Author's manuscript of an article subsequently published as:

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, **34**, 159-179.

Indeterminacy in the use of preset criteria for assessment and grading

D. Royce Sadler, Griffith University

Abstract

When assessment tasks are set for students in universities and colleges, a common practice is to advise them of the criteria that will be used for grading their responses. Various schemes for using multiple criteria have been widely advocated in the literature. Each scheme is designed to offer clear benefits for students. Breaking down holistic judgments into more manageable parts is seen as a way to increase openness for students and achieve more objectivity in grading. However, such approaches do not adequately represent the full complexity of multi-criterion qualitative judgments, and can lead to distorted grading decisions. Six anomalies in the ways assessors approach the grading task are identified, together with several likely contributing factors. Overall, the conclusion is that explicit grading models do not have as strong a theoretical foundation as is commonly supposed, and that holistic appraisal merits further investigation.

———

Introduction

Judgments that teachers make about the quality of student works are, in a profound sense, definitive. There is no objective, methodologically independent way of validating them. Regardless of how they are actually made, they also represent high stakes for learners. Grades matter because they form the foundation for the certification of learning. Transcripts of academic achievement have a permanency about them, and the levels of student performance often have significant career implications. Grades also matter because they have a substantial affective impact on learners and their learning, influencing both students' sense of achievement, and their motivation and level of engagement in future courses.

Over the past two decades, grading schemes that employ fixed sets of criteria have become firmly established in higher education. They are widely advocated in books such as those by Freeman & Lewis (1998); Huba & Freed (2000); Morgan et al. (2004); Stevens & Levi (2004); Suskie (2004); and Walvoord & Anderson (1998). The various systems go by such names as rubrics, grading schemes, scoring keys or guides, criteria sheets and primary trait analysis. This article is about the general validity of using these approaches, which have become mandatory in some institutions. The argument is that they can, and in many cases do,

fail to meet the conditions for sound assessments of complex student works, and that this deficiency is inherent in the method. Although Broad (2000), Shay (2005) and others have expressed reservations before, the critique mounted in this article analyses the issue from a different angle.

Scope of this article and use of terms

This article is relevant to a specific range of approaches for marking or grading student responses to a particular class of assessment tasks. These tasks all require divergent or 'open' responses from students. This means that there is no single correct or best answer, result or solution. Response formats include term papers, essays, written assignments, field and project reports, seminar presentations, studio and design productions, specialized artefacts, professional performances, clinical consultations, creative works, and client interviews. Divergent tasks dominate in a range of disciplines and professional academic programs, and account for a significant proportion of assessment activity in higher education. They are intended to provide opportunities for students to demonstrate sophisticated cognitive abilities, integration of knowledge, complex problem solving, critical opinion, lateral thinking, and innovative action.

Producing a response requires abilities in both design and production, the objective being to allow considerable latitude for creative solution, analysis or expression. There are no formal techniques or recipes which, if followed precisely, would lead to high-quality responses. Divergent responses are referred to in this article as *works*, regardless of their material form. Constructed responses to examination items may, if they are substantial enough, also qualify as 'works', but responses to objective test items do not.

Divergent works are typically complex, in the sense that their quality can be explained only by reference to multiple criteria, possibly including some that are abstract in nature (Sadler, 1983). Throughout this article, a *criterion* refers to a property, quality, characteristic or attribute of a student response. It is distinct from a *standard*, which refers to a particular degree or level of quality (Sadler, 1987). An important distinction is to be made between *quality* as an integrative concept which characterizes a work as a whole, and *a quality*, which is synonymous with a property. Quality characterizes works holistically, often seemingly with a certain detachment from particular qualities. Criteria that are advised to students with, or as part of, an assessment task are referred to as *preset*.

Determining the quality of complex works requires skilled, qualitative judgments. A qualitative judgment is one made directly by the appraiser, the person's brain being both the source and the instrument for the appraisal. The judgment is not reducible to a set of measures or formal procedures that a non-expert could apply to arrive at the 'correct' appraisal (Sadler, 1989). Within each of the response genres listed above, student works that differ considerably from one another in character and structure may nevertheless be judged to be comparable in quality. This type of variability in completed works is regarded as highly desirable in many domains of higher education. For these, variability in quality results primarily from student-related factors, such as how they interpret the task (which is assumed to be clearly enunciated), and the choices they make in designing their works and putting them together.

In this article, a *mark* generally refers to a numerical score on a fine-grained scale, such as 0–30, or 0–100. *Grades* are expressed on a coarser scale. The anchor points may be denoted by words (Distinction, Merit, Credit…), numerals (7, 6, 5…) or letters (A, B, C…).

A mark or grade is assigned to a student work primarily to represent its level of quality, and possesses meaning within a relevant framework. In situations where the distinction between marks and grades does not matter, the terms are often used interchangeably.

Analytic and holistic grading

For a wide range of student responses, grades are commonly decided in one of two ways. Both are based on qualitative judgments, the two approaches differing primarily in their granularity. In so-called *analytic* grading, the teacher makes separate qualitative judgments on each of the preset criteria. If no criteria are provided to students in advance, the teacher determines them prior to marking. After the criterion-by-criterion judgments are made, they are combined, usually by way of a rule or formula. The resulting aggregate is either used as it is, or converted into a grade. The final mark or grade is thus built up from a series of smaller-scale decisions. When the steps are followed systematically, the grade follows as a logical outcome.

In *holistic* (also called *global*) grading, the assessor progressively builds up a complex mental response to a student work. This involves both attending to particular aspects that draw attention to themselves, and allowing an appreciation of the quality of the work as a whole to emerge. The appraiser then makes a qualitative judgment as to its overall quality, and maps that judgment directly to the appropriate point on the grading scale. In addition to assigning the grade, the assessor may provide a rationale for it, perhaps in summary form for the work as a whole, or as running comments on various features of the work. Rationale and feedback statements necessarily invoke one or more criteria, because criteria are constitutive elements of all evaluative explanations or advice. In analytic grading, criteria play a clear front-end framing role. In holistic grading, the assessor's emergent global judgment dominates. In principle at least, the global judgment is made first; references to criteria follow from reflection on that appraisal.

Summary grades are often useful even for assessment episodes that are intended as purely formative or diagnostic, because they give students an indication of the overall quality of their works. However, research on the impact of providing a grade only, specific feedback only, or both, shows that there may be differential effects according to contextual factors (Black & Wiliam, 2004). For example, there may be differences that reflect the extent to which students use their grades as a means of establishing their performance rankings among peers (Dweck, 1986; Butler, 1987, 1988), or differences that are specific to discipline (Hodgen & Marshall, 2005).

Historically, holistic grading held sway in many areas of higher education until the mid-1990s. Since then, a steady swing towards analytic grading has occurred. The overt intention of this movement has been to put more structure into the grading process, reducing the apparent arbitrariness with which faculty decide grades, and making the procedures explicit for students. This sort of codification aims to ensure that markers not only award appropriate grades but also, in the course of making them, automatically reveal the grounds for them to students.

