Document status. This paper is based substantially on a Keynote Address to the Annual Conference of the Education Research Group of Adelaide in September 2007. It may be shared with colleagues but may not be cited until published. Accepted for the journal *Assessment & Evaluation in Higher Education*, it was in press as of Mar 2009. Its final form may have small copy editor changes. r.sadler@griffith.edu.au

Fidelity as a precondition for integrity in grading academic achievement

D. Royce Sadler
Griffith Institute for Higher Education, Griffith University, Brisbane, Australia

Abstract

If a grade is to be trusted as an authentic representation of a student's level of academic achievement, one of the requirements is that all the elements that contribute to that grade must qualify as achievement, and not be something else. The implications of taking this proposition literally turn out to be far reaching. Many elements that are technically non-achievements are routinely incorporated into grades and thereby act as contaminants. A variety of credits and penalties are often included with the intention of helping shape student behaviours or improve their learning. Reversing the situation has ramifications not only for assessment and grading practices but also for the ways in which curriculum and teaching are conceptualised, designed and engaged in.

Keywords: grading, academic achievement, fidelity, validity, continuous assessment

Introduction

Internationally, it is common though not universal practice for degree programs to be made up of more or less discrete components, each of which is taught over a defined time period such as one semester. Each component course (alternative terms being module, subject, unit or paper) carries a fixed credit value which reflects its relative contribution to the whole degree program. Achievement in courses is assessed and the levels typically recorded as summative course grades on student academic transcripts issued by the teaching institution. As representations of achievement, course grades may be derived purely from formal examinations at the end of semester, or by weighting and combining marks (points) or scores arising from a series of assessment events administered at different points in the teaching period.

Decisions made or conclusions drawn on the basis of course grades can have significant consequences for students, teachers, academic departments, institutions, policy makers, and employers. Depending on the higher education and wider social contexts, grades and their derivatives can serve as input data for administrative decisions on progression through degree programs; admission to advanced studies; rankings for prizes, medals, honours and scholarships; determinations of degree classification; and accreditation and quality assurance. Grades can facilitate national and international student mobility with credit transfer, especially when they are expressed in, or converted to, a common metric. The information entered on academic transcripts can affect the employment and career prospects of graduates, and is especially important in the years immediately after graduation. In research, grades can be used as the criterion variable for investigations into such issues as the predictive power of entry scores, or the effectiveness of different approaches to teaching, learning and assessment. Using grades to draw conclusions or make decisions inevitably places a value both on grading as a practice, and on grades as commodities.

Grades may also be used for reporting performance on discrete assessment tasks, especially those that require extended or complex responses from students. These responses come in a wide variety of forms, and include term papers, field and project reports, seminar presentations, studio and design productions, clinical consultations, and creative works. Both course grades and grades for discrete works are directly relevant to the theme of this article. Degree classifications and levels of honours are not the main concern, even though they are often based on course grades.

Foundational to the theme of this article is the premise that grades should represent learners' attained levels of academic achievement, either in a course, or in relation to an extended response to an assessment task. In a scoping article by Sadler (In press), this property was termed 'grade integrity'. In essence, grade integrity implies that each grade should be strictly commensurate with the quality, breadth and depth of a students' achievement. The first of three requirements for this aspiration to be realised is that the assessment evidence be of a logically legitimate type. This aspect is termed 'fidelity' and provides the focus for this article. The second condition is that the assessment evidence be of sufficient scope and soundness to allow a strong inference to be drawn about the underlying achievement. The third condition is that decisions on grades should be made by evaluating the quality of student work against fixed external anchor points (standards) so that grading is not influenced by such factors as how other students in the class perform, or each student's individual history of previous achievement (Sadler, 1987). The full ramifications of the second and third conditions form the subject of ongoing research.

The structure of this article is broadly as follows. First, the concepts of fidelity and academic achievement are defined, specifically as they relate to grading. This is followed by an investigation of their joint implications for determining which elements could logically contribute to a grade, and which should be ruled out. This is important because of the considerable number of questionable components that have been frequently incorporated into grades, as if they constitute evidence of the level of achievement. The final section touches on the types of challenges which would have to be faced and addressed if all non-achievement components were systematically purged from the grading process.

Fidelity and academic achievement

Fidelity can be thought of as the extent to which something actually is what it purports to be, and is therefore true to type, concept or label. Fidelity is about the identity of the object concerned and how it is constituted, not about its utility for practical purposes. In the context of grading, fidelity is the extent to which elements that contribute to a course grade are correctly identified as academic achievement. In general, establishing fidelity involves a process of classification.

Standard text and reference books on assessing and grading student learning almost always refer to the reliability and validity of scores and grades. They seldom if ever raise the issue of fidelity. It also receives relatively little attention in both research and practice, except in the specialized domain of performance testing. In brief, performance testing is used when students attempt tasks that are designed to resemble closely significant real-life activities, and student performance on these tasks is appraised. An artificial situation may be necessary when it is impractical to test competence in real settings for such reasons as cost, risk, or testing efficiency. In executive training, for instance, an inexperienced person cannot be permitted to manipulate affairs or make important decisions in an existing firm primarily to assess capability. In performance testing, the correspondence between the artificial and real situations is referred to as fidelity (Downing & Haladyna, 2006). In training aircraft pilots and astronauts, fidelity refers to how closely the training simulator reproduces the characteristics of real craft in flight (Lee, 2005). Essentially the same meaning applies in contexts unrelated to education and training. In medical and social intervention research, fidelity refers to how closely the

treatment given to each patient, client or subject follows the specified regimen or protocol, or how faithfully an implementation of a program follows the original design (Calsyn, 2000).

