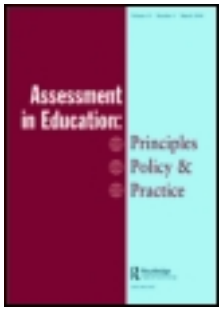

Assessment in Education: Principles, Policy & Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Assuring academic achievement standards: from moderation to calibration

D. Royce Sadler [a]

[a] Teaching and Educational Development Institute, The University of Queensland, Brisbane, Australia

Version of record first published: 21 Aug 2012

To cite this article: D. Royce Sadler (2012): Assuring academic achievement standards: from moderation to calibration, Assessment in Education: Principles, Policy & Practice, DOI:10.1080/0969594X.2012.714742

To link to this article: http://dx.doi.org/10.1080/0969594X.2012.714742

Assessment in Education: Principles, Policy & Practice
2012, 1–15, iFirst article

Assuring academic achievement standards: from moderation to calibration

D. Royce Sadler*

Teaching and Educational Development Institute, The University of Queensland, Brisbane, Australia

(*Received 22 December 2011; final version received 4 July 2012*)

The course (module) grades entered on higher education academic records (transcripts) purportedly represent substantive levels of student achievement. They are often taken at face value and accepted as comparable across courses. Research undertaken over several decades has shown that the underlying standards against which student works are appraised are poorly understood and can vary widely from assessor to assessor. At the same time, it is commonly held that academic judgements should be respected and form the basis of any quality assurance scheme. This article is about some of the conceptual foundations relevant to a particular approach to assuring academic achievement standards. The final concept discussed is that of 'calibrated' academics who are able to make grading judgements consistent with those which similarly calibrated colleagues would make, but without constant engagement in moderation. The overall aims are to achieve comparability of standards across institutions and stability of standards over time.

Keywords: moderation; teacher judgement; peer review; academic standards; higher education

Introduction

Moderation to improve marker consistency (inter-scorer reliability) is widely practised in education when two or more markers appraise complex student responses to assessment tasks. Early studies into the judgements made by different markers, when acting largely autonomously, showed that the different sets of scores were typically poorly correlated or characterised by substantially different means and variances (Starch and Elliott 1912; Hartog and Rhodes 1935). Subsequent research has produced essentially similar results, although a number of techniques are available that do lead to improved consistency, among them following a common set of guidelines closely. Moderation is intended to ensure that the mark a particular student is awarded is independent of which marker does the marking. The English verb 'to moderate' dates from about 1400, and originally meant to regulate or abate excessiveness, to smooth out extremes. Simply averaging the scores from two or three markers literally does that, but without delving below the marks to find the reasons for the differences. Other approaches take a different tack. Linn (1993) reviewed a number of moderation models, one of which involved different assessors

*Email: d.sadler@uq.edu.au

ISSN 0969-594X print/ISSN 1465-329X online

http://dx.doi.org/10.1080/0969594X.2012.714742
http://www.tandfonline.com

reaching consensus on how marks should be awarded. Consensus moderation (Linn's term was 'social moderation') provides the starting point for the analysis in this article.

At a much higher level of generality, concern has been expressed about the comparability of course grades in higher education. The development below is based on the premise that certain aspects of consensus moderation may be repurposed, transformed and supplemented for improving the comparability of course grades. These aspects are: the centrality of academic judgement; the principle of consensus; and distributed responsibility among academics for attaining the end. The goal is to produce course grades which are: based on academic achievement standards; commensurate with the respective levels of achievement they represent; comparable over time and across course boundaries; and consistent with disciplinary, professional and societal expectations of higher education graduates (Sadler 2009a). A complementary objective is to contain the associated workload within reasonable limits by replacing the concept of moderation (as an essentially consensus-seeking activity) with a more general and sustainable capability that draws on the concept of calibration.

In his classic article on grading, Urmson (1950) wrote that, 'grading is something which you cannot in a full sense do without understanding what you are doing' (147). In the spirit of that, this article is a contribution to thinking about some fundamentals. The theme concentrates on the conceptual and cognitive demands engaged in during moderation and also necessary for calibration to serve the wider agenda. Certain aspects relevant to the design of full systems for assuring academic achievement standards lie outside the scope of this article, specifically the politics and logistics of system design, standard setting procedures, and ways of expressing achievement standards in a material form which provides appropriate anchorage.

Terminology

Course refers to a unit of study as a component of a taught undergraduate or postgraduate degree program, also called a module, subject or paper. *Grade* used by itself refers to a course grade. No distinction is made between *marks*, *scores* or *points* used for coding and recording the quality of student responses to assessment tasks. In referring to worth or value, *equivalence* and *comparability* are used interchangeably.