Debates about the relative merits of analytic and holistic grading have had a long history, with most attention being focussed on their respective reliabilities. Whole volumes have been written about reliability in assessment. Although much of it had its origins in psychological testing, its importance in both formative and summative educational assessment is widely recognized (Brown & Knight, 1994; Black & Wiliam, 2006). At its

most basic, reliability refers to the extent to which an assessment score or grade could be reproduced under different conditions. Three of the many factors that affect reliability are: the content and form of assessment items actually used (particularly, how well the set of tasks samples the course domain and its intended outcomes); the decisions students make (particularly, how they choose when options are available among assessment tasks); and the level of accord among markers (particularly, how well they agree on methods and standards).

Of these three, scorer reliability is the one that has dominated comparative research into the application of holistic and analytic grading. Scorer reliability can be quantified as a statistic that measures the degree of consistency between grades assigned by different teachers to a set of student responses (inter-grader reliability), or by the same teacher grading the same works on successive occasions (temporal reliability). Although achieving scorer reliability is a desirable goal for all grading, a comprehensive review of the relative merits of different grading schemes would approach the task from a variety of other perspectives as well, including validity, objectivity, efficiency, formative effectiveness and the student experience.

The theme in this article centres on a particular aspect of validity. As with reliability, validity has a considerable literature, and again, its roots lie in the field of psychometrics, where various types have been identified according to the purpose the measurements are meant to serve (Messick, 1989). The psychometric interpretations of validity, and the relation between intrinsic validity and 'validity for a purpose', have only marginal bearing on the present topic. Reliability and validity are interconnected; to the extent that assessments are unreliable, the validity of any inference drawn from an appraisal is weakened. Notable in the field of education is the analysis of various threats to validity identified and discussed by Crooks et al. (1996), and by Stobart (2006) in his chapter on validity as it relates to formative assessment.

The concern here is with the relative validity of analytic and holistic appraisals. These approaches clearly differ in the processes they apply, and their equivalence cannot therefore be taken as axiomatic in the technical sense. When the respective procedures are followed meticulously, an appraisal by either approach could be regarded as soundly based methodologically and therefore able to withstand internal criticism or objection.

The working presupposition here is that there exists an underlying 'true' appraisal for a particular work to which both holistic and analytic approaches may provide an approximation. The question is, Which of the two yields the truer portrayal of the level of quality in a student work? Or do they simply use different methods to produce equally valid results? If the latter, either one could be considered a legitimate substitute for the other. The issue has been posed in this way in full realization that it begs the question of what could conceivably be considered the 'true' appraisal anyway. That point will be touched on briefly at the end of the article.

Restricting the range to this aspect of validity has the advantage of making the analysis manageable within a single article. The disadvantage is that complex relationships between validity and other aspects of grading remain unexplored. The proposition to be argued is that, in the hypothetical case of all student works being assessed without human error, both holistically and analytically, the two approaches would produce different end grades for many works. This comes about basically because analytic grading often fails to capture special characteristics, which properly managed holistic appraisals can. The corollary is that holistic grading deserves to have its status reviewed.

The types of grading schemes relevant to this analysis all have the characteristics mentioned above, namely, use of a fixed set of criteria, with a separate judgment made on

each criterion. Students are supplied with the set of criteria before they commence responding to the assessment task. The teacher may prescribe the criteria; students and teachers may negotiate them together; or the teacher may require students to develop their own as a way of deepening their involvement in the assessment process. The issue of who decides the criteria is not important to this analysis. If the component judgments on the criteria are to be combined using a mathematical formula to arrive at the grade, all the matters raised in the article are relevant. In some analytic schemes, the component judgments are not assembled mechanically into an overall grade but by means of a separate high-level qualitative judgment. Some of the matters related to mechanical combination do not then apply. Of the criteria-based schematic forms that are relevant to this article, two of the most common are now outlined.

The first is the 'rate-weight-sum'. These systems start with a list of the specified criteria, and assign a weight to each one according to its importance in the appraisal decision. The same weights are used for all students. Each criterion has its own scale line, with fixed numerical anchor points. The assessor marks the position that best represents the 'strength' of a particular student's work on each criterion. The associated numerical values are then multiplied by their respective weights. Adding the weighted scores produces an aggregate, which is either reported directly or converted into an overall grade. A modified form is sometimes used for strictly formative purposes. It employs scale lines without numerical indicators or weights. For each student work, assessors then record only the positions on the respective scale lines for the criteria.

The second form is the *rubric*, which is a cross-tabulation of criteria against so-called 'standards'. (The nomenclature varies. Some rubrics use the terms 'qualities/criteria' instead of 'criteria/standards' as headings.) Suppose the criteria are arranged so that there is one row for each criterion. Each cell in a given row typically contains descriptive text that spells out the characteristics of a particular level or 'standard' for that criterion. This text can take several forms, including 'verbal quantifiers' and 'typical features'. The verbal-quantifier form can use a variety of words, but all amount to much the same thing. They range in meaning from the equivalent of 'little if any evident' to 'an extremely high level present'. The typical-features form of cell entry consists of a qualitative description of the corresponding 'standard', often with reference to sub-attributes of the main criterion. Regardless of the form of the text in the cells, the assessor nominates the cell that best characterizes the quality of each student work on each criterion. Each cell in the tabulation may be given a numerical value, which is then processed mathematically along with all the other applicable cell scores to arrive at a total score for that work and, if required, a summary grade. Less commonly, the assessor may grade simply by reviewing the pattern of cells identified in the matrix, and making an overall judgment.

Exclusions

Not all rubrics, and not all divergent works, qualify for analysis in this article. The rubrics described above are sometimes called *analytic rubrics*. So-called *holistic rubrics* are different, in that they use extended verbal descriptions to set out the characteristics that are typical of, or expected for, each mark range or grade level. For example, an assessor may decide that, of the set of verbal descriptions available for the different grades, the following one (for a written work) fits better than any other: 'The work shows clear mastery. Arguments are presented clearly and coherently, and all key statements are backed by solid

evidence. Readability and technical competence are of a high order.' Each such description is reasonably full, and is intended as indicative rather than definitive or prescriptive. The descriptions invariably contain embedded or implied references to multiple criteria, but do not necessarily refer to the same criteria at all grade levels. Assessors do not apply the descriptions literally to student works, and there are no judgments on individual criteria. The limitations of holistic rubrics differ from those of analytic rubrics, particularly with respect to their usefulness for feedback. They are not dealt with further here.

For certain divergent works, which tend to be more common in technological fields, it is virtually impossible to conceive of quality as a global, high-order attribute in its own right. The idea of quality certainly makes sense, but only as it is constituted by putting together performance data from several sources. An example is an assessment task that requires students to develop a solution strategy for a class of problems for which clear performance criteria are specified. The product design is left open, the deliberate intention being to allow considerable scope for creativity in the development of solution strategies. Student works may be assessed on how they measure up on the following three criteria: effectiveness or technical correctness (that is, the degree to which the strategies actually work or produce authentic solutions), efficiency (the level of overheads, including time, with which the solution processes operate), and robustness (the ability of the strategies to cope successfully with difficult test cases and known producers of anomalies). Performance on these criteria may correspond to physical characteristics that can be assessed by such direct means as counting, measuring, or applying standardized test procedures, and then applying mathematical operations to the data. The appraisal criteria are set beforehand, by necessity. Assuming no dimensions of value exist outside these criteria, they not only specify but also effectively exhaust the meaning of 'quality' in the context. In general, the criteria are not arbitrary, but reflect those that are accepted in the relevant discipline, profession or industry. This article does not deal further with that class of works.