The concept of fidelity has a certain structural resemblance to definitions of validity found in introductory textbooks on educational measurement and assessment. Here are three examples, the emphases being in the originals: 'The *validity* of a measure is how well it fulfils the function for which it is being used' (Hopkins, 1998); '*Validity* is an evaluation of the adequacy and appropriateness of the interpretations and uses of assessment results' (Linn & Gronlund, 2000); and '[T]o be valid, an assessment procedure must measure what it claims to measure' (Falchikov, 2005). Of these three, the first two are the most expansive in their implications; they emphasise measurement function and use. The third is the least expansive; it emphasises the veracity of measurement. What they all share is a central concern with measurement. This is the main characteristic that distinguishes validity from fidelity in assessment. However, fidelity and validity are connected. Any lack of fidelity places an upper bound on the maximum achievable level of validity.

As a term, academic achievement is used freely in the literature on teaching, learning and assessment but its interpretation is almost always taken for granted. It is rarely defined, explored as a concept, or listed in book indexes. The approach here is to start from first principles, drawing on both the common meaning and the etymology of achievement rather than to trying to infer an interpretation from existing assessment practice. In ordinary discourse, to achieve means to bring to fruition some significant act not previously accomplished, or to attain some significant performance status not previously reached. It means being successful in bringing about a desired end as a result of substantial effort, and clearly involves challenge. Normally, the magnitude of an achievement is proportional to the challenge involved.

Climbing Mt Kilimanjaro, rescuing a company from looming insolvency, or winning a major literary award would be significant achievements for most people. In the higher education context, completing a degree counts as a significant achievement for most students, as does having a scholarly article accepted in a reputable journal for an academic. In each case, an act is completed, finished or finalised and the result is clearly evident. Whether something should be classified as achievement depends to some extent on the context. Something that is accepted as an achievement for one person may not be so for another. Thus threading a needle would not count as an achievement for the ordinary person, but could be a significant achievement for a person with poor fine motor coordination who has struggled to master the technique. The following definition of achievement has been compiled from dictionary definitions and is used in this article:

> *achievement* (noun). A goal or level reached; an enterprise completed, accomplished, attained successfully, or brought to a successful end – especially by means of exertion, effort, skill, practice, or perseverance. [Etymology: Middle English, from Old French phrase (*venir*) *a chef*, and Late Latin *ad caput venire*, come to a head.]

This definition is close to what is generally understood in higher education, especially when the discussion is less about what actually constitutes achievement than on other aspects such as educational effectiveness, improvement of learning, or methods of assessment. In this article, achievement is taken as the attainment of an identifiable level of knowledge or skill as determined through evaluating performances on assessment tasks, or through observation of relevant behaviours in specified settings. Students are generally aware when data about their levels of achievement are being collected. The process of assessing and grading the quality, breadth and depth of learning is an inference based on primary evidence such as works that students construct, or sequences of physical steps.

Achievements can often be graded in terms of the level of accomplishment. In some domains, the challenges themselves are graded. Success at accomplishment counts as an achievement at the relevant level, and is often all or nothing. Mountain climbers, for example, classify peaks in terms of difficulty, and the novice works successively through progressively greater challenges. Completing a

task of specified difficulty just once qualifies as achievement. This also applies to setting a significant new record in an athletic event, or achieving a personal best. Routinely completing a particular task, even a difficult one, is not necessarily treated as a new achievement each time it is completed.

In other domains, a task (or set of tasks) is fixed, and the grading is in terms of the degree or quality of accomplishment. Achievement in higher education courses reflects this pattern. Depending on the context, the acquired learning may be referred to as knowledge, skill, proficiency, capability, competence or performance. To say that a person has learned something generally means that they can do, on demand, something they could not do before; that they can do it independently of particular others (such as a tutor or specific group of students, but not necessarily in strict isolation from others); and that they can do it satisfactorily. The last of these implies the existence of a minimum threshold, as reflected in the two-point performance scale, Pass and Fail. Partitioning the performance scale into labelled bands produces the familiar multi-point (A, B, C…) grading. Two-point and multi-point scales are used for both simple and complex achievement outcomes. Grade scales and symbols vary according to institutional or higher education system traditions.

This depiction of achievement clearly has an outcome or product orientation, and includes both knowledge and skill, extending to proficiency in handling complex design challenges or command over sophisticated technical procedures which need to be carried out with high levels of reproducibility. Such outcomes typically come about through prolonged high-level learning experiences, and form part of the professional capital of practitioners, which is portable from context to context. The events leading up to achievement, regardless of type or level, are open rather than fixed, and are strictly irrelevant to judgments about acquired levels of attainment. In particular, the processes, speed, conditions, and student experiences of learning play no role in grading it. That said, some paths might be more effective or more efficient than others, at least for some learners. It is often important to find out what these are, and then refine them progressively. These considerations are pertinent to the design of strategies for stimulating and guiding learning, but not to grading.

Achievements versus non-achievements

Determining whether a particular element is a legitimate component of achievement is a classification rather than a measurement issue. In taxonomic terms, conformity with the definition would be the (singular) character that has only two possible states: eligible or ineligible. Without entering into a discussion about whether borderline cases technically exist, the claim made in this article is that all or nearly all of the elements identified as non-achievements are unambiguously so, even though many of them have an associated rationale as to why engaging with them could be useful or important for learners. The deliberate focus on non-qualifying elements is because many of the things that are eligible for inclusion are, and should remain, matters that are constitutive of the relevant disciplines or professions. They are best debated and, to whatever extent is possible, resolved among experts in their respective fields.

