controlled by any set of procedures that would pass technical scrutiny, primarily because of the softness or unreliability of the underlying 'measurement' scale that arises from the subjectivity of assessment processes, as indicated earlier in this article in connection with Grading Model 2. Numerical cut-offs are therefore intrinsically problematic.

Tacit knowledge refers to the expertise people carry around with them, mostly in their heads but also, in some academic and professional fields, embedded in psychomotor skills. Standards based on tacit knowledge commonly exist in unarticulated form but can be shared among experts, or transmitted from expert to novice, by joint participation in evaluative activity, including moderation of grades. Reliance on standards that are based purely on tacit knowledge has a certain mystique about it that reduces transparency and often supports a dependency relationship between learner and teacher. Students are then forced to rely heavily—or even exclusively—on the judgments of others. Unless students are enabled, through the design of the learning environment, to develop appropriate evaluative expertise themselves, they cannot self-monitor and thereby control the quality of their own work while production is in progress. Tacit knowledge nevertheless has a key role to play.

Verbal descriptions consist of statements setting down the properties that characterise something of the designated levels of quality, much after the style of the criteria statements in Grading Model 4 above. They are intended to be precise enough to allow unambiguous determinations of quality without reference to particular examples of student work. Obviously, the ideal would be for the resulting verbal descriptions to be refined to the point where they constitute formal definitions. Were this to be achieved, it would make them highly portable and give them a currency that transcends particular cases. It turns out, however, that the nature of the elements in verbal statements makes this impossible. For example, Grading Model 4 contains the phrase 'sound level of critical thinking and discrimination' for the grade of **B**. What constitutes a 'sound level'? Is there some purpose (beyond the award of a grade of **B**) against which soundness is to be ascertained? How well would experts agree on what is 'sound'? Constructing answers to each of these questions sets up new verbal terms that in turn call for more elaboration, and so on in infinite regress.

Exemplars are key examples of products or processes chosen so as to be typical of designated levels of quality or competence (Sadler, 2002). Because exemplars are more concrete than they are abstract, they are especially convenient for direct 'viewing' by academic colleagues and students. When multiple criteria are used in appraisals of quality, which is a common situation in higher education, a single case cannot constitute a standard, although it may exemplify it. Other examples relating to the same standard ordinarily may be expected to differ from one another.

Theoretically at least, the most promising approach to setting and promulgating standards would probably be to start with a set of qualitative grading decisions made by teachers, and tease out the substantive reasons for them. This is basically an

inductive process which requires a phase in which relevant characteristics (including criteria and, where appropriate, degrees or levels on those criteria) are abstracted from real judgments, and then clarified and codified. The tacit knowledge that accounts for these judgments would then need to be explored and articulated in relation to the works graded, some of these works eventually becoming designated as exemplars. The goal should be to find an economical set of verbal descriptions and exemplars that are not intended to stand alone but together can be embedded within the context of the tacit knowledge of a relevant group of academics so that explication of key terms, concepts and understandings can be constructed and communicated on demand. The overall purpose should be to identify, and then convey as fully as possible, the essence of the 'knowledge' that makes up the standards.

Conclusion

Internationally over the past two decades, higher education institutions and educators have become increasingly committed to making assessment and grading more effective in promoting student learning (that is, in fulfilling a significant formative function) and to making less mysterious, more open and more explicit the grounds upon which student productions are graded. This has resulted in a strong interest in grading criteria and so-called criteria-based assessment. All of the grading models and versions analysed in this article studiously try to avoid any form of norm-referencing. In addition, all make a legitimate claim to being criteria-based because they employ criteria either explicitly or implicitly. Apart from satisfying these minimalist conditions, however, they have little in common operationally. Additionally, academic standards are typically not seen as common professional property within cognate disciplines in the higher education environment, and ultimately as common ground between teachers and learners.

Taken together, the current situation is characterised by (1) shared aspirations for criteria-based assessment and grading, (2) a multiplicity of interpretations of what they mean and imply for practice and (3) a lack of coherent discourse and research. A particular source of confusion is that the terms 'criteria' and 'standards' are often used interchangeably, as if they were equivalent. A key conclusion reached through this study is that these two terms are distinguishable, and that considerable benefits would accrue from using them consistently in their distinctive ways because both criteria (as attributes or properties) and standards (as fixed reference levels of attainment) lie at the heart of high quality assessment and grading. The final stage is to explore ways of conceptualising standards, to negotiate and arrive at a consensus on appropriate levels for the standards, to provide opportunities for students to make judgments using the standards and finally, to apply the standards consistently in summative grading.

Notes on contributor

D. Royce Sadler is Professor of Higher Education at Griffith University, Brisbane, Australia. His main research interests are in assessment of student learning,

achievement standards, grading policies and how assessment can be used to improve learning.

References

Alverno College Faculty (1994) *Student assessment-as-learning at Alverno College* (Milwaukee, WI, Alverno College).

Biggs, J. (1999) *Teaching for quality learning at university: what the student does* (Buckingham, UK, SRHE & Open University Press).

Brooker, R., Muller, R., Mylonas, A. & Hansford, B. (1998) Improving the assessment of practice teaching: a criteria and standards framework, *Assessment & Evaluation in Higher Education,* 23, 5–24.

Brown, S. & Knight, P. (1994) *Assessing learners in higher education* (London, Kogan Page).

Coombs, C. H. (1964) *A theory of data* (New York, Wiley).

Knight, P. (1995) *Assessment for learning in higher education* (London, Kogan Page & SEDA).

Lloyd-Jones, R. (1977) Primary trait scoring, in: C. R. Cooper & L. Odell. (Eds) *Evaluating writing: describing, measuring, judging* (Urbana, IL, National Council of Teachers of English).

Mager, R. F. (1962) *Preparing instructional objectives* (Belmont, CA, Fearon).

Montgomery, K. (2002) Authentic tasks and rubrics: going beyond traditional assessments in college teaching, *College Teaching,* 50, 34–39.

Palomba, C. A. & Banta, T. W. (1999) *Assessment essentials: planning, implementing and improving assessment in higher education* (San Francisco, CA, Jossey-Bass).

Rust, C., Price, M. & O'Donovan, B. (2003) Improving students' learning by developing their understanding of assessment criteria and processes, *Assessment & Evaluation in Higher Education,* 28, 147–164.

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

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. (2002) Ah! … so that's 'quality', in: P. Schwartz & G. Webb (Eds) *Assessment: case studies, experience and practice from higher education* (London, Kogan Page), 130–136.

Walvoord, B. E. & Anderson, V. J. (1998) *Effective grading: a tool for learning and assessment* (San Francisco, CA, Jossey-Bass).