Six observations

The typical account of how explicit models can be used to make complex appraisals portrays the process as quite straightforward. In practice, it does not always proceed so smoothly. In this section, six types of significant non-standard events are identified and described under the heading *Observations*. They manifest themselves across a wide range of university teachers, in a wide range of disciplines and fields, and for many types of assessment tasks and student works. Each observation has been investigated in conversations with academics, then elaborated and refined. When a certain type of anomaly has been raised as 'something that may happen on occasion', some colleagues have been able to identify with it immediately. Others appeared initially not to be focally aware of the particular event type, but recognized it either when it was described in detail, or subsequently as it occurred in their own practice. The phenomena, by their very nature, remain mostly invisible to learners. Some of the observations presented here manifest themselves in analogous ways in other evaluative decision settings, such as when selecting from among applicants for an employment vacancy using a fixed set of selection criteria. Some of the Observations are linked. For example, the 'dual agendas' referred to in Observation 1 are taken for granted in several later Observations.

Observation 1

At least implicitly, the portrayal to students is that analytic grading first involves a series of judgments made at the level of individual criteria. After these are completed, they are processed into a grade using a well-defined algorithm in an essentially linear sequence of steps. As a procedure, this is easily understood and appears rational, systematic and essentially straightforward. But as many teachers testify, the actual process is much more convoluted.

Markers generally do not scrutinize an entire work multiple times, once for each criterion. As mentioned earlier, they run simultaneously with dual agendas, one being to develop a perspective on the work's overall quality, and the other to take note of particular characteristics or deficiencies that are worthy of special focus or attention. These particular 'noticings' are useful both in shaping the emerging perspective, and in generating raw material for reporting. As most academics can attest, working systematically through the criteria one at a time would be highly labour intensive. Apart from the labour aspect, a significant reason for not operating in this way is that assessors are initially more interested in how the work comes together as a whole than in performance on individual criteria.

In reporting the outcome of the appraisal process, the substantive content of the global judgment is analysed through the lenses of the preset criteria. Anything not visible through those lenses is filtered out and discarded. The levels on the individual criteria are determined retrospectively and, in many cases, creatively. Apart from general reflection, the processes may involve rescanning the entire work, or focussing on physical parts of it or specific characteristics. One form of this creativity was described in 1963 by Braddock et al. in relation to analytic marking of school-level written composition. Beginning assessors 'first establish the total number of points according to [their] general impression of a [work's] merit, and *then* apportion the total points among the various categories so that they add up to the total. Such a practice … undermines the basis of the analytic method…' (p. 14; italics in the original).

Observation 2

During the process of grading certain works, experienced assessors often become aware of discrepancies between their global and their analytic appraisals. A work judged as 'brilliant' overall may not rate as outstanding on each criterion. This would be necessary, logically and arithmetically, for the work to be assigned the top grade. Conversely, another work that comes out well on each criterion may be judged as only mediocre overall. Although the directions of discrepancies between holistic and analytic judgments may go either way, only the existence of this phenomenon is significant here, not the relative incidence of the directions.

Language use also recognizes this type of anomaly as a general phenomenon in a range of contexts outside teaching and assessment. This claim can be easily tested through an Internet search for the occurrence of the two key phrases that follow. If the global appraisal comes out higher than the analytic, the global property, which in the teaching context is quality, is commonly said to be 'more than the sum of its parts'. In the reverse direction, the global appraisal is judged to be 'less than the sum of its parts'. (For some reason, the first of these events appears to be more commonly noted than the second.) Because both holistic and analytic approaches aspire to represent the (idealised) 'true' grade for the work, the two

should correspond, within a small margin of error. It is clearly imperative that they do correspond if analytic and holistic gradings are to be considered equivalent.

The matter of interest under the heading of this Observation is assessors' approaches to dealing with these perceived discrepancies. Take first the simplest case. Assessors who accept the formula-based grade at face value, and therefore as authoritative, typically just ignore any holistic 'appraisal' without further ado. If they put more faith in the holistic appraisal, they often apportion the total score across the criteria, as mentioned in the previous Observation.

The situation becomes more complicated with assessors who aim to form a valid holistic appraisal, and an equally valid analytic assessment, and have them tell the same 'story'. The typical reaction is to question why a judgment that is painstakingly 'built' from careful component judgments fails to produce the 'true' assessment, the holistic appraisal. (Faculty generally seem to have more intuitive confidence in their holistic appraisals than in the analytically derived grades.) A common way to reconcile the two is to respond pragmatically: adjust the reported levels on the criteria until the analytically derived grade, as reported to students, agrees with the grade the marker believes is justified. As a cognitive operation, this is somewhat distinct from the apportionment approach referred to above. To adjust the holistic appraisal so that it agrees with the analytic judgment is not normally possible, because the holistic judgment has already been formed using extended engagement with multiple cues. It is therefore less malleable at that point.

Assessors who adjust the analytic scores and attribute the need for it to their lack of skill in implementing analytic grading may nevertheless remain uncomfortable about the process. Basically, there are two reasons. First, they are not totally sure which one should take precedence, the holistic or the analytic. Second, an action that could be interpreted as arbitrary – or 'fudging' if it were ever divulged – raises an ethical issue, at least on the surface of things. For these reasons, teachers are generally reluctant to talk about their experiences with this type of anomaly, or their actions to resolve them, to either their students or academic colleagues. However, many are prepared to discuss the phenomenon quite openly in secure research environments. This phenomenon has also been observed at other levels of education. In a study of English teachers at school level, Marshall (2000) found that teachers who found difficulty with the use of explicit criteria instead trusted their global judgment, despite viewing this as 'transgressive, an act of defiance against a growing orthodoxy [of analytic appraisal], which they [were] desirous to resist' (p. 164).

Observation 3

When assessors identify works for which the analytic and global judgments differ, they may or may not be able to account for the discrepancy. Many teachers can identify with situations in which a work exhibits an indefinable 'quality', inherent in its wholeness, which simply cannot be passed over as irrelevant or inconsequential. In some cases, this quality is sufficient to set the work apart, almost in a class of its own. However, in other cases the assessor knows the reason: it lies in a positively identified criterion which was not included in the preset list distributed to students but turns out to be crucial to the judgment. Clearly, 'distributing' the influence of any non-standard criteria across the specified criteria is not acceptable. In order to keep the discussion simple, assume there is only one of these unspecified criteria, although in practice there could be several. Also assume that the assessor has checked that this criterion is not included on the preset list in some implicit form, such as an extreme level on a

stated criterion that goes under a different name, or a blend of particular levels of several nominated criteria. When the appraisal protocols are followed strictly, such an 'emergent' criterion cannot be admitted, formally or informally, into either the process of grading, or any subsequent explanation. To do so would breach the implicit contract with students, which is that only the preset criteria will be used in deciding the grade. This contract is formalized simply through promulgating a fixed set of criteria. Almost invariably, no escape clause accompanies these, and this implies that the set is binding.

