# Author Query Sheet

| Manuscript Information | |
|---|---|
| Journal Acronym | cSHE |
| Volume and issue | |
| Author name | Sadler |
| Manuscript No. (if applicable) | 370825 |

**AUTHOR:** The following queries have arisen during the editing of your manuscript. Please answer the queries by making the necessary corrections on the CATS online corrections form. Once you have added all your corrections, please press the SUBMIT button.

| QUERY NO. | QUERY DETAILS |
|---|---|
| | 0 author queries |

*Studies in Higher Education*
Vol. 00, No. 0, Month 2009, 1–20

Routledge
Taylor & Francis Group

# Grade integrity and the representation of academic achievement

D. Royce Sadler*

*Griffith Institute for Higher Education, Griffith University, Mt Gravatt CampusNathan, Queensland 4111, Australia*

In this article, grade integrity is defined as to the extent to which each grade awarded, either at the conclusion of a course or module of study, or for an extended response to an assessment task, is strictly commensurate with the quality, breadth and depth of a student's performance. The three basic requirements for this aspiration to be realised are, in order: assessment evidence of a logically legitimate type; evidence of sufficient scope and soundness to allow for a strong inference to be drawn; and a grading principle that is theoretically appropriate for coding the level of a student's performance. When further developed, the general approach outlined could produce positive side benefits, including ways of dealing with grade inflation.

## Introduction

A grade is essentially a symbolic representation of the level of achievement attained by a student. The grade symbols are usually alphanumeric characters or short verbal descriptors, such as distinction, merit, credit or pass. The main focus in this article is on grades that represent achievement in courses, which are the basic components of degree programs (some institutions refer to these as modules, units, subjects or papers). In this article, grade integrity refers to the extent to which grades are strictly commensurate with the quality, breadth and depth of students' academic achievement. It is the correspondence between the actual level of achievement and what the assigned grade is assumed to stand for, as judged from either explicitly stated intentions, or inferences from practice and usage. The integrity of grades has to do with their authenticity and provenance rather than their utility for various purposes, although logically any improvement in grade integrity should lead to an improvement in practical utility.

A secondary focus for the article is on grades that are assigned to student responses to assessment tasks, especially responses of an extended or complex nature. The classification or grading of a student's performance across an entire degree program comes within range only as grade integrity at the course level flows through to soundness of the degree classification. Internationally, using grades to represent levels of academic achievement is almost universal practice. In the USA, grading dates from the early 1800s (Smallwood 1935), and shows no sign of declining. Despite its general acceptance, it has long been the subject of critical comment on various fronts (See Milton, Pollio, and Eison 1986). Many of the enduring issues were summarised in a

*Email: r.sadler@griffith.edu.au

ISSN 0307-5079 print/ISSN 1470-174X online

DOI: 10.1080/03075070802706553
http://www.informaworld.com

1971 article by Cureton. Although based on an address she had delivered 12 years earlier, this article continues to speak to the present.

By definition, all student works that contribute to course grades are summative. These grades are formally recorded on academic transcripts issued by teaching institutions, and are then available for future reference and use. Grades may be pressed into doing double duty: formative and summative. Decisions made or conclusions drawn on the basis of grades can have significant consequences for students, teachers, academic departments, institutions, policy makers and employers. With one notable exception, grades are typically taken at face value, their integrity being presumed rather than tested. Part of the reason for this acceptance is that it is not a simple matter to detect in a conclusion or decision any impairment which can be traced back to a lack of grade integrity. The particular context in which grades are not taken at face value is when allegations of grade inflation are made: grade integrity is then the main consideration. Many assessment practices that are routinely employed in higher education institutions compromise grade integrity. As with the grades themselves, these practices are rarely challenged theoretically or empirically.

In terms of theme and structure, this article provides an account of a particular approach to improving the quality of grading practice. The first main part is an outline of why integrity in grading is important, both inherently and in how grades are used. Four particular threats to grade integrity are then identified and discussed. The last main section of the article is about how improving grade integrity would help address some grade inflation issues that have so far proved intractable. Parts of the article are analytical; other parts reflect pockets of existing practice. It is essentially a scoping statement that lays out the broad agenda for achieving grade integrity, but it is necessarily limited in what it can cover. A blueprint for a fully fledged system would require detailed attention to each aspect, and would need to evolve in the light of experience. Some current practices that pull in a diametrically opposite direction would need to be abandoned. Others could provide a useful basis for what is required if they were re-purposed and reconfigured as part of an institutional change in assessment and grading culture.

The overall aim is to provide a fresh perspective on an important problem, using terminology that is both intuitive and pertinent. Some of the development could have been set within the framework of measurement validity. However, since the early 1980s in particular, the literature on validity has become progressively wider in scope, and both more complex and more technical. The decision to follow a different route has been taken for several reasons, among them the wish not to be constrained by the discourse of validity, and the need for the basic ideas to be accessible to readers who have little or no background in measurement theory. Although some of the specific terms and concepts have so far had little if any profile in the literature on assessment and grading, they have clear parallels in the literature on evaluation theory and practice.

## Why the intrinsic value of grades matters

Using grades to draw conclusions or make decisions inevitably places a value both on grading as a practice, and on individual grades. This type of value is extrinsic, because it depends on the uses to which grades are put, and the practical consequences of doing so. Intrinsic value is different in kind. It is about how well a grade represents what it is supposed to represent, and any associated philosophical implications of that. The intrinsic value or merit (of a grade) lies at the heart of grade integrity. In the evaluation

literature, merit has been usefully distinguished from worth, which is the term applied for convenience to extrinsic value (Lincoln and Guba 1980; Scriven 1967, 1978, 1991). Scriven (1991) argued that merit is a necessary but not sufficient condition for worth. In real terms, merit is not an all-or-nothing affair, and data of even moderate merit can play a useful role in decision making under certain conditions, especially when selections are made from a pool of eligible candidates. The ideal, however, is high merit. This fundamental attribute makes no assumptions about existing or possible future uses. To illustrate why merit is important in its own right, consider the following propositions, all of which relate to fairness for students:

(1) Students deserve to have their work graded strictly according to its quality, without their responses on the same or similar tasks being compared with those of other students in their group, and without regard to the students' individual histories of previous achievement.
(2) Students deserve to know the bases on which judgments are made about the quality of their work. There should be few if any surprises.
(3) Students deserve their grades to have comparable value across courses in the academic program in which they enrol, and across the institution. Courses should not exhibit characteristically tough or lenient grading.
(4) Students deserve grades that are broadly comparable across institutions and maintain value over time, so that the standing of their educational qualifications is protected not only by the college or university in which they study but also by higher education as a social institution.

