Grading Student Achievement in Higher Education

Measuring or Judging?

Mantz Yorke

A “Natural” Activity

Grading attracts less attention than it merits because it involves practices which almost everyone will have experienced in one way or another. It is a “natural” part of the educational landscape. There is hence a tendency for assessors and users of assessments to believe that they understand grading and, perhaps, to place an unwarranted belief in the robustness of the grades that are awarded. Grading probably receives less attention in pedagogic development programmes than is merited when its significance for students is taken into account.

From time to time grading has been subjected to critique (e.g. Milton et al., 1986; Yorke, 2008), though the force of such critiques has hitherto become dissipated against the breakwaters of tradition. Continuing the critique, this chapter argues that grading is in general more complex and less precise than many take to be the case. For these reasons it is necessary to consider the merits of the reliance that is typically placed on grades, and how some of the problems with grading might be addressed.

The Significance of Grades to Society

The Importance Assigned to Grades

Grading scales of one form or another pervade higher education, apart from in a handful of institutions (such as Alverno College in the US) where student achievement is described in qualitative rather than quantitative terms. Grade-point averages [GPAs] and degree classifications are of importance to students. A GPA of 3.00 or an upper second class honours degree opens doors to graduates that would otherwise remain closed. The “first sift” undertaken by employers is often on the basis of the overall index of the graduate’s performance, and applicants for postgraduate awards in the UK almost inevitably need to have gained an “upper second”.

Grade Inflation

Despite the weight that society places on grades there are persistent claims of grade inflation, and hence some scepticism about the worth of grades. There is a substantial literature, mainly
from the US, arguing whether or not grade inflation exists. The argument is made complex by
the variation in the way that “grade inflation” has been defined (see Hu, 2005, and Yorke, 2008,
p.108ff).

Stone (1995), Rosovsky and Hartley (2002), and Johnson (2003) are amongst those who have
argued that grade inflation is a significant feature of higher education in the US, whereas Adelman
(2008), who has examined grades on a large number of transcripts, argues that the proponents’ case
is itself inflated, and that the general level of grades in the US has remained much the same over a
period of 20 years. Adelman suggests that the allegations of grade inflation are based on a highly
selective reading of the evidence, in which performances in elite institutions drown out those from
the less prestigious institutions. Following his review of the literature, Hu (2005) arrives at an
intermediate assessment of the position.

In the UK there has been a steady rise in the proportion of “good honours degrees” (i.e. first and
upper second class honours degrees) in recent years, which has predictably excited press comment
(from Clark, 2008, amongst others) about the “dumbing down” of standards. Yorke (2008) found,
from a study of awards made between 1994 and 2002, that grades were tending to rise more quickly
in elite institutions than in others in the UK—a finding that is consistent with Adelman’s reading of
US data.

Grade inflation is not a primary focus of this chapter, but Hu (2005) and Yorke (2008, chapter
5) provide convenient starting points from which to approach the issues involved. In particular,
they point out the difference between pernicious grade inflation and a rise in grades that may
be attributed to causes that are not necessarily deplorable, such as changes in the ways in
which curricula are designed and implemented, and in approaches to the assessment of student
achievement.

Standards

Standards are based on disciplinary and/or professional norms, implying often a qualitative
judgement on the part of the assessor (based on sedimented understandings of what the expected
standards are) as to whether the student has reached a particular level of achievement. Evidence
from reviews in UK higher education, whilst painting a generally very positive picture of pro-
vision, has nevertheless consistently pointed to weaknesses in the area of assessment. There are
relevant summary observations in Quality Assurance Agency (QAA) publications (2004, paras
13–14), while the archive of reports on the QAA website (found by navigating from www.qaa.ac.uk/
reviews/academicReview/default.asp) provides plenty of more detailed evidence on this point.