The teacher then faces the dilemma of whether to disclose this criterion to the student, and with it the weakness in the system. It is ironic that a substantial part of the rationale for using preset criteria and a combination rule is to make visible the basis for the teacher's appraisal, something that can be kept more or less private with holistic judgments. Yet the rationale for sticking rigidly with the specified system, without acknowledging its inadequacies when they occur, can be achieved only at the expense of non-disclosure of significant information once the grading event is over.

Observation 4

A necessary condition for any formal system that uses discrete criteria as logical entities in their own right is that the criteria are conceptually distinct from one another. Each criterion is assumed to have an established interpretation that, at least in theory, represents a property that is different from those signified by the other criteria, taken singly or together.

This is a strong assumption that represents an idealized situation which is not necessarily realized in practice. As grading proceeds, specifically at the stage when teachers nominate the strength of the work on individual criteria, it is not uncommon for teachers to experience considerable difficulty in differentiating some criteria from others. After several of the criteria have been scored, the properties of the work that were initially assumed to be mutually exclusive often seem to blur into one another. At some point, one criterion may seem to have had all of its meaning accounted for by application of the criteria already attended to. This is not to say that some criteria are essentially the same as others. What seems to happen is that some of the criteria that are used early in working through the scoring process explain virtually all of the significant characteristics of the work, leaving little to be associated with criteria further down the list.

It would be convenient if criteria were also functionally uncorrelated, with performance on one criterion not being associated, in a structural sense, with performance on others. Were functional independence to apply literally, a student could improve a work according to any one criterion without necessarily affecting the levels on other criteria. This can occur in some contexts, but is rare in many others. Students in the process of improving their work must then use multiple criteria simultaneously in order to monitor the development of overall quality.

Neither conceptual distinctiveness nor functional independence is an indispensable condition for an analytic model to produce a grade. However, both have significant implications if a completed grading matrix is to be used as the main vehicle for feedback, or for students trying to improve the quality of their work based on this information.

Observation 5

Different university faculty regularly use different sets of criteria, even for student works in the same genre. The sets may be derived in different ways. They may consist simply of those that the teacher feels most comfortable with and uses regularly. They may be compiled from surveys of existing lists by selecting those that are most frequently used. They may be adopted or adapted from a published set. Or they may be developed through direct negotiation with students to involve them in the assessment process.

In many contexts it is not practical to nominate all the criteria that could conceivably be used. Choices therefore have to be made. Selecting particular criteria is a decision to exclude others. Given the variety in criterion sets, questions naturally arise. Is one set as legitimate, or as effective, as another? Do the most commonly used or heavily weighted criteria from different sets have essentially the same 'coverage'? What effect would the application of different sets to the same piece of work have on the grade assigned? What signals do variations in the criterion lists from different sources send to students? In situations for which a range of criterion lists is possible, the fact that this multiplicity leads to little if any critical scrutiny of their respective validities indicates that the situation is not generally perceived as problematic.

Observation 6

Different assessors often agree on the overall grade for a particular work, but differ substantially about levels of performance on the separate criteria. In terms of grading, this may be tolerable, but differences at the level of criteria have considerable pedagogical significance. Among the arguments for using analytic approaches to grading is that they substantially reduce the labour involved in providing detailed diagnostic and explanatory information to students. In theory, students should therefore receive feedback that is both more substantial and more specific to guide future work. However, if the precise meaning of the feedback is dependent both on how assessors interpret the criteria and on where they set their thresholds for deciding on different levels on the criteria, one of the claimed benefits for analytic grading is significantly weakened. A corollary of this is that research into the reliability of analytic grading needs to go beyond the results it produces (the grades) if formative effectiveness is to be understood and promoted.

In most analytic schemes, there is no role for distinctly global primary appraisals. At most, there may be a criterion labelled 'overall assessment', but this often enjoys essentially the same status as all the other criteria. Taken at face value, overall assessment necessarily includes aspects also covered by the preset criteria, leading to double counting. Finally, assessors who run with the holistic and the analytic agendas simultaneously draw on their knowledge of the work as a whole in deciding on the levels reported in the tabulation. To the extent that this occurs, it serves to integrate decisions on the criteria with their holistic views. Students, on the other hand, see the criteria as separate qualities contributing to a grade, in isolation from a critical holistic perspective of the entire work. The reality behind the message sent is not the message that is received.

Contributing factors

Diverse though these Observations may appear, they are all symptomatic of structural inadequacies collectively referred to in this article as *indeterminacy*. This is a technical

condition denoting that a proposed solution system is incapable of producing, wholly within its own parameters, complete solutions for a given class of problems. Although it may be possible to obtain a 'solution' of sorts, this is usually an output that has the appearance of a solution but is actually sub-optimal. In this article, the problem is to determine the 'true' grade for a complex student response. The main elements of the solution system are the preset criteria and the rule for combining separate qualitative judgments on those criteria. The inputs are the assessor's judgments. (If the assessor decides on the grade simply by inspecting levels on the criteria, the combination rule is not relevant.) Indeterminacy in this sense is not the same as uncertainty, ambiguity, or lack of clarity, all of which have the potential to be resolved by providing more information of the appropriate kind. Indeterminacy has to do with a system deficiency that prevents fully valid solutions being obtained for the general case. For this predicament to be resolved, either the solution system must be changed, or the problem must be reformulated.

Significant aspects of the anomalies and irregularities outlined in the Observations can be attributed to a number of known factors, which provide the basis for a critique of analytic grading as the preferred solution system. The factors outlined below come from published research on human judgmental processes and the nature of knowledge. Some factors may partially account for specific anomalies, or for several anomalies. The list of factors demonstrates that most types of published grading schemes are based on simplistic models that fail to do justice to the task of evaluating complex student works. The factors are listed in two groups. In the first group are those that relate specifically to the use of preset criteria. The second group applies to grading models that use not only preset criteria but also mathematical combination. This latter group therefore does not include forms in which the overall grade is derived through visual inspection of the pattern of nominated cells within a rubric.

Factors related to criteria

The sampling effect. In many assessment situations, the set of specified criteria is actually a sample from a larger pool or 'population' of criteria. Such selections are not random samples, of course. They are usually made carefully, but the characteristics of samples do not match precisely those of their corresponding populations. In the current context, it follows that any selection of criteria cannot possess the same rich appraisal potential as does the population. Applying a sample of criteria to all student works increases the likelihood that a proportion of those works will exhibit significant features that lie beyond the scope of those criteria. For these works, bias is necessarily introduced by this narrowing of scope. Fitting a work to the template so that the completed rubric is made to 'account for' the quality of the work is reductionist in principle, and to that extent, artificial, with validity lowered as a result.

As a concrete example, consider a piece of written work such as an essay or term paper. Numerous lists and rubrics are publicly available to teachers, and commonly contain from 6 to 12 criteria. Regardless of which criteria are included in any particular list, it remains a sample. Behind it sits a much larger pool of latent criteria (Sadler, 1983). One such collection, which was assembled from published lists, contained over 50 criteria (Sadler, 1989). It was not exhaustive, and could have been extended further. Clearly, working with a manageable number of criteria has to involve selection, but at least for written works, any sample of reasonable size leaves out the majority.