Rationale

The first of two reasons for pursuing the assurance of academic achievement standards and comparability of course grades through the approach outlined in this article is that it is product based. This refers to how well grades in different courses correspond with the nature, breadth and depth of learning as inferred directly from an integrative and holistic evaluation of all the raw evidence of achievement (student works). The necessary qualitative judgements are made by competent persons, their brains being both the sources and the instruments for appraisal. In assessing the quality of a student's response, there is often no independent method of confirming, at the time when a judgement is made, whether the decision or conclusion is correct. Indeed, it may be meaningless to speak of correctness at all (Sadler

1989). The final court of appeal is to another qualitative judgement, or better still, consensus-based collaborative judgements based on academic standards. That is what this article is leading to. According to Brown (2010), putting and keeping academic judgements in the foreground should be 'one of the cardinal principles' (135) in developing a standards system.

The second reason is related to the first in that the alternatives to a product-based assurance approach are process based. Two common methods require obtaining student scores for all assessment tasks (regardless of format), aggregating them and applying a rule to convert aggregates to course grades. A traditional rule involves dividing the aggregate range into bands defined by cut-off scores, one band for each grade level, and then simply classifying all student aggregates. A second common rule is to fix in advance the proportions of grades that are allowable (within some tolerances) and divide the order of merit list (based on aggregates) according to those proportions. Such grades are awarded competitively in a zero-sum game. The second rule may be used as a fallback provision in the event that the first rule produces too many high grades or too many failures. Both of these rules have long been criticised as logically flawed in principle and relying on untested statistical assumptions (Oppenheim, Jahoda, and James 1967; Elton 2004). A third and more recent method is to rely on codifications, which are word-based descriptions of standards, including rubrics. Interpretations of the key terms in such statements are not fixed universals but context dependent and, for that reason, are elastic.

The sufficiency or effectiveness of these three grading approaches (cut-off scores, fixed proportions and codifications) cannot be empirically demonstrated in the absence of an independently derived variable which reflects directly the nexus between actual achievement as inferred from evidence and the course grade awarded. Such a variable is known in test theory as a 'criterion variable'. The rest of this article is directed towards conceptualising a product-based criterion variable based on professional, high-level, informed qualitative judgements that extend beyond the traditional limits of both course-based moderation processes and the use of external examiners.

The base model: single assessment task, multiple markers

Consensus moderation is commonly used for extended complex assessment responses when course enrolments are large, and when a course is taught on different campuses or in different modes. Marking extended complex responses typically involves an assessor assigning a score by making a qualitative judgement, giving partial credit as appropriate. Complex responses include term papers; essays; written assignments; field or project reports; solutions to mathematical, scientific or technological problems; seminar presentations; studio or design productions; specialised artefacts; clinical consultations; creative works; client interviews; and professional procedures or performances. Constructed responses to examination items may, if they are substantial enough, also qualify as complex works. Qualitative judgements about these works are not reducible to rules that non-experts can apply (Sadler 1989).

Consensus moderation starts with a sample of student responses drawn from the course pool. Working independently, all assessors mark all responses in the sample. For each, they record their provisional judgement and their reasons for it. Markers

then convene as a group, individually present their decisions and rationales, and deliberate them until consensus is reached. Abercrombie's (1969) research demonstrated the advantage of markers recording provisional marks and reasons prior to discussion over simply forming general impressions about individual works. Creating physical records formalises commitment to the decisions so that they can function as concrete data for reaching consensus on marking. Formalisation also has a positive influence on group dynamics, making it less likely for one assessor to dominate discussion. After discussion, assessors mark their allocated batches of responses more or less independently, with further cross checking and review of atypical cases as necessary. The quality of works and the marks awarded are linked specifically to those responses, to that assessment task, from that group of students and for that team of assessors. They are negotiated locally for the express purpose of improving consistency and fairness in scoring, and the basic procedure is typically repeated for subsequent assessment tasks. In the next two sections, the concept of quality and the representation of a level of quality by a code are analysed. These two form the main cognitive tasks involved in consensus moderation.

The concept of quality

'Quality is something I do not know how to define but I recognise it when I see it'. This statement captures the essence of an experience many people are familiar with. Furthermore, the same sentence structure can be true when the word 'quality' is replaced by any number of other words, each referring to an abstract concept. Examples are fairness, love, freedom and chauvinism. Although explicit definition may be difficult or impossible, this does not affect the legitimacy or functionality of such concepts. People acquire and use them all the time. Extensive research has been carried out on the psychology of how children appear to acquire concepts, some of which are concrete, others abstract. Significant work on this aspect of thinking was carried out by Bruner, Goodnow, and Austin (1956). Theirs was not by any means the only direction from which the general problem has been investigated but, for the purposes of this introduction to a broad field of inquiry, the discussion that follows uses their terminology.