Each of the elements listed below merits at least rudimentary treatment, but to go through them individually would be both tedious and unnecessary. Student effort, however, is treated in some detail to illustrate the type of thinking required. (An ancillary issue is how effort should be estimated, whether by hypothesising how much would most likely have been needed to produce the work, or by inferring from proxy variables such as class attendance or on-time completion of learning exercises. This would be relevant as a measurement issue but only if effort were a legitimate aspect of achievement.) Many academics cannot help but be impressed by the prodigious time and persistence that some students apparently invest in producing responses to assessment tasks. These teachers are intuitively disposed to take high levels of commitment into account in deciding on grades, because such commitment is not only commendable but also relatively uncommon. However, effort is clearly an input variable, and therefore does not fall within the definition of academic achievement.

The inclusion of effort in grading decisions has generated comment and controversy in the USA (Scott, 2005). This came about through the promulgation in 2004 of a grading policy at Benedict College, South Carolina, which required that student effort become a substantial component of all student grades in the freshman and sophomore years. The debate over the Benedict College policy arose from a clash between the principle of institutional authority and the principle of academic freedom, but this is not an aspect germane to the analysis of fidelity. Adding a mandatory effort component into a grade may be intended to solve the apparent equity problem that arises when extremely hardworking students achieve poorly. However it gives rise to a different equity dilemma: students who achieve extremely well but are able to do so without the high levels of effort required by others would have their potentially highest awardable grade capped by default. Their grades would then no longer be commensurate with their levels of academic achievement.

Nothing in the preceding comments calls into question the desirability of students' striving, persisting, recovering from temporary failures, and generally working hard to achieve. Effort is merely the cover term for what students put into reaching the goal. It is instrumental in attainment but not an identifiable part of proficiency reached or learning acquired. To treat effort as a goal in itself is a contradiction in terms: effort cannot come about through effort. There is no issue with the virtue of providing strong positive encouragement for students; the issue is with the appropriateness of incentives. Put in these terms, it is difficult to argue for mislabelling, for using logically irrelevant rewards, or for improperly including particular elements, all on the extraneous grounds that they assist in promoting motivational or other non-achievement objectives. The implicit rationale for rewarding student effort as an explicit component of the course grade is that effort takes on extrinsic value through its identification with true achievement, which is what the grade is supposed to represent. Besides, if the connection between effort and achievement were as direct and strong as is sometimes claimed, compounding them would amount to double counting.

Transactional credits and debits

Groups of aspects that are often included as achievement are listed below under the two headings, transactional and bestowed, both with credits and debits. Credits are extra marks or higher grades; debits are penalties applied. The boundary between the two categories can depend on the context, so are somewhat permeable. For instance, in a context where credit for effort is mandated and students know the rules of the game, the component is transactional; where credit for effort is rewarded entirely at the assessor's discretion, it is bestowed. The purpose of using accounting terminology is that appropriate naming helps to portray their essence and commonality, and to distinguish them from true achievement.

In general, elements that do not qualify as achievement are those that are socially or pedagogically enabling behaviours; matters of compliance that are not directly part of student works; student activities that enhance the learning environment for groups of students; or other activities intended to facilitate learning. Achievement data and learning process data need to be kept disaggregated, because their character and purposes are different. Process here refers to sequences of steps or activities which, when followed or engaged in, should (and often do) lead to achievement. Apart from being inconsistent with the definition of achievement, rewarding students who follow a particular learning strategy privileges those for whom it works well. This is incompatible with the empirical evidence that students learn in many different ways.

Academics and institutions know that students' voluntary behaviours may be shaped or steered by incentives (which operate prospectively, towards compliance) or penalties (which operate retrospectively, for breaches). Incentives and penalties can both be thought of as transactions between the teacher (or marker) and the student. The goods being traded are compliances and non-compliances; the prices and terms of trade are specified in advance; and the currency (being a mark, point, score or grade) is accepted by both parties. Because the rules of exchange are stated explicitly, the level of

compliance automatically generates rights for the student and obligations for the marker. Barring the outcome of any appeal, the award of summative credit completes each transaction, and all credits, whether for non-achievements or true achievements, thereafter have the same standing for purposes of the grade. Although the behaviours generally may appear to have a legitimate educational rationale in that they are effective in promoting positive behaviours or learning, the problem is that for the goods listed below, none actually qualifies as academic achievement. The history of giving credit for them has had the effect of continuing to legitimate the practice.

Transactional credits are marks or points awarded for such components as:

- Attendance at, or participation in, a minimum proportion of classes, group discussions, laboratory sessions or e-learning chat rooms, including contributions that enhance the learning environment for other students;
- Completion of specified activities, including practice exercises, log books, reflective journals on the learning experience, posts to online forums and discussion boards;
- Inclusion of a specified component in a work submitted (such as 'at least 20 references') unless the quality of that component is an essential element in the evaluation of the whole work;
- Completion of interim drafts or project stages, with emphasis on passage through fixed points in a preset sequence (believed to be the most effective learning path); and
- Engagement in particular processes used as the means of producing a particular outcome.

Transactional debits are marks or points deducted for such components as:

- Late submission of a response to an assessment task (perhaps on a sliding scale according to lateness);
- Non-conformity with regulative specifications, such as maximum word length (for an essay), using non-complying materials or media (for an artistic work), or other breaches of order or protocol; and
- Plagiarism.

Plagiarism is a special case. In many institutions, certain aspects of academic misconduct, namely fabrication or falsification of data, cheating, and abuse of confidentiality, tend to be treated differently from plagiarism. The latter tends to have a high profile and be treated somewhat separately, no doubt influenced by the advent of digital technologies. A common way of responding to plagiarism is to impose mark penalties, but these are inappropriate for dealing with it. An alternative is outlined towards the end of the article.