These propositions are underpinned by a set of ethical considerations. The first disallows the practice of assigning grades on the basis of inter-student comparisons. This is because each student ordinarily has no influence over the membership and achievements of the other students in the reference group. Knowing relative standing may be important for some purposes, but rank ordering should follow from, not lead, the determination of grades. This proposition also rules out comparing a student's current achievement with previous achievement, because that would be rewarding improvement (or penalising a drop in performance), not assessing achievement itself. Although bringing about improvement is the very point of teaching, if improvement as a variable needs to be reported separately for some purpose, it should be labelled as improvement, not as achievement. Finally, the first proposition disallows the assessor's knowledge of a particular student's typical quality of work from having any influence over the grade assigned in a particular course or for a particular work.

The second proposition bars assessors from appraising student works against the backdrop of their own personal 'standards', tastes or preferences, especially if students have no option but to infer these from scanty evidence (there is a particular reason for using quote marks here, which is explained later). This proposition also rules out any inclination on the part of assessors to exploit the imbalance of power that can be exerted over learners through the grading process. As a constitutive element of their higher education, students deserve to be given opportunities to understand the bases on which grades are assigned, and to develop proficiency in appraisal themselves. The last two propositions are self-explanatory and cover comparability across courses and programs, and consistency over institutions and time.

Each of the propositions is consistent with academic integrity, responsibility to the academic community and accountability to society at large. Each states a matter of

principle, leaving open the method of achieving it. All of the propositions have been couched in terms of what students deserve. They could have been expressed in at least two other ways, both of which have similar implications for teaching, learning and assessment. Casting them in the form of what students have the right to, instead of what they deserve, sets up the propositions to form part of a student Bill of Rights. Alternatively, expressing them as teacher or college obligations allows them to form part of an institutional code of practice. Regardless of how the propositions are expressed, a commitment to their substance makes grade integrity both an ethical and a practical imperative.

### *Assessing merit*

Any determination of merit needs to be made on two fronts. The first is the extent to which what is graded (the object of grading) qualifies as academic achievement, and is not something else. The second is the choice of the rule or principle governing how grades are to be assigned. Both of these are dealt with in more detail below. For grades to be true representations of academic achievement, the singular consideration should be how the level of achievement inferred from evidence compares with the minimum levels required for the different grades.

## Why the extrinsic value of grades matters

In practical terms, the worth of grades depends directly on the significance of whatever decisions or conclusions are based on them. Apart from the real costs of generating supplementary or alternative data, two rough indicators of worth are: the extent to which grades are trusted and relied on for important decisions or conclusions; and the opportunity costs of making decisions without using grades at all, or by using them in a way that results in poor decisions. In each of the decision contexts outlined below, grades play a major role. In some cases, the value they add to decision making can be subjected to statistical investigation.

### *Goals and motivation for students*

Grades form a concise method of conveying levels of academic achievement and expressing them in a common currency. Although some teachers, departments and institutions are philosophically opposed to the practice of grading, partly on the grounds of its extreme reductionism, there seems to be something holistic, unitary and integrative about a grade which transmits a message that is not necessarily conveyed by detailed evaluative description. Grades can have a profound and positive impact on a student's sense of achievement, acting as goals that provide motivation to engage productively with, go deeper into, or push beyond course material.

The effect of goals on performance has been the subject of a great deal of research for over 40 years, one decade of it being reviewed by Locke et al. (1981). Recent extensions have included the six sigma phenomenon in industry and commerce (Linderman et al. 2003; Smith 1993). In general, research in a wide variety of field and laboratory settings has shown that hard goals have the greatest impact on performance. Hard goals are specific and clear rather than general or vague, difficult and challenging rather than simple or easy, and closer to the upper limit of an individual's capacity to perform than to their initial level of performance. Hard goals act to focus

attention, mobilise effort and increase persistence at a task. Research has also shown that do-your-best goals are little better than having no goals at all.

Grades by themselves can provide only extrinsic recognition and reward, but the learner's internal psychological response to grades can provide a platform for intrinsic motivation. The magnitude of this effect depends on the credibility of the grades. Many students are capable of discerning among inflated, token and high-merit grades, especially (but not only) when they are able to gauge for themselves the quality of their own work, independently of external judgments. The experience of arriving at a significant achievement destination flows from a personal sense of reaching a level that is either high in absolute terms, or considerably higher than the level previously attained. Grades can act as a support for this platform only when they accurately represent genuine achievement. The worth of a grade (in this case its utility in promoting motivation) is then contingent on its merit. To paraphrase words from an ancient playwright, 'For those who experience hard-won success, the joy of achievement makes it worth all the effort' (Aeschylus, 458 BC).

### *Institutional administration*

In many institutions, course grades provide the basis for a summary statistic for each student known as the Grade Point Average (GPA). The GPA is a weighted mean of course grades calculated over a defined period of study, such as one semester. The cumulative GPA takes into account all studies completed from the time of enrolment in an academic program up to the time of calculation. The weights reflect the relative contributions of courses to the program, measured in arbitrary units such as contact hours, student workload, course credits or credit hours. GPAs can serve as input data for 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 facilitate national and international student mobility with credit transfer. In all these decision contexts, the merit of grades is taken for granted. Furthermore, basing decisions on grades gives the appearance of being both objective and meritocratic (the latter term being used in its non-technical everyday sense). This almost always delivers administrative solutions when required, so there is usually little incentive to place the grades under close scrutiny.

### *Employment and professional membership*

Transcripts and grades affect the employment, career and other life prospects of graduates. They also facilitate admission to licensure or membership of professional registration bodies and learned societies. Knight (2006) referred to these external uses of grades as feedout, because they provide achievement-based information for the use of third parties. Although formal documentation has lifelong significance for some purposes (such as proof of a qualification), it has reducing worth for others (such as career and employment decisions).

### *Research*

Grades are used as the criterion variable in research into various aspects of higher education, including the effectiveness of different approaches to teaching, learning

and assessment. Used in this way, grades are important research commodities. However, researchers regularly express frustration in using grades as the criterion measure. This is because grades typically lack any form of standardisation. Their merit is indeterminate, and the credibility and generalisability of research findings is thereby constrained.

