Education Inquiry
Vol. 3, No. 2, June 2012, pp.201–216

Assessment, evaluation and quality assurance:
Implications for integrity in reporting academic achievement in higher education

*D. Royce Sadler**

Abstract

The terms assessment, evaluation and quality assurance have various interpretations in higher education. The first two, assessment and evaluation, share considerable conceptual ground and interconnected histories. Quality assurance, on the other hand, is a more recent development. The issue of academic achievement standards in particular has significant implications for quality assurance. The first half of this article provides a selective broad-brush outline of the topics just described. The second half is about an emerging concept, grade integrity, which is focused on the trustworthiness of course grades recorded on student academic transcripts. This focus serves as a platform to illustrate: how difficult issues can be analysed; why established conventions and assumptions need to be challenged; and how ways forward can be sought out and thought through. The context for the paper is higher education but the principles also apply to other educational sectors.

Keywords: quality assurance, higher education, grade integrity, assessment, evaluation

Introduction

Over the past 15 years, international interest in assuring the quality of teaching and learning in higher education has intensified. The influences have included: political decisions to widen participation rates (the so-called massification of higher education); lowered minimum entry requirements; financial constraints (higher costs and student fees; lower public funding); increased dependence in many countries on fees from international students (a major income earner and export industry); and university rankings useful in marketing academic programmes.

A strong interest has also developed in the cross-border recognition of professional qualifications with two foci: the substantive content of degree programmes, and the mobility of students who undertake some of their degree studies in other countries with credit points transferred back to their home institution. Specific developments include the creation of the European Higher Education Area, and the Bologna and Tuning initiatives on the harmonisation of degree programmes (Enders & Boer, 2009; Kehm & Teichler, 2006). Explicit concerns have been expressed about academic

*Teaching and Educational Development Institute, University of Queensland, Australia. Email: d.sadler@uq.edu.au
 ISSN 2000-4508, pp.201–216

standards generally, possible grade inflation, and the standing and comparability of course grades and degree classifications (Brown, 2010; Hill, 2010; Hunt, 2008). (*Course* here refers to one unit of study in a degree programme.)

The first half of this article sets out an analysis of some general issues related to appraising, judging and valuing. The starting point is educational assessment and evaluation, because of their strong conceptual, historical and methodological links. Quality assurance is a more recent entrant into debates, practice and accountability and now occupies an important space of its own. The second half of this article is concerned with an emerging topic termed *grade integrity*. Focused on academic achievement standards, grade integrity constitutes a fundamental aspect of quality assurance. This topic is introduced to illustrate: how underlying goals and assumptions can be teased out and clarified; how longstanding traditions and practices can be critiqued; and how a forward agenda can be conceptualised and implemented. For the purpose of addressing the theme, this article selectively draws on, adapts and integrates relevant material in three notionally sequential articles (Sadler 2009b, 2010, 2011). This specific material is linked to the wider agenda of evaluating quality and quality assurance in higher education. However, in the interests of flow and readability, references are not necessarily made to the earlier work at every possible point. Interested readers are referred to the source papers for the full arguments with references. Although the broad context is higher education, many of the principles are potentially applicable to other education sectors as well.

Methodologically, the issue of quality determination or assurance is set against the backdrop of educational assessment and evaluation. The terms evaluation and assessment have been used differently in different countries and contexts. For example, evaluation in the United States in the 1970s covered appraisals (a neutral term) of student learning as well as of curriculum reforms and educational projects and programmes. Assessment meant something quite different. Especially since about 2000, assessment in US higher education has come to include everything to do with appraising the effectiveness, worth or value of institutional characteristics and functions – teaching, research, facilities, services, student support, organisational systems and student learning. This terminological diversion is meant to signal that, in referring to assessment and evaluation in international discourse, meanings have varied across countries and time, and may continue to do so. Regardless of particular contexts and interpretations, it is therefore important to be clear on the scope and nature of the object that is to be appraised or assured. Naturally, this influences the approach to be taken. A general distinction needs to be made between "process" and "product" so that the evaluative activity does not confuse or substitute one for the other, especially if one is less fraught politically, less expensive to do, or just simpler to operationalise than the other.

A prominent feature of higher education in many Western countries has been a considerable degree of academic flexibility within and across degree programmes

of a given type. Generally, degree curriculum has not been standardised across universities, or even necessarily within one university across its campuses. Specific minimum content and structure may be mandated by external agencies as part of accreditation processes that lead to registration or licensing for practice as a professional. However, once these are satisfied, considerable variation is accepted, even expected. Institutions, academics, learned societies and professional bodies would not wish strict uniformity to be imposed on curriculum, teaching and learning. Much the same applies on the research front in terms of freedom to inquire, but there is at least one notable difference.