Standards are difficult to pin down with precision—the fuzziness of constructs relating to
standards and criteria has been noted by a number of writers (e.g. Sadler, 1987, 2005; Webster et al.,
2000; Tan and Prosser, 2004; Woolf, 2004). If standards and criteria are fuzzy, then grading against
them cannot be other than a fuzzy process. Earlier writers, encouraged by the “instructional object-
ives” approach that was stimulated in the second half of the twentieth century, argued for the
specification of achievements in precise terms, with Mager (1962) being a prominent advocate. The
apotheosis of this approach was arguably the system of National Vocational Qualifications [NVQs]
in England and Wales, whose assessment schemes involved evidencing performance to such a
finesse of detail across a variety of contexts that they proved unsustainable in the manner envis-
aged by Jessup (1991). Whilst many practical achievements at the basic NVQ Level 1 can be
evidenced with precision (see www.qca.org.uk/14–19/qualifications/index_nvqs.htm for general
descriptions of Levels 1 to 5), most of those desired of higher vocational levels (and a fortiori of
higher education) are inherently fuzzier.
Assessment Criteria and Standards

Lists of assessment criteria and assessment rubrics (for samples of the latter, see Walvoord, 2004: Appendix A) are common in higher education. They are intended to make clear to both students and assessors the standards that are expected. However, assumptions of interpretability may be rather too easily made. Wolf (1995), amongst others, has shown that statements of expectation and criteria need to be accompanied by exemplifications if meaning is richly to be communicated.

Assessment criteria are often specified with reference to excellence (as would be expected for the particular point in the programme reached by the student), which means that a low level of pass can be indexed against a set of negativities. The following example illustrates the expectation, in one UK institution, regarding the standard of third class honours at bachelor’s level:

Third class (40–49%)

A basic understanding of the main issues but not coherently or correctly presented

Third class answers demonstrate some knowledge or understanding of the general area but a third class answer tends to be weak in the following ways:

- descriptive only;
- does not answer the question directly;
- misses key points;
- contains important inaccuracies;
- covers material sparsely, possibly in note form;
- assertions not supported by authority or evidence.

(Higher Education Quality Council, 1997a, p. 26)

One could be forgiven for not understanding why a performance of this kind would actually merit an honours degree.

There is a broader weakness in the use of lists of performance descriptors. Performances rarely fit neatly into the “boxes” which define levels of achievement. A student may achieve above the specified level in one aspect of their work, and underachieve in respect of another. The assessor is faced with making a judgement about the general level of the work, which may not be achieved satisfactorily by adding up points for various components, as in “menu marking”. Sadler (2005) suggests subdividing criteria into categories of primary and secondary importance, but this does not provide an unequivocal solution to the problem, because of the unevenness of performances and the implicit restriction on achievements that are worth valuing but fall outside the formal specification. Studies, such as that by Saliu (2005), have been undertaken using a “fuzzy set” approach to grading, which allows that a performance can exhibit, simultaneously, different “degrees of membership” of the various assessment categories. However, whilst the logic of the approach has much to commend it, the challenges of implementing it seem insurmountable in practice.

Grade Distributions

Normativeness pervades grading. It lies in the background when topics and items for assessment are being selected. It comes to greater prominence whenever doubts are expressed about the frequency with which high grades are being awarded, since there is often an underlying presumption
that grades should approximate a statistically normal distribution. Yet any curriculum which is based on a conception of “mastery learning” is likely to give rise to a distribution which is negatively skewed (i.e. the bulk of the grades lie towards the upper end of the scale, accompanied by a thinner “tail” of lower grades) and not “normal”. This may be a contributory factor where a rise has been observed in awarded grades. When the curriculum is specified in terms of “expected learning outcomes” (or similar, rather than in more general terms), the students know what is expected of them and can target their efforts appropriately. This shifts upward the likelihood of obtaining a high grade—indeed, there should be no objection in principle to all students obtaining the highest grade for their work, since they could all show that they have achieved to a high standard the specified learning outcomes (the price may be a limitation on the amount of non-specific study undertaken by the student, but that is a matter that will not be pursued here). However, some external examiners and assessment boards look askance at such “top-ended” distributions of grades. Further, a problem arises in the computation of GPA and the like when the study units taken by a student are a combination of those with a bias towards norm-referencing in their grading practice and those where the bias is towards criterion-referencing.