Bestowed credits and debits

In many settings, assessors exercise their prerogative to grant or reduce marks or grades for particular students, operating within the real or presumed scope of their professional academic authority. In other situations, changes may be authorised by the department or institution through explicit policy provisions, or implicitly through practices that are well established. In both cases, decisions are usually made for putatively good reasons, so the variations are neither arbitrary nor gratuitous. For example, an assessor may be generous in appraising the work of a student whose external circumstances are exceptionally difficult, the assessor deciding whether an allowance should be made and, if so, how much.

Referred to in this article as bestowed credits or debits, these types of adjustments are differentiated from the transactional variety because of their discretionary aspect; the student does not engage deliberately in actions that are rewarded by marks or grades. In many cases, it is understood

that a defensible case must exist, and that the assessor making the adjustment may be called upon to explain a decision to an institutional authority or to the student.

Marks, grades or grade cut-offs may be varied as the explicit means of:

- Conveying tangible praise for exceptional effort and persistence, as an indirect way of increasing motivation; or conversely, of conveying disapproval of uncooperative behaviour or apparent laziness;
- Rewarding significant improvement in performance to boost self-esteem, confidence, or the student's sense of accomplishment;
- Acknowledging risk taking, lateral thinking or new ideas, even when these are off target;
- Compensating for under-performance that is attributable to exceptional circumstances, such as acute health events, bereavement, or computer failure;
- Filling in missing data resulting from non-completion of some assessment tasks that would have provided clear evidence of learning (achievement) had they been available, a common method being to extrapolate from work completed to (hypothetical) performance on the whole;
- Obtaining a more acceptable distribution of grades, so as to maintain their perceived value, or to compensate for poor teaching or assessment;
- Improving the retention rate, or in other ways achieving a satisfactory throughput of students;
- Making a concession based on comparative disadvantage, such as limited competence in mathematics or the language of instruction;
- Forestalling negative personal consequences for the student, such as cancellation of a sporting scholarship, liability for additional fees, or delay in graduation; and
- Facilitating access to advanced studies, or enhancing career prospects.

None of these bestowed credits and debits is in the nature of academic achievement, neither do they provide evidence of it. Employing them therefore lowers fidelity.

Continuous and cumulative assessment

So-called continuous assessment is widely (but by no means universally) advocated and practiced throughout higher education. Also known as progressive assessment, it involves the use of periodic tests, assignments or other forms of assessable coursework spaced over the period of study rather than administered just at the end of it. (The strict interpretation of continuous refers to incessant or non-stop activity. Language purists may have preferred 'continual'.) Continuous assessment thus stands in clear contrast to assessment that is compressed into a single, major, end-of-course assessment event such as an invigilated examination. Many textbooks on assessment in higher education deal with continuous assessment. These include Brown & Glasner (1999); Bryan & Clegg (2006); Falchikov (2005); Freeman & Lewis (1998); Harris & Bell (1994); Knight (1995); and Miller, Imrie & Cox (1998). Generally, the authors discuss the relative merits of progressive versus terminal assessment from the viewpoints of students, academics and student learning.

The main advantages claimed for continuous assessment are that it: provides opportunities for formative feedback for students; motivates students to learn throughout a course rather than to cram at the end; reduces the emphasis on time-limited, make-or-break terminal examinations and the stress that goes with them; and affords opportunity for a wider variety of assessment task types and response formats than can be accommodated with formal written tests. In some courses, continuous assessment has replaced end-of-course tests altogether. Among the disadvantages is that continuous assessment can result in high workloads for academics and constant pressure for students. Despite these drawbacks, a common view is that, on the whole, continuous assessment is fairer for students and more facilitative of their learning than final tests. In addition, most technology-based learning management systems facilitate continuous assessment through online tests and quizzes, automated test

scoring and record keeping. Various styles of grade books and spreadsheets simplify the progressive accumulation of marks. At the end of the course, scores are weighted, added and converted into grades.

Without going fully into the pros and cons of continuous assessment, the issue specifically related to fidelity occurs whenever marks are accumulated across the learning period. Accumulation is so much part of the embedded practice and rationale of continuous assessment that the terms continuous and cumulative are often used interchangeably. Miller, Imrie and Cox (1998) explicitly define continuous assessment as that for which 'results for each piece of work contribute to the final result' (p. 34). Clearly, the everyday meaning of continuous does not imply this association. By definition, accumulation does require at least periodic assessments, but periodic assessments do not logically entail accumulation. Because the issue of accumulation lies at the heart of this discussion about fidelity, from this point forward continuous assessment is mostly referred to as cumulative assessment.

Consider the following portrayal of learning in a higher education course. Different students begin with differing prior experience and knowledge levels in the field. Through strategic participation in teacher-initiated learning activities, interactions with peers, and private study, students' knowledge and skill becomes progressively deeper, more extensive, more interlaced, and more sophisticated as the course proceeds. Time and effort are required for the learning to become integrated and mature, at least, as far as this is possible within the restricted period allocated to a single course. Obviously, the length of the learning period limits the degree to which progress towards mastery can be accelerated. Other things being equal, the peak levels of knowledge and capability are reached at or near the end of the learning period. The argument about to be mounted is this: to the extent that this portrayal reflects the way learning actually occurs, cumulative assessment in which early understandings are assessed, recorded and counted misrepresents the level of achievement reached at the end of the course.