In research, certain ongoing quality assurance procedures are widely accepted in principle, and well established in practice. Particularly relevant to the theme of this article is the conduct of peer review for evaluating research proposals for funding and manuscripts for publication in academic journals. Peer review, which for journal articles is often implemented in double-blind form, is regarded as fully consistent with professional collegiality and academic values, and in no way a challenge to academic freedom. Although acknowledged as less than perfect, peer review is not just accepted but actively endorsed as providing an adequate level of quality assurance. That same level of mutuality and collegiality is less common in teaching, learning and especially grading.

The quality agenda: Basic methodological distinctions

Quite separate from the distinction between assessment and evaluation in different contexts lies a distinction in kind between the purposes of assessment and evaluation on one hand and quality assurance on the other. Regardless of the object being appraised, the purpose of assessment and evaluation is predominantly to gather and distil information for reporting a conclusion or providing recommendations for decision makers. Both may also serve a monitoring and improvement purpose, feeding information into a programme or activity for enhancing the ongoing operation. These two purposes are commonly termed summative and formative respectively, following Scriven's (1967) terminology. His original context was the evaluation of curriculum projects and educational programmes, but the distinction is useful elsewhere as well. Quality assurance, on the other hand, is typically directed towards certifying, warranting or guaranteeing the quality of some entity. The concept of assurance implies that decisions or judgments are made against some background system of standards which are stable over time and accepted by relevant bodies as authoritative. Examples are those set for workplace and food safety; manufactured components intended to be interchangeable; and the performance of building and engineering structures.

The major and rapid changes in higher education over recent decades is reflected in the growing number and variety of higher education institutions in many countries, changes in the public—private divide, and a relative shortage of highly qualified academics as educators. Given this expansion, governments, institutions, employers,

academics, students and the public at large have raised questions about whether the quality of higher education is being adequately protected, that is, being properly controlled or assured (Westerheijden, Stensaker & Rosa, 2007). A key question is: What is the entity to be assured?

An obvious candidate is the quality of teaching, learning and assessment (Ryan, 2000). Before analysing those in detail, it is useful to examine some other perspectives drawn from the same era and context as Scriven's. In 1971 and also in the United States, Stufflebeam identified four specific domains helpful in evaluating educational and social programmes: Context, Input, Process and Product (CIPP). The point of this was to indicate to the evaluation community that an exclusive focus on programme outcomes was too narrow because outcomes are dependent to a considerable extent on factors outside the control or influence of programme operators. Translating Stufflebeam's domains into the current higher education context, CIPP comes out roughly as follows. Context includes: social settings of institutions; student characteristics; previous history, and internal and external cultures of colleges and universities; parameters set by independent agencies or government laws and regulations; and formal aspirations for higher levels of social inclusion and participation. Inputs include: financial, physical and electronic resources; student entry levels; teacher/student ratios; qualifications of academics and their subject matter knowledge; and degree and course structures. Processes include actual teaching activity, that is, the performance (competence) of teachers; student participation and engagement; online, campus-based and other forms of interaction; ethical practice; and assessment methods and grading decisions. Products (outcomes) include rates of retention, progression, graduation, and attrition; employability; starting salary on entry into the workforce; and student and employer satisfaction.

As a broad generalisation, the emphasis in higher education to date has been predominantly on Inputs and Processes plus all of the Products listed above. Considerably less prominence has been given to two omissions from the Product list: actual student learning (what students know and can do as a result of their higher education experience); and the integrity and comparability of grades, transcripts and qualifications. In some contexts, the rationale for these omissions has stemmed from sensitivities to, and potential reactions from, academics and institutions that may feel exposed or vulnerable. In other contexts, the rationale has been that individual academics must be free to grade students as they see fit, this being viewed as a constitutive element of academic freedom. This latter situation has been particularly the case in the USA and Canada (Hill, 2010). Whatever the reason, the conventional approach has been to prioritise inputs, processes, and a subset of outputs that downplays actual levels of achievement.