## Four threats to grade integrity

Of the four threats dealt with here, the first two, random error and bias, receive only brief treatment. For over a century, they have been the subject of intensive research, much of it statistical, into the reliability and validity of measurement in the social sciences, and each has a substantial literature. Although statistical investigations may not be much help in identifying cause, they are able to indicate the presence and likely extent of error or bias. The second two threats, contamination of the object and confusion in the grading principle, are crucial issues specifically in relation to grade integrity.

### *Random error*

To some degree, random error always plays a role when measuring the characteristics of objects or phenomena. Pure measurements of perfect accuracy are impossible to produce consistently, something that has long been appreciated in both the physical and non-physical sciences. In the latter, random error is a particular problem when subjective judgments cannot be corroborated by independent means. Sometimes referred to as noise, random error comes about through chance effects that are unpredictable, uncorrelated and devoid of patterns. Error may have its origins in the instrumentation used, in human observers, or in external physical or other conditions that prevail at the time of measurement. In the context of educational assessment, these sources may include such student variables as physical or emotional stress and tiredness, misunderstanding of directions and guessing (especially on objective tests). Teacher sources include poor sampling of course content, ambiguous task specifications, fatigue and marking against tight deadlines. Institutional variables include external distractions or failures in administrative or service systems.

### *Bias*

Grading bias is systematic, and often manifests itself in some form of partiality which advantages certain categories of students and penalises others on grounds other than achievement. Bias may result from assessors' particular inclinations or points of view that dispose them towards or against certain students who may: display certain traits or attitudes (such as cooperativeness and willingness to contribute to class discussions); belong to particular ethnic, religious, racial, cultural or socio-economic groups; or have reputations of previous achievements ('this is an A+ student'). Some of these biases may be addressed by anonymising student responses, but this is not always feasible when works are produced under open conditions. Bias may also result from assessment tasks that are not equally accessible to learners from different cultures, or are couched in language forms not used in the course. Some grading biases apply across the board. An example is the practice of treating all students similarly but

for ulterior motives, such as giving generous grades to mask poor teaching, or to encourage enrolments in future course offerings.

***Contamination of the object to be graded***

The object includes everything that is directly or indirectly credited towards the course grade. Acceptance of the premise that grades should represent levels of academic achievement implies that only elements that qualify as academic achievement should also qualify for inclusion in the object. Non-achievement elements contaminate the object, and damage the integrity of the grade. Common non-achievement inclusions are credits and penalties for engagement in some activity, or compliance with some requirement. Incorporating ineligible components is now so thoroughly part of the assessment culture in many institutions that it is perceived by both teachers and students as normal and unproblematic. In some cases these practices have been actively advocated and written into institutional policies.

The converse of contamination is fidelity, a term used in a variety of contexts that have similarities to grading. High fidelity requires that all influences from non-achievement sources, including implicit and explicit inter-student comparisons, be assiduously excluded. Were they to be included, no subsequent processes or operations could correct or compensate for them. Furthermore, when course credits are awarded for routine student actions or behaviours that are not achievements, they establish a raised floor upon which appraisals of genuine achievement are placed, thus artificially boosting grades. Conversely, non-achievement penalties depress grades. Fidelity in the object rarely if ever features in the literature on assessment and grading, but is a precondition for grade integrity. The test for fidelity requires taking a literal interpretation of academic achievement, and testing potential contributors against this interpretation. It is, therefore, a matter of definition and classification, not of measurement, and can generally be determined in straightforward ways.

To demonstrate how an analysis might proceed, consider the case of participation. In almost all ordinary discourse, participation and achievement are distinct concepts, with little in common. On the surface of things, participation would therefore not qualify. However, the various arguments for including participation also need to be scrutinised to make sure no aspect has been overlooked. To go through all the explicit and implicit arguments would take more space than can be allocated here, but the arguments generally fall into three broad classes, with some overlap among them: the instrumental, the ethical, and the compensatory.

Instrumental arguments are those to do with how participation increases the probability of learning for individual students, through both its motivational and its instructional impact. Instrumental arguments can also apply to cohorts as whole entities. High levels of student participation can help create a rich, interactive peer-to-peer learning environment, which enhances learning prospects for all learners. The focus is on establishing the required critical mass of students. The fundamental problem with instrumental arguments is that they use credits and penalties as leverage mechanisms for shaping student behaviours, the latter being the means to various learning ends. The result is conflation of the end itself (achievement) with the means to that end (participation). Ethical arguments take the form that those students who are conscientious participators and put in a lot of hard work should have that commitment formally recognised in their grades. The compensatory argument is that weaker or disadvantaged students, if they are rewarded for participation, have greater opportunity to

access the higher grades. Acceptance of the compensatory argument amounts to an endorsement of participation as a substitute for achievement.

Observe that none of these classes of arguments confronts the issue of whether participation, as an activity, is part of the concept which carries the label achievement. There are nevertheless special circumstances in which participation can constitute achievement. Some people suffering from severe social anxiety disorders find it impossible to interact with others in almost any way at all. Treatment may enable them to participate in interactions with other people, and doing so would then be a significant achievement. But this special case is a far cry from teaching and learning in higher education.

Participation is but one of many common inclusions in the object for grading. The analyses both of the essential nature of participation and of the typical rationale for its inclusion serve to illustrate the processes by which its status can be determined. In the general case, whenever non-achievements are counted in, the grade represents a mix of compliances, activities engaged in and genuine achievement, or what Brookhart (1991) called a 'hodgepodge' of achievement, effort and attitude. This diminishes the merit of grades and makes it impossible to unscramble them later. Obviously, the term academic achievement no longer then applies, and it is difficult to come up with a satisfactory alternative collective term. Despite the practical difficulties of insisting on fidelity in the face of established assessment culture, the process of assuring fidelity is fundamentally simple and does not require extended analysis for each contender. It involves combing through all the elements that have customarily contributed to course grades, and resolving to exclude those that do not match the literal interpretation of achievement.