The “Measurement Assumption”

There is a widespread—often implicit—belief that assessments of student achievement are tantamount to measurements, even if, at the time of their award, they are seen by markers as general signals of the level of achievement. The “measurement assumption” is made wherever grades are treated statistically, as if they are from measurement scales with common properties (such as when they are summed or averaged). Disciplinary and departmental cultures invest grades with particular meanings which undercut an implicit assumption of commonality in scalar properties. For example, does a percentage of 55 carry the same meaning in law (where, in the UK, grades tend to be low) and in some aspects of healthcare (where they can often be quite high)? Superficially similar, the scales on which such grades were recorded are in practice likely to be quite different. As Knight (2006, p. 438) pointedly put it: “True measurement carries invariant meanings.” The variance inherent in the use of grading scales means that grades cannot be treated as if they are measurements akin to those of physical properties such as mass and length (see also Dalziel, 1998, on this point).

Certifying

Knight and Yorke (2003) argued that considerations of reliability, validity and generalisability raised questions about the extent to which institutions could certify, or warrant, their students’ achievements:

- Some achievements can be assessed cheaply and reliably: these are often to be found at the lower taxonomic levels.
- Some achievements can be assessed with tolerable reliability, but only if the assessments cover a variety of situations. The costs that tolerable reliability implies may not be fully appreciated. An assessment centre may be better placed, in resource terms, to undertake some aspects of assessment on behalf of a putative employer.
- Some achievements, especially those that are complex, cannot be assessed reliably by an institution (e.g. placement performance), and all that may be possible is for the institution to attest that the student has undertaken the activity. It would be for the student to
represent their achievements to an interested party, perhaps drawing on a self-constructed portfolio of relevant evidence.

Not every educational achievement that is valued is susceptible to formal grading, since the grading is likely to lack technical robustness. Knight (2006) further argues that much summative assessment is context-bound and hence "local" in nature, thus prejudicing its generalisability (formative assessment is likewise "local", but its generalisability is of considerably less significance).

**Grading Practice**

There are many sources of variability in grading, whose interconnectedness cannot receive justice in a necessarily linear narrative. Many of these sources are to be found within individual institutions, whereas others reflect origins that are more broadly social in character (for a schematisation, see Hu, 2005, p. 11).

**The Fragility of Grading**

The disquiet expressed by some commentators regarding grade inflation relates in the main to overall indexes of student achievement. Rather less disquiet is evident at the level of the component grades that are aggregated in some way to provide an overall index. Yet a consideration of the practices involved in grading the work produced by students suggests that grading is a more fragile operation than many take to be the case. This may not matter very much when the purpose of the grading is formative—i.e. that it is intended as a signal to students of the level of their achievement (accompanied, one would hope, by some indication of ways in which the student might do better next time). It is when the grade acts as a summative statement of achievement that the fragility becomes a matter of particular concern. In the discussion that follows, the focus is on summative assessment at the level of the bachelor’s degree, though much of the argument applies, mutatis mutandis, to other levels of study.

**Sampling from the Curriculum**

What is assessed is a sample from the range of possibilities inherent in a curriculum or curricular component. Some intended learning outcomes are more challenging than others—for example, analysis, evaluation and creativity (at the upper end of Anderson and Krathwohl’s, 2001, revision of Bloom’s, 1956, *Taxonomy of Educational Objectives*) are generally more demanding than remembering, understanding and applying. The grades gained by students will be influenced by the assessment tasks that are specified. One cannot interpret what a grade might signify unless one knows what was being assessed. Not only is sampling from the curriculum an issue here, but so are the circumstances governing the extent to which the student’s performance was helped by a pre-structuring of the task (in contrast to the student independently having to work out an appropriate structure for tackling it).