Cumulative assessment and course objectives

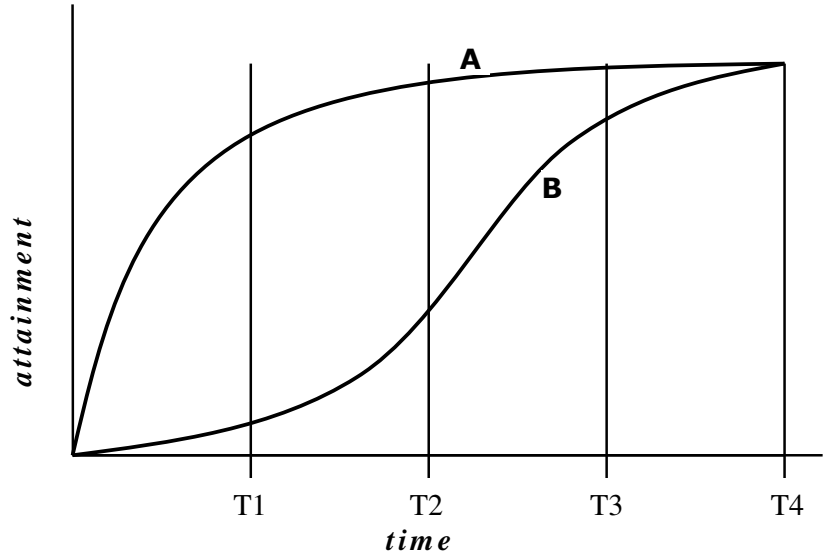
Statements about what a course, unit or module is basically about (its content or subject matter) and an outline of how students are expected to develop, change or grow as a result of the studies undertaken (what they will know or be able to do) form the aims, goals and objectives for the course. Aims and goals are generally broad expressions of aspiration; objectives are typically numerous, narrow and more specific. The usual way of expressing objectives is to prefix each set by a stem similar to this: 'At (or by) the end of the course, students are expected to …'. This provides a smooth grammatical flow into various similarly structured statements of intended learning outcomes. That lead-in stem is the focus of attention here because it refers specifically to the level of learning reached at the conclusion of the learning period, and hence as the culmination of learning in that course. Except for courses composed of independent sub-courses, cumulative assessment undermines that intention.

Cumulative assessment and the shape of attainment path

Every teacher in higher education is aware that attainment paths differ from student to student. Consider two hypothetical students in the same course who reach the same high level of knowledge and skill by the end of the teaching period. Suppose that one student (referred to as A) is able to grasp the material rapidly, more or less as quickly as it is introduced. The second student (B) initially struggles with the material presented early in the course. However, that situation (for B) turns out to be temporary and, as the course proceeds, the development of knowledge and skills accelerates, especially from about half way through the teaching period. Eventually, B reaches the point of being able to fit all the pieces of the knowledge jigsaw together, and achieve at the same level as A.

Figure 1 illustrates this two-student scenario. The variable on the X-axis is the elapsed time from the start of the teaching period and on the Y-axis, the level of attainment of the course objectives. The two curves labelled A and B represent the progressive attainments of the students over the course.

Suppose that the course provides for four progressive assessments, equally spaced across the teaching period. These are denoted by T1, T2, T3 and T4 in the figure. When the marks for these assessments are weighted equally and added, B's aggregate turns out to be about 65% that of A, even though their achievement levels at the end of course are identical. Student B's initial difficulties (reflected at assessment points T1 and T2) create permanent mark deficits that cannot be offset by later catch-up performances. The aggregate scores for these two students clearly reflect not only the final level of attainment but also the shapes of their respective paths.



Two hypothetical attainment paths

The 65% figure is for equal weightings for the particular curves shown. However, in many implementations of cumulative assessment, the weightings are unequal, being lower for early assessments (to reflect partial learning) and substantially higher for later assessments. The objective is to make due allowance for slow starters. For attainment paths that have the same beginning and end points, variation in the choice of weightings merely change the relative contributions of the attainment path shape.

In practice, attainment paths come in a wide variety of shapes, of which two more are now described (but not sketched). The first is for students who exhibit an attainment path that is essentially flat and at a low level for most of the learning period, and then shows an extremely steep late rise between assessment points T3 and T4. This follows abrupt insights into how everything falls into place. Referred to in the literature by such terms as Ah Ha experiences or sudden all-or-nothing learning events, these radical discontinuities and repositionings are well established phenomena which are recognised and documented both in relation to threshold concepts (Meyer & Land, 2005), and in the wider literature on learning and problem solving (Hadamard, 1945; Reber, Ruch-Monachon & Perrig, 2007; Ruggiero, 2002; Simon, 1983; Smith & Kounios, 1996). Regardless of whether these events have their roots in the conceptual subject matter of the discipline (as for threshold concepts) or in the cognitive processes of particular students (and so are idiosyncratic), they are not uncommon. Moreover, their occurrence, suddenness and timing are largely unpredictable. Some students who experience them are the very ones who complete a course with the deepest and most sophisticated knowledge of all. Ah Ha events are exciting for teachers to observe as and when they become aware of them in their students. (They are also exciting to experience for oneself.)

The second example of a different attainment path is for students who begin a course with considerable background knowledge and experience in the area but who do not demonstrate much

growth during the course. Such students find it relatively easy to perform well on the early material, but the course adds little to their initial levels of knowledge and skill. Putting aside the issue of whether recognition of prior learning could have been appropriate, cumulative assessment can generate aggregate scores for them that exceed those of other students whose levels of knowledge were lower to start with but who perform at considerably higher levels by the end of the course.

It is clear from just these four examples that a grade based on cumulative assessment reflects a mixture of entry knowledge level, final knowledge level, and attainment path shape. The relative contributions of each are unspecified and cannot be deduced from the grade. Cumulative assessment therefore raises practical, ethical and interpretive problems that flow inexorably from the decision to accumulate, not from any weaknesses in implementation. That they cannot be addressed by refining the processes is a matter of logic, not of empirical demonstration.