### *Inappropriate grading principle*

By a grading principle is meant a theoretically coherent, explicitly articulated set of ideas that forms a bridge between fundamental educational values (such as fairness and authenticity) and the techniques or methods that can be legitimately applied to raw assessment evidence to produce grades. The principle should be general enough to admit new types of assessment evidence or processes, but specific enough to allow for testing against the principle itself, and to permit adjudication among competing approaches to implementation. On the assumption that fidelity in the object has already been assured, grade integrity requires a decision framework that produces grades which can withstand critical scrutiny from a variety of philosophical, technical and practical standpoints, and as a result be safely accepted as trustworthy.

Fully developed grading principles along these lines are rarely found in the assessment documentation of higher education institutions. Similarly, books on assessment and grading typically outline various grading strategies, possibly with some discussion about their respective pros and cons, but are notably non-committal about the need for a coherent guiding principle (see, for example, Davis 1993; Forsyth 2003; McKeachie and Svinicki 2006). Grading practices nevertheless reflect institutional or departmental policies, conventions and traditions. In this section, some common grading approaches are briefly described and evaluated in relation to how well they could serve as grading principles. They have been grouped under five headings, the first of which (grading policies) has considerable currency. The others are headings of convenience. Some aspects of the various approaches are mutually

incompatible in principle, but this is often glossed over in developing hybrids for implementation.

*Grading policies*

These are essentially tables of equivalences that allow translation from one set of symbols to another. Typical grading policies associate alphabetic grades (A, B, C, …) with numerical values (4.0, 3.0, 2.0, …), and short verbal descriptors (excellent, very good, average, …). The use of plus and minus grades simply expands the tables. In some cases (discussed separately below), percentages or aggregate score ranges (85–100, 75–84, 60–74, …) are also included, as are fixed proportions of students (10%, 20%, 35%, …). Whereas alphabetic grades are purely nominal, the numerical values are intended to possess enough of the properties of a measurement scale to justify mathematical operations for calculating GPAs and carrying out quantitative research.

Although common in higher education, grading policies make no claim to be grading principles, but there is nevertheless a point of contact. The short descriptors are intended to provide broad meanings for the alphabetic grades, and thereby offer guidance for assigning them. However, their utility for this purpose is more limited than is commonly supposed. Short descriptors are open to individual interpretation and, depending on how literally the terms are taken, can be drawn from different registers, even within the same policy. Specifically, achievement may be judged: relative to other students (outstanding, excellent, above average); according to absolute but undefined levels (fair, very good, proficient, competent); or sufficient for gaining course credit (satisfactory, adequate, passing). In addition, the boundaries between adjacent descriptors are left open.

*Grading by aggregate score ranges*

Grades are often 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, is probably the most extensively adopted in higher education, and is commonly supported by institutional spreadsheets and grade books which simplify data entry and management. The technique is sometimes called absolute grading, presumably 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, giving the impression of 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 are cause for concern. This follows from a basic property of measurement in education: the aggregates are not composed of standardised points or units. The most obvious source of variability lies in how assessors allocate points when they make judgments subjectively, specifically whether they are generally liberal, generally parsimonious, or drift one way or the other during the scoring process.

Another source is that a given score increment does not necessarily represent the same achievement increment at all positions on the scale. This nonlinearity occurs for a variety of reasons, such as the guessing factor on objective tests (especially at the lower end of the scale). In addition, aggregates are usually made up of scores derived from all summative tests and tasks in the course, and the equivalence of score units

derived from different instruments is left unexamined. There are probably as many underlying scales and units as there are assessment instruments. Measurement in the physical sciences is quite different from this. Properties such as length and mass are measured in basic units that are standardised and have identical value across their respective scales. Inches are inches, and kilograms are kilograms.

*Grading by proportions*

Applied at the class or cohort level, aggregate scores are first arranged in order. This list (or the corresponding raw frequency distribution if enrolments are large) is then partitioned into bands that contain predetermined proportions of the cohort, and grades are assigned 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 idea being that this makes grades comparable across courses. Although often called grading on the curve, the shape of the frequency distribution is irrelevant; it could be symmetrical or skewed, bell-shaped or rectangular. What matters is that the proportions are controlled. This method is simple to apply, and produces results with a minimum of fuss except where course numbers are small or different cohorts are demonstrably unequal in overall ability or diversity, in which cases variations may be permitted. Although grading by score bands and grading by proportions both make use of aggregate scores, they partition a given set of aggregates according to different rules, and so generally do not produce identical results.

Regardless of the rule used to assign grades, provisional grade allocations made at the course level may be reviewed at a higher organisational level, such as a departmental board or panel. Grade distributions that are in line with the specified proportions are ratified; others are either negotiated with course convenors or summarily modified by the panel until they conform. Grading by proportions thus assumes an override status for the review of grades. The thinking is that, without surveillance of the final course grade distributions, the proportions of high grades could (and probably would) migrate upwards, thereby devaluing those grades. The Bologna Process (European Commission 2005) mandates grading by proportions as its grading rule, so that achievement records originally expressed on different scales can be rendered onto a common scale which will have currency across institutional and national boundaries.

Although rarely stated explicitly, the rationale behind grading by proportions is the classic market approach to regulating value when there are no stable, independent reference points. Limiting the supply of a desirable commodity in the face of steady demand generally maintains that commodity's market value. Although the combination of relative scarcity and market forces is commonly assumed to maintain real value, signifying merit is inherently different and it is merit, not scarcity, which is the key to grade integrity. By allocating grades through ranking student performances, works or aggregate scores, each grade comes to represent relative position in the cohort, not necessarily the actual level of achievement.

To rationalise grading by proportions by saying that achievement naturally distributes itself more or less according to a set pattern such as the bell curve is to argue that other factors make no net difference to the levels of achievement reached. Holding grade proportions constant makes the award of grades structurally blind to a variety of factors that affect achievement and its assessment: admission policies and student entry levels (and therefore the academic abilities of particular cohorts); the demo-

graphic profiles of cohorts; student–teacher ratios; academics' qualifications; resources for teaching; the availability and nature of support services; the quality of teaching; and the quality of assessment programs or tasks. Grading by proportions is robust essentially because it is fully self-adjusting. With each new cohort, the grading parameters are reset, and so the meaning of the grades is also reset. With no external anchorage, grading by proportions is hardly an option for achieving grade integrity.