Quality assurance of teaching processes

How do teachers in higher education promote learning? A considerable number of researchers have explored this since the 1970s, examples being Biggs & Tang (2007),

Elton (1998), Ramsden (1991, 2003), Sherman et al. (1987) and Skelton (2005). In various ways, they have all attempted to identify salient dimensions or aspects of teaching excellence. As one might expect, the findings have overlapped considerably. Examples include: having clear objectives; being well prepared and organised; knowing the subject matter thoroughly; communicating skilfully; appropriately using information and communication technologies; sequencing material and managing time; teaching and testing for higher-order cognitive outcomes; establishing empathy with students; aligning objectives, teaching and assessment; actively engaging students; designing for research-led teaching; providing good quality feedback promptly; and being responsive to student evaluations of teaching.

Backing for these has typically come from analyses (some of them multivariate) of what excellent teachers actually do, or empirical studies of the effectiveness of specific changes or innovations, such as a particular way of giving students feedback. An alternative to empirical approaches is to view all aspects through the lens of a developed theoretical model or framework such as constructivism, and evaluate teaching practice according to its correspondence with the theoretical model (Bostock, 1998).

There are several advantages of a strong focus on teaching processes. First, it is consistent with an emphasis on actual proficiency or performance, and opens up directions for providing professional development for academic teachers. Professionals generally like to find ways of doing things better, more economically or preferably both. Second, students can be surveyed to rate teachers on particular aspects, such as the provision of feedback. Development and research on this front goes back at least to the 1980s in the USA (Braskamp, Ory and Pieper, 1981). In many countries, such aspects are regularly tapped into by national surveys of the students' experience, satisfaction or engagement, an example being the questionnaire developed by Ramsden (1991) which became the national survey instrument for Australian students – the *Course Experience Questionnaire*. Finally, teachers and institutions seem to feel more comfortable about improving their teaching strategies than they are about putting their grading judgments and standards under the microscope.

However, concentrating on a variety of separate dimensions or processes has its limitations. First, the validity of appraising a collection of observed aspects of teaching as a means of evaluating its overall effectiveness is not as well established as is the effectiveness of the individual aspects. The whole may be – and frequently is – more, or less, than the sum of its parts. This phenomenon is common to many areas of human experience but also, as Ford (1992) showed, a matter of straightforward logic. Further, although certain strategies, practices or techniques may work well for teachers and students on average, they may not work well for all, and conceivably could interact in counterproductive ways in some contexts, with negative consequences for students.

The philosophy of putting specified aspects together as a way of 'constituting' excellent teaching makes a deterministic assumption, which is that if all aspects were attended to meticulously, students would learn satisfactorily and achievement would

follow. The fact is that some teachers teach brilliantly but manage to break many of the standard principles of good teaching. They do not necessarily deliberately ignore the principles; they achieve teaching excellence in other ways, some of them highly idiosyncratic. Conversely, some teachers who display all the standard principles of good teaching are not effective teachers overall. Logically, one cannot disregard these two fronts if one is interested in appraising the effectiveness of particular teachers. As explained in Sadler (1985, 2009a), when judgements reached according to meticulous application of analytic (componential) processes do not agree with the holistic judgements made by competent experts, a strong case can be made that the latter should prevail because of the indeterminism inherent in using what amount to 'common-factor' criteria applied to all cases. This position is also consistent with that of Dewey (1939) and other philosophers that 'valuations' are primary acts of recognition, from which criteria are derived.

This leaves unanswered the question of how to identify effective or brilliant teaching. Could this be illuminated by looking closely and holistically at what brilliant teachers actually do – their professional practice – and emulating it holistically? Alternatively, could the quality of teaching *processes* be appraised, at least in part, by the quality and amount of student learning that actually occurs? The latter would require a clear idea of what constitutes achievement, high quality data on it, and perhaps evidence that high-performing students attribute a significant part of their success in learning to the teacher. Achievement would obviously need to be measured in some independent way, not by utilising marks, percentages or grades awarded by the teacher (Sadler, 2009b, 2010). Any move towards achievement as the final output is clearly a product rather than a process orientation. However, it presents sizeable challenges, not the least being potential negative reactions from academics and the difficulty of assessing the extent to which the teacher should be held accountable for student achievement. As Stufflebeam's (1971) model implies, the evaluation of outputs is not necessarily an adequate measure because learning depends on many student and other factors outside the influence or control of the teacher. Nevertheless, attempting it could throw valuable light on what effective teaching means.