The type of performance being assessed needs to be factored into an interpretation of the awarded grade. Evidence from the UK (e.g. Bridges *et al*., 2002) indicates that coursework assignments tend to attract higher grades than do formal examinations. This is not surprising, since the completion of an assignment in a student’s own time allows for access to supporting resources that are not available unless the examination is “open book” (of course, supporting resources are not an unalloyed blessing—they may be used in a proper academic manner or plagiarised). Placements
and practicums are difficult to grade (for reasons that are discussed later), and often are assessed on a pass/fail basis.

Approaches to Grading

There are, perhaps, two main kinds of approach to grading: holistic and additive. Holistic assessment takes place “from the top down”, in that the work is first appraised as a whole, this often being followed by a refining as it is weighed against whatever more detailed criteria may have been specified. Additive assessment takes place “from the bottom up”, in that the work is sequentially graded against a set of criteria and the overall grade is determined by addition of the separate grades (which are quite often weighted relative to the importance of the criteria). The difficulty—as those who have used both approaches can attest—is that the two approaches do not necessarily give the same overall grade. Studies show that assessors sometimes juggle the grades they give for components of the task until they arrive at an overall grade which they believe does the student justice (Baume et al., 2004; Hawe, 2003; Sadler, 2008). These assessors are reconciling what the additive approach tells them with their holistic appreciation of the merits of the work (and, perhaps occasionally, of the student).

Matters are complicated where the assessor brings into play criteria additional to those specified in order to take into account aspects of the work that are worthy of acknowledgement (see Sadler, 2008, and, for empirical evidence regarding the assessment of dissertations, Webster et al., 2000). The less tightly the task is specified (particularly an issue where creativity is involved), the more problematic becomes assessment.

Walvoord and Anderson, for example, suggest the use of:

a “fudge factor” of 10 percent or so that you can award to students whose work shows a major improvement over the semester. Or you may simply announce in the syllabus and orally to the class that you reserve the right to raise a grade when the student’s work shows great improvement over the course of the semester. (Walvoord and Anderson, 1998, p. 99)

This conflates two purposes of assessment—to index actual achievement and to encourage students.

Assessors, perhaps particularly in the nurturing professional areas, are quite often reluctant to fail students because they want to give them a further chance to demonstrate that they can satisfy the requirements of the course (see, for example, Brandon and Davies, 1979; Hawe, 2003). A reluctance to fail students can also occur because of the assessor’s awareness of the economic consequences for the institution or organisational unit, or because the assessor is aware of the commitment of time and energy that would be needed in respect of an appeal against the decision to fail (Newstead, 2002).

Assessors who are expert in both their discipline and assessment practice may be able to judge the general standard of a piece of work using a holistic approach, though they may find difficulty in mapping their judgement on to the specified assessment criteria. The difficulty arises, in part, from the desire to include non-specified criteria, and from the “intersections” between different criteria (see, for a critical review, Sadler (forthcoming). For a holistic approach to grading to be effective, it seems necessary for the assessor to be able to demonstrate the level of expert behaviour that is described by Dreyfus and Dreyfus (2005). The less experienced may find the additive approach helpful in assessing—and in developing their capability as assessors. Where the assessment specification covers both subject content and more “generic” matters (for example, the ability
to give an effective presentation), an assessor may have differential expertise across the specified requirements.

The use of the additive approach has an advantage where the assessor may be called to account, such as in institutional quality assurance processes, or when a student lodges an appeal against a grade. Of some relevance to the point, Hodges et al. (1999) found, in a study of medical practitioners, that junior personnel obtained higher scores for their use of a diagnostic checklist than did their more experienced colleagues—presumably because the juniors “followed the script”, whereas the more senior were able to apply learning from their experience to get to the heart of the diagnosis more quickly.