*Grading against referents expressed as text*

Among the approaches developed to avoid the problems of grading by proportions are those that grade according to whether, or how well, students have demonstrated, satisfied or reached any of the following: course (or taxonomy-based) objectives (an approach that dates from Travers 1950); intended student learning outcomes (Biggs 1999); and grading criteria and explicit statements of 'standards', especially those expressed in scoring grids, matrices or rubrics (Freeman and Lewis 1998; Huba and Freed 2000; Walvoord and Anderson 1998). Despite significant differences, these are grouped together because they all make use of written statements in one form or another, and that is the point of interest. The rationale for institutional or departmental advocacy of grading against text-based referents rests on two assumptions. The first is that the 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 assumption is that such statements can communicate assessment expectations to students at the beginning of a course, and so emphasise that grading is not competitive but is against meaningful, objective, external referents that are accessible to assessors and students alike.

On the surface, these methods may appear to constitute a grading principle. There are, however, two significant problems. First, the content, design, construction, interpretation and application of the text is typically devolved either to individual teachers or to teaching teams, and often regarded as partly constitutive of the courses they teach. To the extent that grading decisions are made wholly within the parameters of each course, across-course comparability cannot be addressed. The second problem is the assumption that declarative or propositional knowledge, the kind that can be expressed in written statements and words, is sufficient to the task of putting the grade reference points into an enduring and workable form. This issue is returned to below.

*Miscellaneous heuristics*

Within the discretionary space available to individual markers for assigning grades, a number of other practices are sometimes employed as a means to an end. They are mostly unresearched, but are referred to in some books on the assessment of student learning, on some departmental websites, and in informal communications among academics. As heuristics, they have little if any theoretical substance. Their appeal is that they can help in arriving at grade distributions that satisfy institutional requirements. The four examples mentioned here all run counter to grade integrity. The first rule of thumb is to look for fortuitous breaks in the distribution of aggregate scores, as inspected from the top downwards (Cohen 2000; Davis 1993; McMillan 2007). A markedly lower frequency of aggregate scores in the vicinity of the normal cut-off for

the top grade suggests that resetting the boundary to that position would minimise the number of students who could appeal against the grade awarded, or who would require just an extra point or two to get over the line. The same process is then repeated for successively lower grade boundaries.

The second heuristic is to construct and inspect a trial distribution of grades. If the entire set appears too high or too low based on some presumed knowledge of the abilities and achievements of students enrolled in the course, the whole set is rescaled to a different mean and standard deviation so that, when the regular institutional cut-offs are applied, an improved grade allocation can be achieved. An alternative approach is to see whether particular students would be awarded the correct grade they are somehow known to deserve. If not, individual grade boundaries are adjusted to suit.

The third approach is to use contract grading (McKeachie and Svinicki 2006; Wright 1994). The instructor and the students develop and agree on a written contract which sets out what the student has to complete to a certain 'standard' to be awarded a given grade. The 'standard' is specified in the contract by way of relevant quality criteria. It is not the mere completion of activities or tasks that earns the grade; the quality of how well they are completed matters as well.

The final heuristic is the practice of setting grade cut-offs for a course by using the average of that course's previous cut-offs, or alternatively to set cut-offs for a department by using the average of previous grade cut-offs in all courses in the department. Both are driven by a desire for stability through pattern maintenance. Effectively, the techniques use course or departmental norms instead of cohort data, and clearly do not address the merit of the grades.

**Standards referencing as the grading principle**

In grading approaches that make use of cut-offs, proportions or textual referents, teachers or markers make the grading decisions. Where they exist, panels or boards of review scrutinise grade distributions and take steps to adjust them as necessary. The degree of autonomy delegated to academics depends on the institution, and reflects the institution's philosophical position on academic freedom. At one end of the spectrum, academics effectively have more or less sovereign right to decide grades. Each grade is regarded essentially as a protected one-way communication from professor to student (See *Parate v. Isibor* 1989). Institutions may exhort or advise teachers about preferred grading methods, but, in the end, the approach taken is up to individual teachers or course teams, and the resultant grades are not open to challenge. At the other end of the spectrum, grading must follow a set procedural policy, and the broad patterns of the grading decisions must be satisfactory. Regardless of whether ultimate responsibility rests with individual academics or with review panels, there is little that is geared specifically towards making choices based on an underlying philosophical position that addresses either specific issues, such as achieving comparability of grades across courses, or the general issue of assuring grade integrity. Logically, progress on the comparability and grade integrity fronts is impossible in institutional contexts characterised by high levels of autonomy for academics. The grading principle outlined in the rest of this section assumes that this condition does not hold.

The development of the theme so far sets the boundaries for a grading principle that is consistent with the concept of grade integrity. Specifically, the grading principle should aim to:

(1) facilitate the appraisal of student works strictly according to their quality and, if necessary, the grading of a single work outside the comparative framework of other student works;
(2) facilitate high levels of comparability within courses across different assessors;
(3) produce grades that are comparable in the various courses for each academic program, across programs, and over time;
(4) be accessible and meaningful to students so they can understand the bases on which judgments are made about the quality of their work, and thereby improve their own ability to make refined, grounded judgments about quality; and
(5) generally provide both context and stimulus for maintaining and defending academic standards.

It is imperative that the grading principle focuses on the quality of student achievement evaluated against fixed external anchor points, which are referred to here as academic achievement standards. This is the core concept in the approach referred to as standards referenced assessment and grading as originally set out in Sadler (1987). It is based on the definition of a standard as a 'definite level of excellence or attainment, or a definite degree of any quality viewed as a prescribed object of endeavour or as the recognised measure of what is adequate for some purpose, so established by authority, custom, or consensus' (p. 194). Wherever the term standard is used in this article with a meaning that does not conform to this definition, the other uses are denoted by quote marks. A criterion was defined as a 'distinguishing property or characteristic of any thing, by which its quality can be judged or estimated, or by which a decision or classification may be made' (p. 194).

Each of these definitions corresponds to at least one of a considerable variety of meanings to be found in comprehensive dictionaries, but substantial intersection exists between the two sets of dictionary meanings. In higher education practice, the terms criteria and standards are often used loosely, or even interchangeably. For the purpose of making progress specifically on developing the concept of standards referenced assessment and grading, criteria and standards need to be distinguished, because they play different roles. This has required taking one legitimate meaning for each from their respective dictionary sets and then using them consistently. In short, a standard is a definite level, and a criterion is a property or characteristic.