A currently advocated measure of learning output is to use standardised tests of learning outcomes. These would produce comparative performance data across institutions and national borders. There are many precedents for this in the schooling sector. Although not specifically designed for this purpose, the USA-based Collegiate Learning Assessment (Benjamin, 2009) is an example of a broad-spectrum test of this type. Another has been proposed under the title Assessment of Higher Education Learning Outcomes (OECD, 2010). Both test (or would test) so-called graduate (generic) attributes (competencies, outcomes). Problems with this approach include getting agreement on the outcomes to be tested; operationalising the underlying constructs; motivating students to take the tests seriously or at all; agreeing on a technology for test administration and data analysis; potential distortion of the sub-

stantive curriculum by diverting teaching energies towards explicit coaching for the test; and finding useful ways to communicate the results. A significant problem of interpretation occurs because the same outcome labels or descriptors (such as critical analysis) have connotations which are distinctive in various fields, disciplines and professions (Jones, 2009). Publication of comparative performance data, a common practice in other educational sectors, could inform decision making by funding bodies, governments and consumers. In the worst case, this would result in naming and shaming low-performing higher education providers but with neither paths nor resources for improvement.

Before continuing, reflect again on what the quality turn is about. What is the underlying problem that is or should be addressed? Is it concern about educational outputs, namely, what students have learned, or are supposed to have learned, through formal education? Is it whether the enormous investment in education is providing the yield or dividend expected of it? What is the mechanism by which mass testing (of any kind) leads to improvement on the ground? Would investment in large-scale testing of graduates provide a satisfactory return? Would it lead to system, institution or teaching improvement, or higher student performance? From the point of view of attending to the integrity of course grades, large-scale, high-stakes testing would effectively sidestep the issue of curriculum-based academic achievement standards and, indeed, standard setting in any absolute sense. Finally, putting the issue of methodology to one side, suppose it were decided that the evaluation of teaching should emphasise just three variables, namely levels of student achievement, ethical teaching practice, and the student experience of learning. Quality assurance on the achievement front could then focus more intensively on what students have learned, the integrity of course grades and the standards behind them. A critical factor is the quality of the professional judgments made by academic teachers at the site of assessment and grading. This is the position argued through in Sadler (2009b, 2010, and 2011).

Approaches to assuring academic achievement standards

The primary motive for a focus on assuring course grades is that they should be accurate representations of the levels of students' achievement in courses. Clearly, standards in the sense of firm reference points for grading are a central issue. If standards could somehow be captured and used, they would provide a way to shift the agenda away from the relative and towards the absolute. But what are standards, ideally? What constitutes high achievement? Can standards be identified and made to stick? How could they be specified, shared or conveyed? Without answers to thorny questions such as these, any talk about standards is meaningless. However, the start made in 1987 by Sadler has been developed further in the three articles referenced in the preceding paragraph. A second motive for a focus on assuring course grades is that they then could, conceivably, provide some indication of overall teaching effectiveness, regardless of a teacher's micro-behaviours or strategies.

Traditional attempts to apply standards to course grades have followed one of two paths, both of which are generally canvassed in textbooks on educational assessment and grading. However, both have long been the subject of serious criticism. In a relatively recent contribution, Elton (2004) reiterated a set of longstanding concerns, lamenting that most of them had been raised over 30 years earlier by Oppenheim et. al. (1967). Relatively little had happened by way of systemic progress and improvement. The first path to grading practice was to grade by using cut-offs on aggregate scores. With this approach and assuming a 100-point scale, grades are allocated by assigning an **A** to all aggregates that fall within a fixed range, such as 85-100, a **B** to aggregates 75-84, and so on. This method of grading has a long history and is probably the most extensively adopted in higher education. Nowadays, it is not only still actively promoted but also commonly supported by institutional spreadsheets and grade books which simplify data entry and management. The technique is sometimes misnamed "absolute grading" because the grade cut-offs (85 for an **A**, 75 for a **B** ...) are decided in advance of the assessment results and are the same for all courses. Whitley and Keith-Spiegel (2002) endorsed it as a solution for academic dishonesty. The technique has also been classified as criterion-referenced, being supposedly characterised by openness, objectivity and comparability across courses. Any attempt to modify the cut-offs may be interpreted as meddling with academic standards.

However, the assumptions underlying this grading rule (and the one following) do not hold up when it comes to setting and holding standards (Sadler, 2009b). The fundamental problem with it follows from a basic property of measurement in education: the aggregates are not composed of standardised points or units, neither does a given score increment necessarily represent the same achievement increment at all parts of the scale. In addition, aggregates are usually made up of scores derived from all summative tests and tasks in the course, leaving the equivalence of score units derived from different instruments completely unexamined. Basically, there are as many underlying scales and units as there are assessment instruments. This is radically different from measurement in the physical sciences where properties such as length, mass and time – and all measures derived from or expressible in them – are measured in basic units that are standardised and have values which are either identical across their respective scales or, as is the case with decibels, given a precise mathematical non-linear formulation. Millimetres are millimetres, ohms are ohms, and calories are calories. Further, for any one dimension (such as length), measurements on one scale can be converted into exactly equivalent units on any other scale. Thus millimetres can be expressed in inches, feet or miles.