A stringent test of whether a given set of criteria provides sufficient coverage to serve in the general case would have two components to it. The first would be an inability to find, create or adapt works that are undoubtedly of outstanding quality but do not excel on all the criteria. The second would be an inability to find, create or adapt works that are clearly strong on all the criteria but are actually not of high quality. The extent to which both tests can be satisfied would indicate the sufficiency of a particular sample of criteria in operationalizing the concept of quality.

Non-discrete criteria. Some properties of likely interest to academic assessors are concrete and unambiguous, including length (of a written work), or conformity with an external convention (such as referencing style). These share with many physical properties (such as mass, stiffness or pH) the ability to be measured, counted, or checked for compliance or accuracy. However, a large proportion of assessment criteria are not at all like these. They refer to concepts which are denoted primarily by words and their meanings, not to measurable properties. In ordinary discourse, words need to be interpreted in context. Any dictionary or thesaurus sets out a sample of the wide range of meanings a given word can take on in different contexts. Word-based criteria have the same versatility as words in general, but naturally are subject to the same limitations. Even those that may appear to have obvious meanings and straightforward implications for grading often stimulate debate when their meanings are probed. A consequence of the lack of unique meanings is that, within the same context, criteria may be interpreted differently by different teachers. They can also be interpreted differently by the same teacher in different assessment contexts.

The concept of using preset criteria assumes they are non-overlapping. That they turn out to overlap in practice is because their boundaries are fuzzy, and situationally dependent. In effect, some may be found to be substitutes for, nested within, or redundant in the face of, others. Clusters of criteria may have the same coverage as other clusters, but without one-to-one correspondence. Furthermore, although based on experience, analytic scoring schemes are typically devised in abstract settings prior to, and therefore dissociated from, the context of actually grading student responses to an assessment task. The meanings for particular criteria take part of their form from an assessor's interaction with student work. The initial impression of discreteness and independence (in the abstract) slides over into intersection when they are applied.

Intangible criteria. Criteria currently in use have been progressively identified and refined so that they now function more or less as concepts. Some that appeared at first to represent a single dimension or property have been later elaborated into several criteria. In theory, the process of subdivision could proceed to any desired extent. Once an evaluative characteristic has been recognised, it needs its own brief descriptor, regardless of its scope. Preset criteria are mostly denoted either by single words (such as *coherence*, *structure*, or *originality*), or by short phrases (such as *support for assertions*). The usual approach has been to choose an existing word or phrase that is already a reasonably good fit for the purpose, and if necessary, give it a nuanced meaning. At other times, a new word or phrase has to be created, perhaps by adapting a term or phrase from another language, especially if the concept already exists in the other context. As new criteria materialize and become familiar, they become part of the standard assessment discourse. Their use among students and teachers can be facilitated by a mixture of example, definition, and explicit differentiation from other criteria.

This process of criterion identification and clarification may be difficult in some fields. It makes only slow progress in developing fields where new forms of highly complex

works are being produced. A certain evaluative ‘characteristic’ may be acknowledged as important, but be found difficult to crystallize into a crisp term or set of words. In some cases, an extended description may be able to convey the gist of it. That may not be enough for other situations. When a characteristic is elusive, it can defy all attempts to express it in propositional form. In such situations, a partial solution may be to use metaphors or other linguistic devices. This type of solution is common in everyday conversations in which the speaker cannot think of a precise term that applies, or where the concept does not yet have a condensed form.

Some evaluative characteristics may actually be impossible to articulate, even in principle. ‘Criteria’ that can only be ‘tacitly known’ pose a significant communicative challenge. Tacit knowledge as a concept was introduced and given philosophical substance by Polanyi (1962) in his classic work on the topic. This knowledge does not come about by accessing and internalizing codified accounts expressed through language. It comes about only through experience, and can be communicated among people through sharing experiences in an environment characterized by mutual trust.

Regardless of the reason that some criteria are not expressible either in an economical verbal form, or at all, such criteria obviously cannot be included in preset lists. This type of occurrence has its parallels in many fields of human activity, including program evaluation (Sadler, 1980). The types of criteria classified here as intangible could be analysed in much greater detail, but space does not permit. Suffice to say that their existence and importance are widely recognized in the literature on tacit knowing and ineffable knowledge.

Factors related to numerical representation and combination

Interval properties. A fundamental principle in measurement on a linear scale is that equal numerical increments should correspond to equal increments on the underlying attribute being represented. On a 10-point scale, for example, the difference between the attribute levels represented by the interval 2 to 4 should be the same as the attribute difference represented by the intervals 5 to 7, and 8 to 10. Another way to say this is that all 2-point differences should be worth the same in the appraisal, wherever they occur on the scale. In practice, this is a difficult principle to achieve or even test for, yet the arithmetical operations carried out on the numbers assume constancy of worth across the scale.

The rule of addition. Combining the weighted scores on the criteria simply by adding them together is known as a *linear model*, and is a fully compensatory procedure. The latter characteristic means that low performance on one criterion can be compensated for by high levels on others. Everything is in the same melting pot. In practice, the use of a compensatory rule allows for a middling grade to be composed from moderate levels on all the criteria, or from a mix of extremely positive and extremely negative levels on different criteria, provided the weighted aggregates lie within the same grade range. Addition also produces more middle-level aggregates, because many more combinations of sub-scores are possible there, whereas the highest (and lowest) aggregates require consistently high (or low) levels on all the components. (The Central Limit Theorem in statistics is essentially an expression of the same principle.) This effect can also account for some of the apparent agreement on aggregate scores among different assessors, despite disagreement about levels of performance on the separate criteria.

A strictly applied compensatory rule can restrict a competent assessor. For example, an assessor may come to the view that a particular work is so outstanding on a single dimension that its shortcomings on certain others should be discounted or even ignored altogether. Such exceptional events could be accommodated in holistic appraisal, but not in analytic criterion-based schemes. The additive rule does not have a sound theoretical rationale, but is essentially a device of convenience. The processes for criterion identification do not have any obvious implications for the choice of a combination rule. In particular, the act of 'decomposing' a holistic judgment into 'its' criteria (Sadler, 1985) is not the operational inverse of 'recomposing' component judgments using an additive model. A possible way forward could be to explore different combination rules empirically, starting from a broader base of possibilities. In the field of decision analysis, for example, a variety of other combination rules have been identified and are well documented (Coombs, 1964). Two of these, the *disjunctive* and the *conjunctive*, are outlined here.

A conjunctive grading rule would require that a student work reach an expected (or minimum) threshold level on each criterion in order to qualify for a particular grade. Different threshold levels would normally apply for different grades. A disjunctive grading rule would require that a work reach or exceed a minimum threshold on just one criterion in order to qualify for a particular grade. For example, an assessor may regard outstanding performance on a particular criterion as singularly important for a work to qualify for the top grade. Even high levels on all the other criteria cannot make up for any shortfall.