The definition of a standard incorporates three important elements. First, achievement standards are not out there somewhere, as if they were natural phenomena waiting to be discovered. They have no independent existence of their own, but have to be decided upon, essentially subjectively, after due deliberation and taking all relevant factors into account. Thereafter, they become stipulative. Second, a standard is a preset level that remains fixed under use, normally over a sustained time period. Third, standards are shared understandings that can be accepted as comparable but are, of necessity, manifested differently for different assessment tasks and in different courses and disciplines. They are the property of the academy as a collective, and are not determined or held privately by individual teachers or course teams. Academics have a professional responsibility for participating in the development of standards, for keeping abreast of refinements that occur as experience with them grows, and for knowing how to apply them. Similar types of standards exist within various professions, organisations, societies and guilds.

Standards referencing is also similar in certain respects to the ways in which a wide range of standards are set and applied in modern societies. Such standards cover medical devices, information security, atmospheric emissions, industrial, food and consumer safety, and a host of other aspects of contemporary life. They are devised and administered through such national organisations as the American National Standards Institute (ANSI), the British Standards Institution (BSI) and Standards Australia (SA), and globally through the International Organisation for Standardisation (ISO). At some risk of oversimplification, the standards with which these agencies are concerned fall into three categories. The first covers standards for processes rather than products. The second covers manufactured products that must conform to precise specifications, within defined tolerances, so that items from different sources will be fully compatible or even completely interchangeable in use. The third category covers entities that must conform to specified minimum performance characteristics, but with the actual design left open. Academic achievement standards are closest to this third category.

In the context of higher education, academic standards are fixed levels of quality that are negotiated and recognised as adequate for representing student academic achievement by competent, mutually calibrated discipline and professional peers. These standards ideally function as constant points of reference, but this does not imply that standards must remain fixed indefinitely, even for the same course. They may have to be reset periodically as a result of shifts in curriculum, technology or the underlying discipline or profession. However, the resetting should be done with due deliberation and the same attention to old–new comparability as was given to both setting the original standards, and to achieving comparability across cognate fields. Otherwise, the grades cannot hold their intrinsic value. The standards exist in advance of the students' beginning work in response to assessment tasks. As students are inducted into their nature and use, the standards become accessible to them to guide their responses to assessment tasks. The award of a grade is performed by comparing the quality of a student work directly with the relevant quality-oriented standards, and classifying it accordingly. The standards thus provide an explicit non-volatile decision system, which is why they allow for an individual work to be graded in isolation. Grades can also be interpreted later by reference back to the standards system.

Despite broad similarities between the standards referenced grading principle and standards systems for industrial, commercial and consumer purposes, fundamental differences exist. The most substantial is that achievement standards are, in essence, abstract concepts, whereas the other standards can usually be codified and expressed in objective, concrete form that leaves little leeway for subjective interpretation. Fundamental to academic achievement standards are the qualitative judgments made by markers. These cannot be reduced to set procedures that can be applied by non-experts (Sadler 1989, 2009a). It is well established that experts in a variety of fields can recognise quality when they see it – even when they are unable to define or explain it formally in words. For a range of related epistemological reasons, it is often impossible to express quality-based achievement standards in purely propositional or declarative form. They cannot be written down as detailed verbal descriptions, categories or lists which can then be used by students and assessors alike.

Although explanations, descriptions and their interpretations are critical tools for extending the circle of knowers to other assessors and to students, they are not enough in themselves. The words need to refer to a concrete reality. The main reason for this is that, in general, verbal descriptions do not have unique meanings or fixed,

context-free interpretations. This is a natural feature of language, and a basic source of its power and versatility. When contexts change even slightly, interpretations do as well. Meanings often migrate so gradually that subtle changes are not noticed from one point on the continuum to another. However, a quantum change in meaning may be recognised by scrutinising two interpretations separated by a considerable time period.

To crystallise and convey achievement standards requires a combination of four elements. The first two consist of exemplars, and, for each exemplar, a summary judgment of its quality together with an explanatory description. This description needs to invoke whatever criteria are relevant to that judgment. The exemplars are concrete cases that serve two purposes. They exemplify particular levels or standards, and they contextualise and anchor the corresponding explanations of the judgment. There are technical reasons for expressing the relations between these two elements in this order. In particular, the starting point is an exemplar of a standard, not a verbal description of a standard. In determining the quality of a complex student work, recognition – not definition – is the primary act. The purpose of a description is to account for a particular judgment. It is based on and therefore subsequent to the primary act. In this way, each description is constructed both to portray the properties of the exemplar and to explain the judgment about its quality, at least, as well as words can. The words form the necessary link between an immediate concrete referent and the abstract concept of quality.

The third element consists of focused discourse about the exemplars and their corresponding judgments and descriptions. This is necessary for establishing a common vocabulary appropriate to judgment and evaluation, so as to enable communication among academics with minimum ambiguity. Conversation represents language in use, not language in the abstract. Unless conversation occurs, individuals may attach their own meanings to terms, without being aware that these meanings may not necessarily be those held by colleagues. Ultimately, this would lead to miscommunication. The remaining element is tacit knowing (Polanyi 1962), which consists of subtle and often elusive aspects that people develop and share primarily through engaging in common experiences. By definition, tacit knowing is difficult or impossible to articulate adequately. Connoisseurship typically relies heavily on it.

Of these four elements (exemplars, explanations, conversation and tacit knowing), none is expendable. For particular genres or forms of student works, standards can thus be drawn out of the abstract domain and become known, shared and understood among assessors. All this is a long way from a putative 'standard' that can be conveyed through an aggregate score cut-off or a numerical representation of achievement as some form of measurement. It gets to the core distinction between standards as they apply in standards referenced grading and the 'standards' that are expressed through rubrics, tables of objectives, catalogues of competencies, sets of intended student learning outcomes, lists of skills to be mastered, gained or demonstrated, and various forms of what have loosely been called criteria based and 'standards' based grading (Sadler 2005).

## Design of assessment programs

Assessments of achievement are invariably inferential, so assessment processes need to provide a sound evidentiary basis from which reliable inferences can be drawn. Many, but certainly not all, of the methods discussed in assessment handbooks and

articles are compatible with, and important to, the concept of grade integrity. These include consistency between broad course aims on the one hand and what is taught and assessed on the other; adequacy of sampling across basic substantive content, and across cognitive or practical operations on that content; reliability and validity in scoring and grading; and efficiency in the collection of achievement data. Special attention is required, however, if standards referencing is to achieve its potential for student learning. In particular, students need to be inducted into the meaning and use of achievement standards in ways that mirror how standards can be shared among academics and how grading decisions are made in practice. This section touches on two of a number of aspects that are critical to the success of this induction process.