The second path has been to grade by proportions. Applied at the class or course cohort level, aggregate scores are first arranged in order. This list (or the corresponding frequency distribution if enrolments are large) is then partitioned into bands that contain predetermined proportions of the group; grades are attached to the bands.

The top 10% of students may be awarded an **A**, the next 20% a **B**, and so on. The choice of proportions is essentially arbitrary, but is often the same for all courses in an institution, the assumption being that this makes grades comparable across courses. Although often called grading on the curve, this phrase overlooks the fact that the shape of the frequency distribution is irrelevant. The thing that matters is that the proportions are controlled. Provisional grade allocations made at the course level may be scrutinised at a higher organisational level by a review board or panel. Grade distributions that are in line with institutional norms for the proportions are ratified; others are either negotiated with course directors or summarily modified by the panel until they conform. The same review procedures are not uncommon with grading by cut-off scores. In that way, grading by proportions assumes an override status for stabilising grades.

The philosophy behind grading by proportions is that, without surveillance of the final course grade distributions, the proportions of high grades may migrate upwards, thereby devaluing those grades. Although rarely stated, the rationale behind this methodology is the classic market approach to regulating value when there are no stable, independent reference points (Sadler, 2009b). Limiting the supply of a desirable commodity in the face of constant demand generally maintains that commodity's (market) value. "When the proportion of high grades is tightly controlled so as to keep them in relatively short supply, the worth of these grades is also high" (Sadler, 2005, p. 187). In grading by proportions, each grade represents relative position in the cohort, not an absolute level of achievement. As Guskey and Bailey (2001) point out, the approach actively works against standards. In summary, adopting a combination of relative scarcity and market forces is not an appropriate model for assuring the quality of course grades.

To rationalise grading by proportions by saying that achievement naturally distributes itself more or less according to a set pattern (such as a bell curve) is to argue that other factors should make no net difference to the levels of achievement reached. Holding grade proportions constant makes the award of grades structurally blind to: admission policies and student entry levels; the demographic profiles of cohorts; student–teacher ratios; academics' qualifications; resources for teaching; the quality of teaching itself; the availability and nature of support services; students' motivation to learn; and the quality of assessment programmes or tasks. More significant than all of those, however, is that controlling grade proportions disconnects the grades awarded from absolute achievement levels. This grading principle is structurally robust simply because it is fully self-adjusting. With each new cohort, the grading parameters are reset. However, the meaning of the grades is also reset. Clearly, with no external anchorage, the method is not an option for achieving grade integrity. Despite that, it is the grading rule specified for the European Credit Transfer System (European Commission, 2009), a central tool in the Bologna Process which aims to make national systems convergent.

A different approach which is currently achieving prominence relies on codification, that is, specifying standards by means of text-based (verbal) statements. Examples of codifications are: rubrics; intended learning outcome statements; grade descriptors; marking guides; criteria-standards matrices; and subject or discipline threshold (or minimum) standards. Compact and economical, codification is assumed to provide an efficient and effective means of knowledge transfer, the knowledge being about standards. This is why educators engage in 'writing' standards and nation-wide consultations take place to get consensus on the wording of standards. The rationale for codified standards rests on two assumptions. The first is that objectives, outcomes, criteria or standards can be stated clearly enough and comprehensively enough to enable markers to decide unambiguously on the grades that should be awarded for various levels of achievement. The second is that codification can communicate assessment expectations to students at the beginning of a course. Codification is advocated partly to emphasise that grading is not competitive but is against meaningful, concrete, objective and external referents that are accessible to assessors and students alike. But there is an additional catch. Clearly, the more general the statements, the easier it is to accommodate variation and arrive at consensus. The cost is that increased generality means less direction and specificity for decision making.

Besides those concerns lie two other problems. The first is that declarative or propositional knowledge, the kind that can be expressed in the form of written statements, words or symbols, is inadequate for expressing certain types of standards in an enduring and workable form. Common though this assumption is, its inherently problematic nature is widely recognised in many fields outside education, in particular, knowledge transfer in the commercial world (Cowan, David & Foray, 2000). The details of the argument are available in Sadler (2009a) but not recounted further here. The general invalidity of codification for educational standards persists as an unshakeable premise in the world of education. The second problem for higher education is that the content, design, construction, interpretation and application of codification are typically devolved either to individual teachers or to teaching teams. To the extent that grading decisions are made wholly within the parameters of each course, across-course comparability simply cannot be addressed.