Finally, there is a slew of personal circumstances (some of which are noted by Pascarella and Terenzini, 2005, p. 66) that can affect an individual assessor’s grading, including:

- the number of pieces of work that have to be marked;
- the time available for marking;
- the order in which the submitted pieces of work are marked;
- their level of concentration on the task; and
- how tired they are.

**Disciplinary Variation**

There is plenty of evidence that grades vary across subject disciplines, in terms of means and distributions, whether one looks at the study unit or the overall index: evidence on the point can be found in Yorke et al. (1996), Kuh and Hu (1999), Johnson (2003), Adelman (2008), Yorke (2008a), and on the website of the Higher Education Statistics Agency (HESA) in the UK. In the UK experience, for example, high grades have consistently been hard to attain in Law (see Yorke et al., 1996, for data at module level and, by navigating via “view statistics online” at www.hesa.ac.uk, the tabulation of bachelor’s degree awards for 2005–06 published by HESA). At bachelor’s level in 2006, and not untypical of the previous decade, 5.4 per cent of UK graduates in Law were awarded a first class honours degree. The corresponding percentages for Social Studies, Historical and Philosophical Studies, and Engineering and Technology were 9.0, 12.7 and 18.2, respectively. Whilst entry qualifications across these subject areas differ, it is highly unlikely that they account for the wide variation observed.

The reasons for this variation are not transparent, but are likely to involve, *inter alia*, the kind of assessment task, the extent to which the task admits of an unequivocally correct response (thereby encouraging the use of the full grading scale), the approach taken to grading and the extent to which assessors adopt a “psychological set” in which encouragement or the valuing of a student’s progress is to the fore.

**Complex Achievements**

“Graduateness”

An overall index of achievement masks the profile of achievements of a student, in which some aspects are generally superior to others (whether this is judged by grades or by qualitative criteria). The former Higher Education Quality Council (HEQC) in the UK was challenged by government in the mid-1990s to investigate academic standards. The challenge arose because of concerns emanating from Asia about the robustness of academic standards. The threat to recruitment to institutions in the UK led to HEQC establishing the *Graduate Standards Programme*. The Programme’s
report (HEQC, 1997b) indicated the complexity of the standards issue, illustrating it with a grid in which the concept of “graduateness” was elaborated under the five main headings of subject mastery, intellectual/cognitive, practical, self/individual and social/people. It implied that students would vary in the profiles that would be judged as reaching an acceptable standard of graduateness, and that different kinds of profile might be valued according to how the student intended to exploit the bachelor’s degree qualification.

“Wicked” Competences

Consistent with “human capital” perspectives on higher education (see Becker, 1975, for the original proposition), there is increasing political interest in constructs like employability and workforce development. Such constructs highlight the complexity of achievements in so-called “real world” settings, and demand the display of what Knight (2007) called “wicked” competences. “Wicked” competences have quite a lot in common with the so-called “soft skills” that are valued by employers (such as the application of emotional intelligence in teamworking and other situations). Knight (ibid, p. 2) described them as:

achievements that cannot be neatly pre-specified, take time to develop and resist measurement-based approaches to assessment.

Such achievements often take longer than an individual module of study to develop, and may more appropriately be assessed over a complete study programme. This requires the assessor to resolve a “part-whole conundrum”, which demands a response to the question “what is the relationship between the learning outcomes specified for the programme as a whole, and those specified for its component parts?” A summary grade built up from a number of module grades (e.g. a GPA) may not do adequate justice to learning outcomes that are specified for the entire programme (see, for example, Sadler, 2007, pp. 389–90).

Problems outside the academy, such as those that can be met in work placements, practicums and the like, are often “messy” and unbounded, requiring more than a mono-disciplinary approach to their solution in circumstances in which the information base is less than ideal. In such circumstances there are some similarities with the “mode 2” approach to the production of knowledge that is described by Gibbons et al. (1994): however, some placements turn out to involve little more than menial tasks.