Both conjunctive and disjunctive grading rules are nonlinear. They are also non-compensatory because neither allows unrestricted trade-offs across criteria. Implicit in them is the idea of qualifying (or not qualifying), on either specified or 'floating' criteria, rather than focussing on just the additive composite. In technical terms, the utility value of performance on one or more of the criteria is discontinuous, which means that certain achieved levels count more (or less) than others in specific but identifiable cases. The point being made here is not that conjunctive, disjunctive or blended grading rules should be promoted as potential replacements for the common weight-and-sum rule in criterion-based grading schemes. The argument is that an assessor's intuitive judgmental processes may operate naturally, at least partially, in one of these non-compensatory modes, rather than in the fully compensatory mode that is built into the weight-and-sum approach. A particular case is when a student work barely addresses the actual task specified. Criteria related purely to production quality (such as presentation, technique or finish) cannot then compensate for an inappropriate response, and are strictly irrelevant to that appraisal. To provide 'reward' by giving credit for them misleads the student.

What is more, assessors may be unaware of what their 'ideal' grading composition rule would be for a particular class of works, and may operate with different rules at different grade levels without compromising the goal of broad consistency of grading across students. The intuitive use of non-compensatory combination rules probably contributes to the discrepancies between analytic and holistic judgments.

Interactions. When the criteria are treated as separate variables, each criterion stands on its own. No account is taken of situations in which the co-occurrence of particular levels on two or more criteria contributes more, or less, to the overall quality determination than is reflected in the separate levels, either singly or when combined. Co-occurrence on two criteria is known as a two-way interaction. The typical analytic grading scheme does not allow for interactions at all, yet they can be pivotal to certain grading decisions, and therefore deserve a place in the feedback as well. The number of possible interactions, even for six criteria, is

considerable. Obviously there are the six primary criteria. As well, there are 15 possible two-way interactions, of which some may be highly influential in particular appraisals. (There are also over 40 possible three-way and higher order interactions. The extent to which the human brain is capable of processing interactions to that degree of sophistication is not yet clear.) When the non-interaction rule is relaxed, the way opens for assessors, whether teacher or student, to make what Kaplan called *configurational* (1964, p. 211) rather than componential judgments about quality. Holistic judgments are essentially configurational, placing a high priority on how the work comes together in its entirety. Imperfectly differentiated criteria are compounded as a kind of gestalt, and projected onto a single scale of quality through the integrative powers of the assessor's brain. Implicit in this process is the possibility of interactions of all orders.

Research into formal models of expert judgment

As currently implemented, analytic systems have a broad following and are widely accepted as 'best practice'. By claiming to be objective and open, they project strongly on both ethical and practical fronts. The argument in the preceding sections is that analytic approaches are structurally and practically problematic, the limitations being inherent in the formal procedures that use preset criteria, and more so when used with an additive combination rule. They promise more than they can deliver. Expanding the number of criteria does not solve the fundamental problem. The further such decomposition progresses, the harder it is to make the elements work together to provide an adequate portrayal of the whole. The logic of this phenomenon is obvious: whenever something is analysed into components, whatever originally held it together has to be either supplied, or satisfactorily substituted, if the sense of the whole is to be restored. The relationships and dependencies that should characterize an expert appraisal are lost when the criteria are treated as discrete properties. When component judgments are to be combined using a rule, employing more elaborated rules would also be unlikely to lead to a practical solution because of the degree of sophistication needed. Notwithstanding this bleak outlook, it is appropriate to examine briefly the research into models of human expert judgments.

Weight-and-sum analytic grading schemes bear an outward resemblance to certain formula-based models of expert judgments. Does this similarity provide a sufficient rationale for their use in grading? Much of the research to date has taken place within the discipline of cognitive science (Chi & Glaser, 1988; Ericsson & Smith, 1991; Meehl, 1954/1996), and some of the directions and findings are outlined in the next two paragraphs.

Where the aim has been to illuminate what goes on cognitively as decision makers process relevant information, the models have generally aimed to develop representations of those (often sequential) cognitive processes. Other models have been developed purely as mathematical functions designed to produce essentially the same decision outputs as experts would, without explicit attention to the psychological processes or sequences involved. Instead of specifically seeking to understand these processes, the latter have been intentionally blind to them. The primary focus has been on the accuracy of the 'decision' outcomes. The weightings in both types of decision functions have typically been regression coefficients, calculated from empirical data. This has allowed the models to be tested statistically.

For both classes of decision models, there has been an interest in their potential for practical application. In some cases this has been to facilitate predictive judgments about

future outcomes. A different type of application has been in automated decision making through the development of artificial intelligence systems that could 'make' decisions more quickly or more efficiently than humans do, or flag potential errors in critical judgments made by humans. The success of the modelling processes has turned out to be contextually dependent. In some situations, a formula for combining data has performed as well as, or better than, expert judges. In other situations, experts have consistently outperformed the models. In some contexts, regression coefficients have been replaced by equal or even random weightings with only a minor reduction in effectiveness. Despite the volume of studies completed, extant research into models of human multi-criterion judgments offers little that is directly transferable to the context of academic grading decisions.

Analytic grading of student works is fundamentally different on three counts. First, the criteria are not physical, social or biological characteristics whose measurements possess the properties of mathematical variables. (Comments about the status of criteria as 'variables' and the selection of criteria have been made earlier in this article, so no more is said here.) The same applies to any numerical coding of subjective decisions on individual criteria. Second, the choices of criteria and weights are arbitrary, in the sense that they are discretionary. This does not imply that any particular choice is capricious, accidental or unreasonable; criteria are presumably chosen carefully by individual or group decision. Despite this, the choices are not guided by any accepted principle, rule, convention or evidence.

The third count is that there is no external objective criterion against which to assess the validity of an analytic model. A *criterion variable* is an external benchmark that is not dependent on the elements of the model itself, and which is available to test the accuracy of model-based 'judgments'. The criterion variable may be an outcome (such as an economic index) that will become known only at a time later than when the input data are collected, or it may be the state of some phenomenon that is subject to confirmation or disconfirmation by independent scientific means or measurements (for example, as the result of a CT scan or pathology test). In the case of analytic grading judgments, the result of using the formula is taken as definitive in itself.

Of course, a single assessor's holistic judgments could be used as the criterion variable. The catch is that the weightings would be determined empirically, not by prior decision. This would run counter to the principle of advising students of the weightings in advance. A self-referent criterion variable would also prioritize the holistic grading decision, for which analytic methods are intended to obviate the need. Another option would be to use the average of several judges' appraisals as the external reference point. This would amount to intersubjectivity being interpreted as a form of 'objectivity', which is accepted as legitimate in many contexts of human experience (Scriven, 1972). Averaging holistic appraisals and analysing anomalies has long been a common approach to general-impression and other marking. It is also used in some recent approaches to 'calibrated' online peer assessments. Despite these various possibilities, an important concern remains. Holistic judgments, whether by a single assessor or a group of assessors, are not verifiable through means other than subjective judgments, and therefore do not satisfy the independence requirement for the purposes of validating the regression weights.