Restarting the agenda: The concept of grade integrity

The short answer to the question of where achievement standards exist and reside is: In academics' heads. This may initially appear to provide no workable basis for progress, yet that is not the case, as the remainder of this paper outlines. But first, the special context of higher education is reiterated. Academics and their institutions often have considerable scope and autonomy in designing and delivering academic courses and programmes, within the constraints laid down by accrediting bodies. University teachers in many countries design or select their own assessment plans, items and tasks, and either grade their students' works themselves, or work in course-based teams

to do it. Under such conditions, how can the quality of grades be assured? What are (or should be) the underlying standards? Those are not the only challenges. To what extent can comparable standards be applied in different courses and programmes within a single university? In what sense can they be comparable across different universities if courses do not match one for one? If these appear hypothetical questions which are impossible or pointless to answer, bear in mind that the three grading principles mentioned above (fixed cut-off scores, controlled grade proportions and codification) exist and are widely implemented specifically to address such questions. This is because the aims have been and remain important to students, institutions, employers, governments and society at large. Despite the fundamental invalidity of those three grading approaches, their intention is not thereby rendered invalid.

Clearly, grades are intended as expressions of summative assessment, certifying levels of attainment students have reached by the end of their courses. Grade integrity is a shorthand way of saying that grades represent different levels of student achievement in as absolute a sense as possible, regardless of teaching and learning processes, course design, course sequencing and individual student learning paths (Sadler, 2009b). The logical first step is to clarify the desired end-point for quality assurance purposes. Too often, approaches are pursued with only vague ideas about what the most valued end point would consist of. This opens the way to casting about as to what is happening elsewhere and adopting it, especially if it appears to be novel, or is labelled best practice or evidence-based. It is also tempting to embrace a technological fix which is not entirely appropriate, or to make a quick decision based on political expediency. Sometimes putative solutions are brought to the problem, and the problem is re-jigged to fit the solution strategy. Another line of attack is to trust blindly in consultative or committee procedures which, despite collective knowledge, fall short of identifying potential causal chains of events that could, if they all worked, lead to success (or equally, to failure if they do not). In such situations, pursuit of the proposed approach can become an end in itself rather than the means to an end. It is crucially important first to be clear on the problem, and then let the characteristics of the problem drive the development of a solution. In the present discussion, the goal must remain to award with integrity grades that represent true levels of academic achievement.

Three requirements for grade integrity

As argued in Sadler (2009b, 2010), the key elements are fidelity, commensurability and comparability. Fidelity requires that grades represent levels of student *achievement*, and nothing but achievement. The focus must be exclusively on what students know and can do by the end of a course. This is important because in higher education institutions in some countries it has become common for students to be rewarded for effort, participation, contributions to learning environments, engagement, and producing draft work. Much of what students do in order to learn becomes included

in a judgment about achievement. This is a confusion of categories. The practice also serves to perpetuate extrinsic motivation – nothing is worth doing unless it counts towards the course grade. The practice of accumulating credits during a unit of study towards the end-of-course grade is frequently endorsed without qualm by academics and students alike. It is also required by some university assessment policies. The emphasis in grading should be not on engagement with learning activities or protocols, important though those are, but strictly on the level of achievement that is reached at the end of a course. The test for fidelity requires taking a literal interpretation of academic achievement and checking practice against this interpretation.

Fidelity is therefore a matter of definition and classification, not of measurement, and can generally be determined in straightforward ways. In addition to the activities listed in the previous paragraph, no credit should be awarded for purposes of praise, reward or encouragement. Conversely, no grading penalties should be applied for misbehaviour, lateness or even plagiarism. (The last two do need to be taken seriously and dealt with, but in ways that are completely dissociated from the course grade.) In short, unless all influences from non-achievement sources are deliberately excluded, course grades inevitably become contaminated. Fidelity in the object being assessed (which in this context are student responses to assessment tasks) rarely if ever features in the literature on assessment and grading, yet it is a precondition for grade integrity. Without it, grades cannot be interpreted properly. A second requirement for fidelity is practical: the evidence (student works) must be of sufficient scope and soundness to allow for strong inferences to be drawn about student levels of achievement. This implies that the quality of assessment tasks and how they are specified are also critically important. Poor assessment items and tasks provide low quality evidence at best, and inferences about achievement status are then compromised.