Bringing students into a knowledge of standards requires considerably more than sending them one-way messages through rubrics, written feedback or other forms of telling. It requires use of the same tools as those employed for setting, conveying and sharing standards among teachers: exemplars, explanations, conversation and tacit knowing. Again, none of these is expendable. As students grow in explicit and tacit knowledge of essentially the same type as that held by the teacher, they come to appropriate the standards for themselves and become better able to manage their strategic metacognitive behaviour. This growth can be nurtured by providing learners with evaluative experience in domains that are the same as, or substantially similar to, those in which they are required to produce their own works. The overall aim is to induct students into the guild of knowers, which is simply the group of people who share sufficient tacit knowledge for them to be able to recognise, judge and, to a considerable extent, explain quality when they see it. Students thereby become better able to monitor and control the quality of their own productions while these are still under development (Sadler 1989; O'Donovan, Price and Rust 2008).

With due allowance for the inevitable disparity in experience between teachers and learners, the implicit obligation is to educate students not only in the content and skills of the course itself, but also in the knowledge of what counts as quality in the corresponding field. This obligation extends to all students, not just to those who already know how to weave an informed path through learning and assessment. Specifically, students must become not only involved in making judgments about works or performances of the same genre or type that they are producing, but also engaged in informed discourse about those judgments. This needs to be deliberately designed into an educative environment, for which the stakes for learning are high but no points are awarded and banked towards summative course grades. Such an environment requires radical rethinking about how formative assessment is conceptualised and how feedback is given.

To the extent that explicit provision can be made for students to acquire evaluative expertise through becoming familiar with the academic standards that are relevant to their courses, the artificial ceiling imposed by learners' ignorance of quality is removed (Sadler 1983). A further side benefit is that when students come to a deep understanding of the standards, in an important sense it is those standards, rather than assessors, which pass judgement on the quality of their works. This reduces the pressure for assessors to award grades on grounds other than the work's quality, and to provide extensive feedback which is labour intensive to create. Incorporating the design of these aspects into course teaching and assessment can act as a positive vehicle for achieving the high-order objectives in a course (Sadler 2009b).

**Grade inflation**

In the long history of grading in higher education, specific concerns about grade inflation have surfaced regularly. Grade inflation has been defined as 'an upward shift in the grade point average … of students over an extended period of time without a corresponding increase in student achievement' (Rosovsky and Hartley 2000). Simply put, grade inflation occurs when high grades are awarded for progressively lower and lower achievements. Grades then lose a great deal of their meaning and usefulness. Another effect is that the portion of the grade scale actually used becomes restricted to the upper end.

That grade inflation is a concern at all reflects a deep-seated belief that grade integrity matters. The existence of an ordered set of symbols to represent different levels of academic achievement implies that a high level of correspondence between the two sets is desirable, so far as this is capable of being achieved. Anything less than this undermines the very point of grading. Obviously, it is important to know whether grade inflation is a real phenomenon or part of some general romanticisation of the 'standards' of the past. Research into grade inflation generally analyses frequency distributions of grades awarded at different points in time, in association with a number of covariates which are potential explanatory variables. However, the masses of data that do exist on grades cannot get at the core of the matter, because the problem as formulated and researched is flawed.

Although typical studies may provide indicative evidence for the existence of the phenomenon, they lack a crucial feature, namely a fix on the raw levels of academic achievement which are independent of the grades awarded. For many courses, the concrete first-order evidence of achievement consists in the actual works students produce, and these are either discarded or returned to students soon after the grades are awarded. Obviously, they cannot be reconstructed and interrogated retrospectively, so revisiting the original site of the grading decisions is not possible. Without comprehensive archiving of student works, the responsibility for achieving commensurability between achievement and grades rests on getting the initial grading correct. Only when there is a stable, objective way of identifying and recording genuine achievement is it possible to track real changes over time, and only in the presence of a criterion variable with these properties is it technically possible to produce definitive evidence about the nature and extent of grade inflation.

The issue is not whether grade distributions are creeping inexorably upwards, but whether the grades awarded remain commensurate with actual levels of achievement. The proper place to address grade inflation is at the site where grades are assigned, that is at source. If that could be achieved, with standards remaining external to, and independent of, particular cohorts, the proportions of the different grades would no longer be relevant. In addition, competitive game playing by students would produce no relative advantage or disadvantage. Bids by course convenors to award high grades to all or most of the students in a purportedly elite cohort could be tested to see whether such claims are sustainable.

Institutions that decide on a policy shift away from relatively loose (or effectively no) regulation of grade apportionment and towards strict rationing of grades expose themselves to two conflicting reactions from teachers and students. Strict apportionment certainly raises the relative value of higher grades. In addition, it frees up the grade scale and provides opportunity for markers to use a greater segment of it to represent achievement than is available with a truncated scale. In theory, strict rationing

could, therefore, lead to improved merit overall. On the other hand, it would mean a decrease in merit for any student who legitimately deserves a high grade, but is simply outcompeted by other students. The incidence of events of the latter kind may be low, but this is no consolation for the students concerned when the stakes are high. In any case, such events are certain to occur from time to time, and this structural unfairness makes it both a matter of consternation for many academics and students, and ethically dubious as a policy option. On the other hand, an institution that adopts the grade integrity route as a valid alternative to a simplistic grade deflation policy takes up an approach that is overtly positive towards student learning.

That there are few signs of appropriate data being collected may be indicative of an unwillingness to confront the issue at its roots, a lack of appreciation of the potential alternatives, or resignation to the futility of coming up with any solution at all. Insisting on fidelity in the object accompanied by informed judgments in setting, conveying and applying academic standards offers a way out of the grade inflation dilemma. If these conditions could be met, research into grade inflation could be added to the list of assessment problems that are amenable to research.