The assessment of complex achievements is challenging, but Knight (2007) provides some empirical evidence for academics believing that such assessment is not particularly problematic. Students’ effectiveness in dealing with complexity will be contingent on the prevailing circumstances. Assessments are therefore judgements rather than measurements (which, in reality, are no more than quasi-measurements at best). In Knight’s (2006) terms such judgements are “local”, and hence cannot be force-fitted into a specified list of assessment criteria. Grades can signal in only broad terms the strengths and weaknesses of such multi-faceted achievements, and an overall grade might obscure a significant weakness. A pass/fail approach to grading, supported by a commentary on strengths and weaknesses, may be more informative.

Transcripts and Portfolios

An approach to assessment that relates more to professional judgement than measurement does not fit with the ways in which students’ overall achievements are typically represented (in grade-point averages or in honours degree classifications), and pass/fail grades are typically dropped from
calculations of indexes of achievement. The “measurement problem” may be mitigated by disaggregating achievements, as in transcripts.

Transcripts of achievements in study units have been a commonplace in US higher education for a long time, though they were adopted rather more recently in the UK, as has been the diploma supplement across Europe. These “flesh out” overall indexes of achievement. However, and as noted above, the gradings in particular study units are subject to a variability in which disciplinary norms are influential. A way of dealing with the problem of variability, albeit not wholly satisfactory for various technical reasons, is to set a student’s grade against the generality of grades attained by the relevant cohort. Hence 55 per cent shows up as relatively high when the cohort modal percentage is 50, but as relatively weak when the modal percentage is 65. Other approaches of varying degrees of complexity have been proposed (e.g. Morrison et al., 1995; Johnson, 1997; Felton and Koper, 2005). However, they introduce what many might see as a statistical obfuscation, Johnson’s approach being particularly susceptible to such an objection.

In the UK, there is an expectation that, during their programmes, students complete “Progress Files” (see especially www.qaa.ac.uk/academicinfrastructure/progressFiles/guidelines/progfile2001.asp). These, originally recommended in the Dearing Report (National Committee of Inquiry into Higher Education, 1997, para 9.48), have two components: an official transcript of achievements, and a personal compendium through which “students can monitor, build and reflect upon their own personal development”, and on which they can draw when presenting themselves, for example, to potential employers. The compendium is the place in which students can record, and reflect upon, their experiences—particularly in the more open circumstances of placements and the like. If institutions are unable to certify some kinds of achievement, it falls to the student to make claims for them and to support them with the evidence that they have compiled. The personal compendium is a private document, and not for public presentation—indeed, any recruiter would recoil from a heap of such documents attached to job applications.

**Institution-level Considerations**

Brumfield (2004) conducted a survey of the way in which institutions in the US derived GPAs from the collections of grades achieved by students in curricular components. This indicated that, whilst there was some generality of approach to the computations (and also to the award of honours), there were some quite considerable variations in institutional practice. A smaller survey of 35 varied institutions in the UK (Yorke et al., submitted) and a survey of Australian universities (Australian Vice Chancellors Committee, 2002) also showed marked variation within the respective national systems.

In the UK-based study, for example, sources of variation included the following:

- the study units whose grades “count” in the determining of an overall index of achievement (for example, although students have to obtain the specified number of credits for the award, “award algorithms” vary regarding the number of module grades that have to be included in the computation of the honours classification);
- rules governing the inclusion of grades for modules that have had to be retaken;
- the weighting of performances at different levels of the programme when computing the honours classification; and
- the handling of students’ claims for discretionary treatment because of personal circumstances that may have affected their performance.

It is readily apparent that interpretation of a GPA or an honours degree classification risks being simplistic, since “local” considerations (Knight, 2006) influence what is reported.
An issue which may be of more significance in the UK than elsewhere is the near-complete use of institution-wide assessment regulations. This has evolved as institutional review activity undertaken by the Quality Assurance Agency for Higher Education (QAA) has asked institutions with varied regulations for different subject groupings why this should be so, and has made critical comment where divergence has been evident (critical comment in the educational press may also have been a potent driver of institutional change).