Most of the mark-based means listed in this and the previous sections are accepted as unexceptionable by course conveners and institutions, and also in the assessment literature. Some have been advocated as best practice and shown through empirical research studies to have a positive impact on learning (Isaksson, 2008 is but one example). In many institutions, this suite of practices is normalized to the point of orthodoxy and is rarely, if ever, perceived as problematic. Part of the reason has been the incremental and uncritical accretion of one practice after another. From an almost exclusive reliance on final examinations 50 or so years ago, course assessment programs have been progressively transformed. They now cover a wide variety of assessment types spread over a significant portion of the teaching period.

A final question remains: Should the concern about accumulation be scaled up to cover entire degree programs? Historically, some universities have had only comprehensive finals at the end of the degree candidature, with no accumulation, so it is certainly possible. A pragmatic response that is more in tune with widespread organisational arrangements is that courses and modules clearly dominate at present, and there is little sign that this will change in the foreseeable future. Courses as more or less intact entities provide flexibility for students in planning their degree studies, and generally possess a reasonable level of internal coherence. In addition, courses are generally coordinated by individual teachers or small teaching teams who have considerable power to transform their teaching and assessment creatively and concurrently.

Fidelity and formative assessment

This article is mostly concerned with the summative grading of academic achievement, but fidelity as a concept is equally pivotal in purely formative assessment (with no accumulation). Some of the elements identified above as non-achievements are relevant to formative assessment, and some new elements need to be added. The phrases ‘assessment of learning’ and ‘assessment for learning’ are commonly contrasted in order to reflect alternative emphases. Usually, the two mid-phrase prepositions provide the focus of differentiation, but different nuances of the term ‘learning’ are also implied. In discussions about the assessment of learning, the noun form of learning has a substantial overlap with achievement (as demonstrable knowledge), and reflects essentially a product orientation. In assessment for learning, learning has an active verb form, and the interest is in how assessment can facilitate the process of acquiring knowledge.

Assessment for learning, if it is to be done with integrity, is contingent upon judgments being based strictly on the quality of student works, free from extranea. That is where fidelity comes in. A statement that a student may reasonably assume to signify a demonstrated level of achievement or quality is misleading when its content is confounded by bonuses or penalties that are intended for other purposes but are not identified as such. This applies regardless of the medium used for coding formative judgments – whether grade, mark, or a written appraisal without a mark or grade. Such components may be employed for encouragement; for rewarding effort, persistence, risk taking or substantial improvement; for enhancing motivation or self-esteem; and as compensation for

comparative student disadvantage. On a different note, lowering the thresholds for interim judgments about quality rather than using the expectations that would apply at the end of a course also sends a muffled message that is difficult for the student to convert into action, and thereby achieve a formative purpose.

Challenges of implementation

The difficulty of mounting any critique of existing practice, whether or not related to assessment, is proportional to the pervasiveness of those practices, the status they enjoy, and their perceived role in bringing about desirable ends, especially ends that are related to the promotion of learning, justice, or due process for students. Treating fidelity in the object to be graded as a non-negotiable and overriding condition for grade integrity has major implications for practice. These implications merit a much fuller treatment than can be given in a single article, but a few examples are considered here.

Consider the situation of students who experience unexpected adverse events that prevent them from submitting work for appraisal that is essentially the same, and created under substantially similar circumstances, as that submitted by other students. When grading decisions are based exclusively on the quality, breadth and depth of student learning (rather than on comparisons with the performances of other students), the issue should be not so much whether the evidence submitted is identical in form with what other students submit but whether alternative evidence is of sufficient scope and of an appropriate type for a safe inference to be drawn. (Logically, those could also be the dominant considerations for all students, but practical efficiency generally rules out customising assessment programs for all students.) The common practice of interpolating or extrapolating from limited evidence raises the question of how much evidence is sufficient, and whether it is substantial enough. If a smaller amount is satisfactory for some (special circumstance) students, why would that amount not also be sufficient for all students?

The greatest challenge is probably in creating alternative approaches to the design of course assessment programs that serve both formative and summative purposes. Although including marks derived from inadmissible sources contaminates the appraisal of achievement, these contaminants are neither random error nor systematic bias, both of which are concerns in the reliability and validity of measurement. They result from deliberate decisions made for supposedly good reasons but without appreciation of their full implications. In theory, they could be removed simply by a policy decision that prohibits non-achievements from contributing towards grades. Among other things, this would make it impossible for mark accumulation to serve purposes believed to have learning, ethical, institutional or social justification. If some form of external incentive or reward were then considered crucial for bringing about changes in student behaviour, commitment or effort (an hypothesis which needs to be tested, given the considerable conditioning that has taken place over many years to support current practice), a creative approach to finding replacement strategies would need to be taken. Retaining mark-based incentives and penalties obviates the need to search for alternative strategies.

Apart from the means-ends issue, three deeply rooted assumptions are commonly made. The first is that students will not take seriously any learning activities or practice exercises unless their engagement with them contributes directly towards the summative grade. The second is that marks are a fungible commodity. Provided some defensible rationale can be produced, marks from all sources are deemed to possess equivalent worth and to deserve equal standing. Their provenance is neither tested nor seen as relevant, an observation that is consistent with the fact that fidelity as a concept has been all but invisible to date. The third assumption is that continuous (progressive, periodic), cumulative, and formative assessment are inextricably tied together, or even synonymous. In fact, these are three quite distinct concepts that are often confused; the distinctions among them are relatively straightforward but are seldom made explicit. This article is not the place to unravel the knot, but continuing confusion about them imposes limitations on a proper understanding of teaching, learning and assessment.