Limitations of traditional holistic appraisals

If a way forward is to be found by focusing again on holistic methods, traditional approaches are not up to the task. They are beset by significant shortcomings, which have been documented progressively since the 1920s. Studies specifically into the reliability of holistic judgments show a variety of inadequacies, especially when large numbers of essentially similar judgments have to be made. Irregularities attributable to fatigue, boredom, carelessness or capriciousness have usually been treated as random errors or 'noise', because in the main they are unintended and produce unpredictable variations. More significant are certain systematic effects in which patterns of various types emerge. A positive impression formed on the basis of a sequence of appraisals of works of high quality can carry forward and negatively influence the grade for the next work that is of only mediocre quality. This 'order' or 'serial' effect applies equally in the opposite direction. Extraneous information such as the assessor's personal knowledge of particular students' personalities, cooperativeness or diligence can affect grade assignments in what is called the 'halo' effect. Assessors may exhibit bias on such grounds as race, ethnicity or gender. The intrinsic difficulty of maintaining cognitive constancy over time can produce trends towards progressively greater leniency (or severity) as assessors work through grading large numbers of student works. Finally, in the higher education context, differences in the 'standards' different assessors adopt may be vigorously defended as an academic right.

Most of this research has been undertaken specifically in the context of holistic judgments, but analytic appraisals also depend on qualitative judgments. There is therefore little reason to expect that the smaller-scale judgments undertaken in the process of analytic grading would be exempt from at least some of the same patterns, or that inaccurate decisions on some criteria would somehow be offset by compensating decisions on others.

Up to this point in the article, the argument has been developed primarily in relation to student responses to assessment tasks in individual courses. The argument and its implications are, however, upwardly scalable to other higher education situations. Gaining ground in some provincial and national contexts and, to some extent at the international level, has been a movement towards detailed codification of achievement criteria and standards for higher education outcomes, with reference to both individual courses and full degree programs. The premise is that high levels of 'standardization' will lead to high levels of comparability in outcomes and records of academic achievement. These in turn would facilitate student mobility and credit transfer. Rational though this approach may appear, it suffers from the same structural limitations as appraising student works by the analytic rather than the holistic approach.

Charting a way forward

Sound theoretical bases for the processes of both holistic and analytic grading have not yet been developed, despite the articulation of their respective rationales, advantages and disadvantages. Historically, holistic appraisal was the original approach, probably because it intuitively appeared so natural. It prevailed for decades, more or less unchallenged. The concept of focusing on criteria specifically related to quality appears to have been first proposed by Burt (1920, p. 359). Cast (1939) later referred to this as the method of 'analytic' scoring. This development took place in the context of school-level written composition. Improving scorer reliability was the key objective, and the method then had no formative implications.

In higher education, early scoring keys and marking guides tended to focus on either the inclusion or omission of specific subject-matter content, or the structure of the response. For a written piece, this structure could consist of the following: Introduction, Statement of the problem, Literature review, Development of an argument or position, and Conclusion. A significant subsequent shift has concentrated on properties or dimensions related specifically to quality, in which content or structure may or may not play a part, depending on the discipline. The rationale for criteria-based analytic marking in higher education is of relatively recent origin. Basically, analytic marking purports to achieve certain practical ends, and to be a means for discharging a number of ethical obligations. The main aspects of current rationales are as follows:

a) Informing students of the marking criteria before they attempt an assessment task allows them to shape their responses more intelligently to the task specifications;
b) Comparing the quality of a student's work with fixed criteria and 'standards' is educationally more defensible than making comparisons with how other students in the course perform on the same or equivalent tasks;
c) Fairness requires that all student responses to the same assessment task should be appraised according to the same template;
d) Using explicit criteria enables students to understand the process by which their grades are derived, thus increasing transparency; and
e) Systematized approaches make for efficient and timely provision of formative feedback.

The move towards criteria-based assessment encapsulated in point (b) above represents an alternative to norm-referenced grading, which has typically been the default position. Norm-referenced grading is often considered unfair because students do not have control over the membership of the reference group, yet their grades depend on ranking relative to others. By externalizing the reference framework, criteria-based grading is also a counter to evaluating student works against such other framework as the teacher's unarticulated tastes or preferences, or each student's previous level of performance.

Points (a) – (d) in the rationale are actually worded from the viewing frame of preset criteria and the perceived need to treat all student responses to a given assessment task in identical ways. Setting this frame aside allows identification of the fundamental operational and ethical principles that lie beneath the 'solution' based on criteria and standards. This facilitates explorations into other potential solutions that nevertheless respect the underlying values, and allows competing solutions to be properly evaluated. These values are based on being scrupulously fair to all students. Fairness includes ensuring that students understand, before the grading event, the grounds upon which their work will be appraised (with a focus on openness, freedom from surprises, with no retrospective rationalizations) and assessing each student's work strictly on its merits. A different form of words that does not rely specifically on 'criteria' and 'standards' would be a commitment to ensuring (so far as possible) that students are inducted into an understanding and appreciation of the grounds upon which grading decisions are made.

Reliability is not everything, and rationales are not theories. Using an explicit model has been widely accepted and seen as 'the solution' to the grading problem, but that does not reduce the need for practical and ethical scrutiny. Clearly, indeterminacy in the use of preset criteria does not establish, by itself, a case for holistic appraisal. Given the absence of a proper 'theory of process' for either approach, all options need to be kept open pending further analysis and research.

To the extent that new approaches to holistic appraisal can be devised, there is the prospect of a positive way forward. These approaches should emphasize the creation of environments in which the critical discernment of quality becomes a key aspect of learning, drawing on what is known about connoisseurship in other contexts. Such lines of attack may turn out to be radically different from traditional approaches. In particular, students should be introduced to, and develop facility in, making holistic judgments for which criteria emerge during the process of appraisal. Analytic grading has long been explicitly recognized and advocated in higher education, but the concept of emergent criteria also has a long history. For example, Marshall (1968) referred to it as 'flotation'. He responded to his own rhetorical question 'Why not catalogue pertinent features?' with this: 'Prepared forms or lists destroy the whole principle … The basic idea in flotation is to let such distinctive and definitive features as may be noted to rise to the surface of their own accord. Only one or two of any list of qualities would fit, and important occasional ones would be missing.' (p. 134). (His comments were made in the context of avoiding the use of grades in summative assessment. His overall argument is not relevant here.) In an interesting shift of metaphor, Elbow (1973) referred to essentially the same process of emergent criteria as the *centre of gravity* approach to appraising student writing, in his case for essentially formative purposes.

Further potential for holistic appraisals

This section contains outlines of proposals in which holistic appraisal may have special potential. In many courses, the general sequence is that students attempt assessment tasks, submit their responses, have them appraised, and receive feedback. Analytic grading certainly provides one means of giving feedback. An alternative is to extend holistic appraisal to a context in which students themselves engage in making multiple holistic judgments of complex works, the source material being the work of their peers and the anonymized teacher's response to the same task. After each appraisal is made, students formulate their rationales.

As students make holistic judgments and develop skill in providing justifications, they position themselves to engage in conversations with peers and faculty about the nature and functions of the criteria they employ. Creating contexts in which this is a deliberate pedagogical strategy in its own right has the prospect of keeping the overall workload for teachers within the same total labour envelope as for more traditional teaching. Such an approach is outlined in a companion work to this article (Sadler, Forthcoming).

Potentially, two additional benefits could follow. First, inducting students into the same types of intellectual appraisal processes that experts use would substantially reduce the need for teacher-derived feedback. In the process, it would encourage students to adopt a holistic perspective of each work as a whole, and take into account both identifiable properties that rarely feature on criterion lists, and properties that are difficult or impossible to encapsulate in words. Developing evaluative expertise through guided practice would also equip learners to become self-critical and able to self-monitor their own work while it is in production, which ultimately is the very point at which it can make a difference to the work's quality. This would obviate some or all of the need for explicit feedback from the teacher.