The second requirement for grade integrity is commensurability. Students deserve to have their work graded strictly according to its quality, with no easy grades, favouritism or concessions. Responses to the same or similar tasks should not be graded by comparison with responses from other students in the group (which would be a form of norm referencing) for the reason that students ordinarily have no influence over the membership and achievements of the other students in the reference population. Knowing relative standing is obviously important for some purposes, such as deciding on admission to advanced study or the award of scholarships, but rank ordering should follow from, not lead, the determination of grades. In addition, students deserve to have their work graded without regard to their individual histories of previous achievement. Commensurability implies that grades must be strictly in accord with the quality, breadth and depth of a student's performance as judged from the available concrete evidence and against fixed standards.

The third requirement for grade integrity is comparability. Students deserve their grades to have comparable value across courses in the academic programme in which they enrol, and across the institution. Courses should not exhibit characteristically

tough or lenient marking or grading. Students also deserve that their grades be broadly comparable across institutions and maintain their value over time, so that the standing of their qualifications is protected not only by the college or university in which they study but also by higher education as a social institution.

Achieving grade integrity

In this section, only a short sketch can be given as to the main processes involved in setting standards and assuring academic achievement grades. A slightly fuller outline is set out in Sadler (2011), but the fine details form the topic of ongoing research and development. The main approach is based on peer consensus. The simplest form of this, consensus moderation, is often used when several scorers mark extended responses to a single assessment task in a course which has large student enrolments. The steps are not fixed, but one common pattern runs as follows: a sample of student works is selected; all scorers mark them independently; scorers convene to compare the marks awarded, discuss their reasons, and come to a consensus on marking standards. The balance of student responses is then marked more or less independently, with some cross-checking of special cases, and the results are filed.

At the highest level of generality associated with the comparability of grades across courses and institutions, the process necessarily changes, but retains the essence of the original procedures. First, as before, it is based on the sampling of students. For each student in the sample, all responses to all summative assessment tasks in cognate courses at different institutions are required. They should be obtained in clean form (that is, unaccompanied by any information that could inform a marker about previous decisions or comments). The question to be answered, again by peer consensus, is: What grade should be assigned to the performance of the student, in each course, taking as evidence the student responses to all summative assessment tasks? To the extent that this integrative process can be achieved, it covers both commensurability and comparability (equal worth). The aim is to work towards consensus, calibrating academics against one another and, by an extension of the process, against societal needs and professional expectations. Between the moderation of student responses to a single assessment task and the comparability of grades across courses, programmes and institutions, each institution would be free to organise its own internal processes as it sees fit.

At this point, a return is made to the quality of teaching. If integrity in grading were achieved tolerably well, it would be possible to evaluate the quality of teaching and learning in a hitherto underdeveloped way, namely by accessing actual student achievement. Taking into account the caveats listed earlier in this paper, this would prioritise the actual effectiveness of teaching. It would also legitimate teaching styles in whatever form they take, whether or not they conform to set patterns or specifications. This proposition is underwritten partly by a philosophical commitment to variety in teaching approaches that are tailored to, or reflective of, student

characteristics and individual teachers, and partly by the assumption that the whole (namely, teaching quality understood holistically) cannot be adequately specified by way of defined elements.

To finish on the note of grade integrity and assured grades, the key thrust should be truth in labelling. Grades on student academic transcripts should be able to be taken at face value. This is surely a fundamental ethical imperative. Overall the aim should be to have better quality assurance of higher education through assuring course grades.

D Royce Sadler is Senior Assessment Scholar in the Teaching and Educational Development Unit at The University of Queensland, and Professor Emeritus in Higher Education, Griffith University. He has been teaching, researching and publishing on the formative and summative assessment of student achievement since 1973. Recent work has focused on assessment, grading and academic achievement standards in higher education. A Member of the Editorial Advisory Boards of two international assessment journals, he reviews manuscripts for several others as well.

References

Benjamin, R. (2009) *The CLA: What It Is Today, What It Will Be Tomorrow*. New York: Council for Aid to Education.

Biggs, J. and Tang, C. (2007) *Teaching for Quality Learning at University: What the Student Does*. Maidenhead, UK: Open University Press/Mc Graw-Hill Education.

Bostock, S.J. (1998) Constructivism in mass higher education: A case study. *British Journal of Educational Technology* 29, 225–240.

Braskamp, L.A., Ory, J.C. and Pieper, D.M. (1981) Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology* 73, 65–70.