There is nothing about an assessment task – its intention, its timing, its structure, its substance, or the response format – that marks it out as distinctively formative. Assessment tasks and subsequent student or teacher actions based on them can function formatively or summatively. Many assessments are imagined to have a partly formative and partly summative role. But formative assessment, when interpreted strictly, occurs only when it fulfils its intended role, which is to lead to improved learning (Sadler, 1989). The purpose of formative assessment is not to generate evidence that will ultimately be incorporated into a course grade. Neither the structure of an assessment task nor the teacher's best hopes and intentions makes any difference to this. The defining characteristic for summative assessment in a course is that the result is, or contributes towards, a permanent record of academic achievement. This usually implies that it will have enduring consequences for the student. The crucial practical issue is the extent to which an assessment event can simultaneously achieve both purposes well.

For both academics and students, cumulative assessment banks credits for later withdrawal. The process of banking signals a sense of closure for each event. For students, this may leave them less inclined to look for ways to carry forward the information and implications to future assessment tasks. For teachers, each episode is for assessing particular material and outcomes, making it unnecessary to assess the final level attained. (If it were revisited, double counting would occur.) This situation promotes a teaching and learning environment that is incompatible with a primary focus on learning. Purely formative assessment in which the learning stakes for student are high but the grading stakes are nil frees up the learning environment. It allows students and teachers to be more imaginative and explorative, and to feel comfortable with trial and error, going up blind alleys, and occasionally having Ah Ha experiences, all as normal individual learning processes with a view to the development of substantially integrated knowledge at the end.

Plagiarism and regulative criteria

A strategy for dealing with plagiarism in a way that does not involve penalties scaled according to the seriousness of the breach is to shift the agenda away from the punitive and towards the creation of works that are eligible for appraisal on the grounds that they comply with certain restrictive norms. This approach will be introduced in general terms first, and plagiarism considered as a case in point. The starting point is with what have been termed *regulative criteria* (Sadler, 1983). These are sets of requirements that govern certain concrete aspects of responses to assessment tasks. The simplest regulative criterion for written works is the number of words; publishers' manuals of style usually contain a number of additional ones. For regulative criteria, the nature of the rules is quite straightforward. They generally require an administrative decision, not a qualitative judgment, to detect when a particular rule has been broken or a requirement not met.

In the context of higher education, the requirements can be codified, and cover a limited number of key aspects. Whenever initial screening for compliance detects a problem, the student is not given the full diagnostics and locations but a summary report and a short period of time to rectify the situation and make the work compliant. Learning to use the available tools and figuring out what to do in response is an important skill in its own right. If the additional time expires before the revised work is submitted, a new response to a different but equivalent assessment task is required, again within a specific time frame. If all deadlines expire, evidence that could be appraised is not available. The opportunity to be awarded a grade in the course then lapses and a null result (not a failing grade) is recorded. This in turn may lead to prescribed short-term or long-term consequences for the student, but not to lowered grades. The aim is to decriminalise submitting works that are in default and place full responsibility on the student to work through the submission of a compliant work. If the concept of compliant and non-compliant works were introduced to the assessment culture, students could be inducted into the basic principles involved as part of an institutional or departmental approach to assessment practice.

This approach is now applied to plagiarism. At the beginning of each course, students are advised that plagiarism, of whatever form and to whatever degree, is one of the nominated conditions that automatically makes a work non-compliant, so the procedures for regulative criteria apply. If plagiarism is detected, whether by a software detection tool or directly by the teacher, the appraisal process stops regardless of the extent of the deficiency and the work is returned for further attention according to the established protocols. In institutions that routinely process digitally submitted written student responses using such filters, students may be required to gain familiarity with these so they can check their own work for compliance, whether deliberate or inadvertent, before submission. This outline does not pretend to be a complete solution to the issue of plagiarism, but it does indicate a direction that may be worth pursuing further.

The second example of using regulative criteria is referencing style. For students putting written works together, getting the referencing to the required standard is a straightforward but nevertheless important procedural skill that needs to be mastered. Command over referencing conventions requires considerable attention to detail, but is hardly a high-level intellectual or professional outcome of the type that appears in lists of objectives. It therefore does not justify explicit credit towards a grade. Treating works as non-compliant on the basis of defective referencing style is a quick but powerful way to induct students into the exactness required. Some regulative criteria for written works are so heavily context dependent that automated error detection and decision making is difficult if not impossible. These include text properties such as structure, spelling, punctuation, and language conventions. The same principles apply for works of other types. For instance, creative works may have specific requirements as to the medium to be used. Using the wrong medium makes a work non-compliant; it cannot be graded as if it were a complying work, and the work has to be returned for reworking.

Regulative criteria apply when compliance and non-compliance can be determined quickly and unequivocally, and when the criteria are important in the technical sense but not pivotal to achievement of the main intellectual or professional outcomes of the course. Students in advanced higher education courses need to master, as a matter of course, the mechanical aspects that form part of the established conventions of their professional education or discipline. That many of these conventions are ultimately based on essentially arbitrary decisions does not reduce their importance. At the same time, the core properties of a student's work that should take precedence in appraisal and grading are those that deal with the quality, depth and extent of knowledge or competence. Routine matters need not intrude into the appraisal of higher-order academic achievement variables.