## Conclusion

Grades possess integrity when they are true representations of student academic achievement in courses. Assessment fulfils many roles in higher education, some of which are summative and some formative. Summative grades provide official certification of academic achievement for use inside and outside the awarding institution. Despite certain misgivings about the worth of grades as symbolic representations of achievement, they are nevertheless put to serious use. Part of this is no doubt due to their compactness and the lack of workable alternatives, and users seem to have reasonable confidence in them as aids to administrative decision making. However, the ways in which grades are commonly being awarded are open to criticism. Two significant but mostly ignored issues are raised in this article. The first consists of course-related student activities and behaviours that do not fall within the formal meaning of achievement but are nevertheless routinely incorporated into grades. This practice produces grades that are contaminated, and the resulting loss of fidelity reduces the intrinsic merit of the grades.

The second issue is the common use of grading practices that are not up to the task of representing achievement in anything like an absolute sense. They typically employ either grade cut-offs for use with uncalibrated scores, or each student's relative achievement standing within the cohort. Among other things, these practices make unambiguous interpretation of an isolated grade difficult or impossible. The alternative proposed is to employ standards referencing as the grading principle. This would be an option for higher education institutions that exercise a degree of policy authority to govern the manner in which grades are assigned. An institution that aspires to achieve high levels of integrity in their course grades is likely to find that strategic policy decisions have to be made and defended at the institutional level. If these were left to departments or individual academics to decide and implement, a wide variety of sub-optimal practices could again become so deeply embedded and actively advocated on pragmatic grounds that they become fully normalised and perceived as unproblematic.

The approach to grade integrity set out in this article can have a strong formative impact for students, because it opens up the nature of quality to students so that they

can become conversant with it. This allows them not only to develop their responses to assessment tasks more intelligently during a course, but also to function more professionally after graduation. The pursuit of grade integrity as an institutional priority has implications that are far reaching for teachers, learners and third-party users of grades. This article is intended to provide basic directions for such a pursuit.

## References


Aeschylus. 458 BC. *Agamemnon.* In Aeschylus, ed. and trans. H.W. Smyth and H. Lloyd-Jones. *Aeschylus II: Agamemnon, Libation-Bearers, Eumenides, Fragments,* 67. Cambridge, MA: Harvard University Press, 1988.

Biggs, J. 1999. What the student does: Teaching for enhanced learning. *Higher Education Research and Development* 18: 57–75.

Brookhart, S.M. 1991. Grading practices and validity. *Educational Measurement: Issues and Practice* 10, no. 1: 35–36.

Cohen, W.D. 2000. The Grade Point Average (GPA): An exercise in academic absurdity. *National Teaching & Learning Forum* 9, no. 5: 1–4.

Cureton, L.W. 1971. A history of grading practices. *Measurement in Education* 2, no. 4: 1–8.

Davis, B.G. 1993. *Tools for teaching.* San Francisco: Jossey-Bass.

European Commission. 2005. *ECTS User's Guide.* Brussels: Directorate-General for Education and Culture.

Forsyth, D.R. 2003. *The professor's guide to teaching: Psychological principles and practices.* 1st ed. Washington, DC: American Psychological Association.

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

Huba, M.E., and J.E. Freed. 2000. *Learner-centered assessment on college campuses: Shifting the focus from teaching to learning.* Needham Heights, MA: Allyn & Bacon.

Knight, P. 2006. The local practices of assessment. *Assessment & Evaluation in Higher Education* 31: 435–52.

Linderman, K., R.G. Schroeder, S. Zaheer, and A.S. Choo. 2003. Six sigma: A goal-theoretic perspective. *Journal of Operations Management* 21: 193–203.

Lincoln, Y.S., and E.G. Guba. 1980. The distinction between merit and worth in evaluation. *Educational Evaluation and Policy Analysis* 2, no. 4: 61–71.

Locke, E.A., K.N. Shaw, L.M. Saari, and G.P. Latham. 1981. Goal setting and task performance: 1969–1980. *Psychological Bulletin* 90: 125–52.

McKeachie, W.J., and M.D. Svinicki. 2006. *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers.* 12th ed. Boston, MA: Houghton Mifflin.

McMillan, J.H. 2007. *Classroom assessment: Principles and practice for effective instruction.* 4th ed. Boston, MA: Pearson/Allyn & Bacon.

Milton, O., H.R. Pollio, and J.A. Eison, 1986. *Making sense of college grades.* San Francisco: Jossey-Bass.

O'Donovan, B., M. Price, and C. Rust. 2008. Developing student understanding of assessment standards: A nested hierarchy of approaches. *Teaching in Higher Education* 13: 205–17.

*Parate v. Isibor*, 868 F.2d 821 6th Cir. 1989. (US Federal District Court).

Polanyi, M. 1962. *Personal knowledge.* London: Routledge & Kegan Paul.

Rosovsky, H., and M. Hartley. 2002. *Evaluation and the academy: Are we doing the right thing? Grade inflation and letters of recommendation.* Cambridge, MA: American Academy of Arts & Sciences.

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

———. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13: 191–209.

———. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18: 119–44.

———. 2005. Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education* 30: 175–94.

———. 2009a. Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education* 34. (Online DOI: 10.1080/02602930801956059)

———. 2009b. Transforming holistic assessment and grading into a vehicle for complex learning. In *Assessment, learning and judgement in higher education,* ed. G. Joughin. Dordrecht: Springer.

Scriven, M. 1967. The methodology of evaluation. In *Perspectives of curriculum evaluation: Vol. 1.* AERA Monograph Series on Curriculum Evaluation, ed. R.W. Tyler, R.M. Gagné, and M. Scriven, 39–83. Chicago: Rand McNally.

———. 1978. Merit vs. value. *Evaluation News* 8: 20–29.

———. 1991. *Evaluation thesaurus.* 4th ed. Newbury Park, CA: Sage.

Smallwood, M.L. 1935. *An historical study of examinations and grading systems in early American universities.* Vol. 24. Harvard Studies in Education. Cambridge, MA: Harvard University Press.

Smith, B. 1993. Six-sigma design. *IEEE Spectrum* 30, no. 9: 43–47.

Travers, R.M.W. 1950. *How to make achievement tests.* New York: Odyssey Press.

Walvoord, B.E., and V.J. Anderson. 1998. *Effective grading: A tool for learning and assessment.* San Francisco: Jossey-Bass.

Wright, D.L. 1994. Grading student achievement. In *Handbook of college teaching: Theory and applications,* ed. K.W. Prichard and R.M. Sawyer, 439–449. Westport, CT: Greenwood Press.