Brown, R. (2010) The current brouhaha about standards in England. *Quality in Higher Education* 16, 129–137.

Cowan, R., David, P. and Foray, D. (2000) The explicit economics of knowledge codification and tacitness. *Industrial and Corporate Change* 9, 211–253.

Dewey, J. (1939) *Theory of Valuation*. International Encyclopedia of Unified Science, Vol.2, No.4. Chicago: University of Chicago Press.

Elton, L. (1998) Dimensions of excellence in university teaching. *International Journal for Academic Development* 3, 3–11.

Elton, L. (2004) A challenge to established assessment practice. *Higher Education Quarterly* 58, 43–62.

Enders, J. and Boer, H. (2009) The mission impossible of the European university: Institutional confusion and institutional diversity. In A. Amaral, G. Neave, C. Musselin and P. Maassen (eds.), *European Integration and the Governance of Higher Education and Research*. Dordrecht, The Netherlands: Springer.

European Commission. (2009) *ECTS Users' Guide*. Luxembourg: Office for Official Publications of the European Communities.

Ford, M.E. (1992) *Motivating Humans: Goals, Emotions, and Personal Agency Beliefs*. Newbury Park, CA: SAGE.

Guskey, T.R. and Bailey, J.M. (2001) *Developing Grading and Reporting Systems for Student Learning*. Thousand Oaks, CA: Corwin.

Hill, D. (2010) A contentious triangle: Grading and academic freedom in the academy. *Higher Education Quarterly* 65, 3–11.

Hunt, L.H. (ed.). (2008) *Grade Inflation: Academic Standards in Higher Education*. New York: SUNY Press.

Jones, A. (2009) Redisciplining generic attributes: The disciplinary context in focus. *Studies in Higher Education* 34, 85–100.

Kehm, B. and Teichler, U. (2006) Which direction for bachelor and master programmes? A stocktaking of the Bologna process. *Tertiary Education and Management* 12, 269–282.

Oppenheim, A.N., Jahoda, M. and James, R.L. (1967) Assumptions underlying the use of university examinations. *Higher Education Quarterly* 21, 341–351.

Organisation for Economic Co-operation and Development (2010) *AHELO: Assessment of Higher Education Learning Outcomes*. OECD: Newsletter – December 2010.

Ramsden, P. (1991) A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education* 16, 129–150

Ramsden, P. (2003) *Learning to Teach in Higher Education*. London and New York: Routledge-Falmer

Ryan, K.E. (ed.). (2000) *Evaluating Teaching in Higher Education: A Vision for the Future* (New Directions for Teaching and Learning Series, Vol. 83). San Francisco: Jossey-Bass.

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. (2005) Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education* 30, 175–194.

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

Sadler, D.R. (2009b) Grade integrity and the representation of academic achievement. *Studies in Higher Education* 34, 807–826.

Sadler, D.R. (2010) Fidelity as a precondition for integrity in grading academic achievement. *Assessment and Evaluation in Higher Education* 35, 727–743.

Sadler, D.R. (2011) Academic freedom, achievement standards and professional identity. *Quality in Higher Education* 17, 103–118.

Scriven, M. (1967) The methodology of evaluation. In R.W. Tyler, R.M. Gagné and M. Scriven (eds.), *Perspectives of Curriculum Evaluation*. (AERA Monograph Series on Curriculum Evaluation, Vol. 1). Chicago: Rand McNally.

Sherman, T.M., Armistead, L.P., Fowler, F., Barksdale, M.A. and Reif, G. (1987) The quest for excellence in university teaching. *Journal of Higher Education* 58, 66–84.

Skelton, A. (2005) *Understanding Teaching Excellence in Higher Education: Towards a Critical Approach*. Abingdon, UK: Routledge.

Stufflebeam, D.L., Foley, W.J., Gephart, W.J., Guba, E.G., Hammond, R.L., Merriman, H.O. and Provus, M.M. (1971) *Educational Evaluation and Decision Making*. Phi Delta Kappa National Study Committee on Evaluation. Itasca, Illinois: Peacock.

Westerheijden, D.F., Stensaker, B. and Rosa, M.J. (eds.). (2007) *Quality Assurance in Higher Education: Trends in Regulation, Translation and Transformation* (Higher Education Dynamics Series, Vol. 20). Dordrecht, The Netherlands: Springer.

Whitley, B.E. Jr. and Keith-Spiegel, P. (2002) *Academic Dishonesty: An Educator's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

EDU.
INQ.
VOL. 3