Conclusion

This article has its origin in concerns about certain classes of assessment practices that have progressively worked their way into many higher education teaching, learning and assessment contexts over a considerable period of time. The analysis shows that, despite their widespread use, many of these practices are inherently problematic. Two particular classes of practices, which are termed 'transactional' and 'bestowed', are outlined. These practices manifest themselves as systems of bonuses and penalties which boost or depress grades by including components that are in fact counted as student achievement, but in theory should not count. Argued as equally problematic is the practice of accumulating marks or scores across a teaching period in order to produce a final grade or result. Almost all of the examples listed can apply both to grading complex responses to assessment tasks, and to course grades. However, many additional possibilities exist, so the practices mentioned do not make up an exhaustive list.

Some of the practices identified have been mandated in institutional assessment policies. In other contexts, they have been uncritically accepted as normal or desirable. As numerous conference papers, journal articles and assessment handbooks can attest, many have been actively advocated and supported. Some have been tested for effectiveness in bringing about a desired end; there is little doubt

that they can be used for influencing student behaviours and attitudes, making allowances for perceived disadvantage, or filling in missing data. Despite what appears to be a broad authorisation for continuance of these practices, effectiveness is not the appropriate criterion. This is because implementing these practices inevitably compromises the integrity of the grades.

At its core, the problem is that a wide variety of non-achievement variables and outcomes have frequently been treated as if they were true elements of achievement. The extent to which the various components incorporated into grades technically qualify as achievement is termed fidelity. Perhaps not surprisingly, a key stage in the development of the concept of fidelity is clarification of the meaning of achievement. For a grade to provide a warrant of academic achievement that can be relied upon, the issue of how the grade is constituted – the 'object' of grading – is foundational.

As pointed out in the Introduction, fidelity is but one of three requirements for grade integrity, the other two being high-quality evidence from which to draw inferences about achievement, and the use of fixed standards of reference for assigning grades. Fidelity is not only a necessary condition for grade integrity; it is also a precondition for the other two. By determining eligibility for inclusion in the appraisal process, it sets sharp boundaries for the object that is to be appraised. Once that is settled, the issues of obtaining evidence for achievement and of appraising that evidence, and then coding the appraisal as a grade, are able to be addressed. Taken literally, this approach focuses attention on – and repositions – achievement as an end status or learning destination of considerable importance.

References

Brown, Sally, and Glasner, Angela. (Eds.) 1999. *Assessment matters in higher education: choosing and using diverse approaches.* Buckingham, UK: SRHE & Open University Press.

Bryan, Cordelia, and Clegg, Karen. (Eds.) 2006. *Innovative assessment in higher education*. London: Routledge.

Calsyn, R. J. 2000. A checklist for critiquing treatment fidelity studies. *Mental Health Services Research,* 2, 107-113.

Downing, Steven and Haladyna, Thomas. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ: L. Erlbaum.

Falchikov, Nancy. 2005. *Improving assessment through student involvement.* London: RoutledgeFalmer.

Freeman, Richard and Lewis, Roger. 1998. *Planning and implementing assessment.* London: Kogan Page.

Hadamard, Jacques. 1945. *The psychology of invention in the mathematical field*. New York: Dover.

Harris, Duncan and Bell, Chris. 1994. *Evaluating and assessing for learning*. London: Kogan Page.

Hopkins, Kenneth. 1998. *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn & Bacon.

Isaksson, S. 2008. Assess as you go: the effect of continuous assessment on student learning during a short course in archaeology. *Assessment & Evaluation in Higher Education,* 33, 1-7.

Knight, Peter (Ed.) 1995. *Assessment for learning in higher education.* London: Kogan Page.

Lee, Alfred. 2005. *Flight simulation: virtual environments in aviation*. London: Ashgate.

Linn, Robert and Gronlund, Norman. 2000. *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice Hall.

Meyer, J. H. F. and Land, R. 2005. Threshold concepts and troublesome knowledge: epistemological considerations and a conceptual framework for teaching and learning. *Higher Education,* 49, 373-388.

Miller, Allen, Imrie, Bradford and Cox, Kevin. 1998. *Student assessment in higher education: a handbook for assessing performance*. London: Kogan Page.

Reber, R., Ruch-Monachon, M. A. and Perrig, W. J. 2007. Decomposing intuitive components in a conceptual problem solving task. *Consciousness and Cognition,* 16, 294-309.

Ruggiero, J. A. 2002. "Ah Ha..." learning: using cases and case studies to teach sociological insights and skills. *Sociological Practice: A Journal of Clinical and Applied Sociology,* 4, 113-128.

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

Sadler, D. R. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13, 191-209.

Sadler, D. R. 1989. Formative assessment and the design of instructional systems. *Instructional Science,* 18, 119-144.

Sadler, D. R. (In press). Grade integrity and the representation of academic achievement. *Studies in Higher Education*.

Scott, J. W. 2005. Academic freedom and tenure: Benedict College (South Carolina): a supplementary report on a censured administration. *Academe,* 91, no. 1, 51-54.

Simon, Herbert. 1983. Alternative visions of rationality. In *Reason in human affairs,* ed. Herbert Simon, 3-35. Oxford: Basil Blackwell. (Series: Harry Camp Lectures at Stanford University, 1982). Edited and republished in *Rationality in action: contemporary approaches*, 1990, ed. P. K. Moser, 189-204. Cambridge: Cambridge University Press.

Smith, R. W. and Kounios, J. 1996. Sudden insight: all-or-none processing revealed by speed-accuracy decomposition. *Journal of Experimental Psychology – Learning, Memory, and Cognition,* 22, 1443-1462.

Note on the author

D. Royce Sadler is a Professor at the Griffith Institute for Higher Education, Griffith University, Brisbane. His research interests are in the assessment of student learning, achievement standards, grading principles and practice, and how assessment can be used to improve learning.