The other potential benefit is as a means of improving the student's ability to tackle the problem or issue in an assessment task precisely as it is set, that is, taking the specifications literally. Academics often find that a considerable proportion of their students do not address the set task, and this causes them considerable frustration. It is of particular

concern for tasks that demand high-order intellectual or professional skills (analysis, critique, design, and synthesis). When students respond to these types of assessment tasks with a focus on either content (rather than what is to be done with the content) or lower-order outcomes (memorized knowledge, understanding, and application of routines), many academics nevertheless feel obliged to give at least some credit for it.

Part of the difficulty is that there are limits to what can be learned through giving and receiving feedback, that is, through being told. Students themselves need to see multiple exemplifications. As students holistically examine a variety of responses specifically through the lens of critique or appraisal, they can discover three things. First, they may be surprised to find for themselves that many works do not actually address the assessment task as set. They may then come to understand more powerfully than if they were to use criterion sheets how some works can barely address the set task at all, and why this creates significant dilemmas for making judgments about the quality of those works. Second, they can come to understand the wide diversity of ways in which valid responses can be constructed, and how quite different responses can be judged as worth the same grade. Third, learners can discover how a work that may be highly polished, technically sophisticated and inclusive of all essential elements can nevertheless fail to come together as a whole.

The 'true' grade and the begged question

One final point remains to be addressed before summing up. It was indicated above that the issue of adjudicating between analytic and holistic appraisal would be revisited. The question was, 'Which approach makes the stronger case for being able to deliver the 'true' appraisal of the quality of a student work?' The answer lies in the form of potential structural adequacy, which is an aspect of validity rather than an aspect of reliability. The argument in this article has been that the analytic approach is theoretically and practically deficient on two grounds. By limiting itself to preset criteria, it cannot take into account all the necessary nuances of expert judgments. Neither can analytic appraisal, when using a simplistic combination rule, represent the complex ways in which criteria are actually used. In principle, properly done holistic appraisals can do both. Therefore, the 'truer' representation is the 'fuller' of the two.

Conclusion

Assessment practice in many areas of higher education makes use of tasks that require students to produce complex works, which are appraised using multiple, often interlocking, criteria. Explicit grading models are in common use. These use preset criteria, and many also employ a formal rule for combining judgments on the criteria. The intended purposes include increasing transparency for students and achieving more objectivity in grading. Six anomalous patterns that arise in analytic grading have been identified, together with a number of known explanatory factors. The picture that emerges is that the overall process is not as robust as is commonly supposed. More sophisticated approaches, including new forms of holistic appraisal, merit further investigation.

Following an explicit model produces outputs (grades) that appear to have been substantially 'validated' through careful attention to all the steps. However, the model itself is characterized by indeterminacy, and is inherently weak. Furthermore, its implementation creates a veil of rigour that makes it difficult for learners to question either the process or the

outcome. Holistic appraisal, on the other hand, is not self-shielding in the same way, but the credentials of its traditional forms are not strong either.

Students need to develop a conceptualization of what constitutes 'quality' as a generalized attribute. They also need to be inducted into evaluating quality, without necessarily being bound by tightly specified criteria. This approach would mirror 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. To simply reach for a rubric or construct a scoring key each time a complex work has to be appraised is both impractical and artificial in life outside academe. Equipping students with holistic evaluative insights and skills therefore would contribute an important graduate skill. Clearly, there is more work to be done on this important issue of grading complex student responses to assessment tasks.

References

Black, P. & Wiliam, D. (2004) The formative purpose: assessment must first promote learning, in: M. Wilson (Ed) *Towards coherence between classroom assessment and accountability* 103rd NSSE Yearbook, (Chicago, National Society for the Study of Education, 20-50.

Black, P. & Wiliam, D. (2006) The reliability of assessments, in: J. Gardner (Ed) *Assessment and learning* (London, Sage), 119-146.

Braddock, R., Lloyd-Jones, R. & Schoer, L. (1963) *Research in written composition* (Urbana, Illinois, National Council of Teachers of English).

Broad, B. (2000) Pulling your hair out: crises of standardization in communal writing assessment, *Research in the Teaching of English*, 35, 213-60.

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

Burt, C. (1920) Tests of educational attainments. Memorandum 3 in C. Burt, (1947). *Mental and scholastic tests* (3rd ed.) (London, Staples), 285-366.

Butler, R. (1987) Task-involving and ego-involving properties of evaluation: effects of different feedback conditions on motivational perceptions, interest and performance, *Journal of Educational Psychology*, 79, 474-482.

Butler, R. (1988) Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on interest and performance, *British Journal of Educational Psychology*, 58, 1-14.

Cast, B. M. D. (1939) The efficiency of different methods of marking English composition: Part 1, *British Journal of Educational Psychology*, 9, 257-269.

Chi, M. T. H., Glaser, R. & Farr, M. J. (Eds) (1988) *The nature of expertise* (Hillsdale, New Jersey, Lawrence Erlbaum).

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

Crooks, T. J., Kane, M. T. & Cohen, A. S. (1996) Threats to the valid use of assessments, *Assessment in Education: Principles, Policy and Practice*, 3, 265-286.

Dweck, C. S. (1986) Motivational processes affecting learning, *American Psychologist (Special Issue: Psychological science and education)*, 41, 1040-1048.

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

Hodgen, J. & Marshall, B. (2005) Assessment for learning in English and mathematics: a comparison, *Curriculum Journal*, 16, 153-176.

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).
Kaplan, A. (1964) *The conduct of inquiry: methodology for behavioral science* (San Francisco, Chandler).
Marshall, B. (2000) *English teachers – the unofficial guide: researching the philosophies of English teachers* (London, RoutledgeFalmer).
Marshall, M. S. (1968) *Teaching without grades* (Corvallis, Oregon, Oregon State University Press).
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).
Messick, S. (1989) Validity, in: R. L. Linn (Ed) *Educational measurement* (3rd ed) (New York, Macmillan/American Council on Education), 13-103.
Morgan, C., Dunn, L., Parry S. & O'Reilly, M. (2004) *The student assessment handbook: new directions in traditional and online assessment* (London, RoutledgeFalmer).
Polanyi, M. (1962) *Personal knowledge* (London, Routledge & Kegan Paul).
Sadler, D. R. (1980) Conveying the findings of evaluative inquiry, *Educational Evaluation and Policy Analysis*, 2(2), 53-57.
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. (Forthcoming) Transforming holistic assessment and grading into a vehicle for complex learning, in: G. Joughin (Ed) *Assessment, learning and judgement in higher education* (London, Springer).
Scriven, M. (1972) Objectivity and subjectivity in educational research, in: L. G. Thomas (Ed) *Philosophical redirection of educational research* 71st NSSE Yearbook, (Chicago, National Society for the Study of Education), 94-142.
Shay, S. B. (2005) The assessment of complex tasks: a double reading, *Studies in Higher Education*, 30, 663-679.
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).
Stobart, G. (2006) The validity of formative assessment, in: G. Gardner (Ed) *Assessment and learning* (London, Sage), 133-146.
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).