“Translation” between Systems

A further issue is the difficulty associated with “translating” grades achieved in one national system into the grading system of another. The way in which percentage grades are awarded in the US, Australia and the UK indicates differing assumptions: percentages tend to be highest in the US and lowest in the UK. Any inference that students in the US are more able than their peers in Australia and the UK is however unwarranted. Karran (2005) conducted a survey of overall indexes of achievement in Europe (where the Bologna Process—see www.europeunit.ac.uk/bologna_process/index.cfm—stresses the harmonisation of programme structures), and showed a diversity of scaling that makes problematic “translation” between one system and another.

Professional Development for Assessment

Approaches to the grading of work vary (e.g. Ekstrom and Villegas, 1994; Yorke et al., 2000; Hornby, 2003), with some of the variance being attributable to disciplinary norms. Most of the sample questioned by Yorke et al had learned their assessment practice from colleagues, a number had attended developmental workshops on the topic, and a few relied on their experience of having been assessed earlier in their lives. They had mixed views regarding the accuracy with which they could grade work, some pointing to “double marking”, moderation or some reliance on external examiners as providing a safety-net against rank injustice to the student. For various technical, social and structural reasons, such a safety-net has a much wider mesh than those who rely on it probably appreciate.

Relatively few academics have made a systematic study of assessment, and hence the institutional (and system-wide) capacity to support colleagues in respect of assessment is limited, despite the growth of teaching development centres in a number of higher education systems. Inexperienced assessors need to undergo a period of induction into assessment within their subject discipline. The development of institutional capacity in assessment should not be left to chance, and it is poor practice to leave the induction process to happen “naturally”: Johnston’s (2004, pp. 405–6) vignette presents a vivid picture of a person struggling to work out how to adjust her assessment practice to suit a context with which she was unfamiliar.

The induction process has to deal with the language in which achievements are represented, since there is plenty of scope for misunderstanding. Students do not necessarily share academics’ understandings of specifications of expected learning outcomes (see Maclellan, 2001), nor do they always understand the quasi-legal phraseology of regulations relating to assessment. Webster et al. (2000) pointed to the different meanings given by academics within a broad subject area to terms like “analysis” and “evaluation”. The potential for variation is increased as one crosses subject discipline boundaries; “analysis” has different connotations in, say, English, Law and Computer Science. Webster et al also pointed to the occasional lack of coherence between the grade awarded and the words used alongside, as in the following two examples (p. 76):
• “Real awareness of the various perspectives” (46% awarded); and
• “this is a clear, well presented [dissertation] . . . which fulfils its specific aims” (49% awarded).

What should a student make of such apparent contradictions?

Inherent Ambiguity

This chapter supports the contention that grades are “inherently ambiguous evaluations of performance with no absolute connection to educational achievement” (Felton and Koper, 2005, p. 562: emphasis added), and provides a basis for understanding the inherent ambiguity. However one goes about grading, the excellent student (according to the chosen assessment criteria) will end up at the top of the lists of performances, and the very weak student at the bottom. It is between the two extremes that the picture is haziest. Two students may end up in the same broad band of grading, but for very different reasons. To be useful, assessment information needs to be more finely grained than a single index. Users need detail if they are to make optimal judgements, as in recruitment: it is a different question as to whether they actually want it.

Walvoord (2004) gave her book the main title Assessment Clear and Simple. Making allowance for the fact that “assessment” in American usage has broader connotations than in the UK (since it encompasses programme evaluation), this chapter has shown that grading is neither clear nor simple, and that the path of developmental work stretches far into the distance.

References


Newstead, S.E. (2002). Examining the examiners: why are we so bad at assessing students? Psychology Learning and Teaching, 2(2), pp. 70–75.