Bruner, Goodnow, and Austin (1956) analysed various characteristics of a wide variety of simple and complex concepts and differentiated classes according to their basic structures. Included in their work were some findings pertinent to the concept of quality. For complex works, quality manifests itself on a continuum according to the different amounts or levels present. In deciding on the level of quality for a particular case, multiple criteria are usually involved. Each criterion may itself be a concept that is difficult or impossible to define, and each is usually manifest to a greater or lesser degree along its own continuum. Consequences of this phenomenon are that a group of criteria which seem in the abstract to name identifiably distinct properties may, when applied to actual judgements, turn out to overlap. As explained in Sadler (2009b), fixed sets of criteria used to formalise the process of making qualitative judgements are often found wanting, not least because the use of fixed sets assumes that all potentially salient properties are known in advance. This assumption manifests itself whenever different sets of criteria are proposed for making appraisals of the same set of objects. A common experience is that some things can be appraised as outstanding but for reasons not covered fully by a fixed set of criteria. Conversely, meticulous application of a fixed set may overvalue or

undervalue other things. These sorts of problems can be traced back to the ways in which different concepts and groups are formed and applied.

Suppose that the appraisal of a particular student response could be satisfactorily explained, after the fact, by appeal to a certain set of criteria. Another work of essentially the same quality may have an explanation which appeals to a somewhat different set of criteria. Although sets of criteria may differ from case to case, each set may be valid for the work to which it specifically refers. In such situations, Bruner, Goodnow, and Austin (1956) would say that the criteria are being used 'disjunctively', meaning that works of comparable quality may be characterised by one group of properties, or another group of properties, or a selection from both groups, or others altogether. The significance of this statement lies in the connective 'or'. A disjunctive concept allows for different things to 'qualify' as instantiations of a group or class through alternative sets of attributes or criteria. Also allowable would be judgements in which the salience of certain properties may be contingent on the levels of other properties that are also evident. Wittgenstein's ([1967]1974) celebrated example of this phenomenon was the concept of a *game*. That some judgements are made by the disjunctive use of criteria provides an explanation as to why a particular level of quality can be easier to recognise holistically from a complex set of properties than it is to define or to deduce analytically. A similar observation may be made about recognising holistically that a particular set of activities can be classified as a game.

Recognition in a specific case works by means of allowing certain of an object's properties to be perceived as salient to a judgement and incorporated into the explanation or rationale for the judgement of its quality. By definition, whatever is salient is noticed; what is noticed may be (without necessarily being) open to verbalisation. Any such verbalisation clearly has its roots in the object itself. A description of certain aspects of an object is intimately connected with the object – and may even be unique to it. A formal definition, on the other hand, applies to a class of things, and functions as the 'decider' as to whether or not an object qualifies as a member of the class. If the borders of a class (say, of objects that are of about the same quality) are established and maintained by other than a set of properties which are held in common (called the 'criterial attributes' by Bruner, Goodnow, and Austin 1956), constructing a formal definition is not possible. In contrast to disjunctive classes or concepts, a conjunctively formed class or concept is characterised by the use of a fixed set of attributes (criteria), with little or no flexibility. Rubrics provide an example of a conjunctive rule that specifies a fixed set of criteria for application to all judgements.

That the recognition of quality is a fundamental evaluative act in its own right (Dewey 1939) is a perspective widely appreciated and applied in many fields – but less so, overtly at least, in assessing student work. A judgement can, and often does, precede rational analysis. It involves 'holistic similarity recognition' (Dreyfus and Dreyfus 1986, 28), which comes about through responding or reacting to the object of interest. This capability is developed to a substantial degree by making global judgements about, and discriminations among, multiple actual cases (Dreyfus and Dreyfus 2005). To construct a justification for a judgement, assessors select from a pool of criteria those that are salient to a particular work and compose statements with the help of qualifiers, modifiers and hedge words. None of these word elements is in itself absolute. Each is open to interpretation. The one thing that has the potential to give the word elements substance and tie them together is the

specific work to which they refer – the referent. The text of the justification draws attention to certain key features of the object that contribute to its valuation while features of the object provide substance and meaning for the text of the justification. The two are therefore in a mutually reciprocal relationship.

When several assessors agree on a judgement but differ in their explanations, it may be that they have attended to different aspects of the work. Alternatively, they may have attended to the same aspects but still come up with explanations that are structured or expressed differently. That is why one assessor's 'coverage' of an explanation of a judgement may have much the same 'coverage' as another judge's explanation which seemingly invokes different criteria. The main point is to share the essence of the reasons, which can subsequently be achieved through open discussion about particular cases. Decomposing a particular judgement by constructing a rationale for it afterwards – to any desired level of detail – is important in communicating achievement levels, standards and the symbols or labels that go with them. However, exhaustive explanation may not be possible.

In judging quality holistically, proficient assessors readily see below the surface features. They typically run dual agendas simultaneously, one of which focuses on the overall quality of the work, the other on particular characteristics. They notice aspects of student responses that are worth noticing and pass over others which are ordinary or expected. Less proficient assessors lean towards following rules (Dreyfus and Dreyfus 1986). Proficiency in making consistent global judgements requires practice not only with multiple cases but also with variety. Numerous cases over the fullest range possible are crucial in building up experience in judging quality as expressed through differing configurations. It is 'through variation that aspects are differentiated within the experience of a phenomenon' (Marton and Booth 1997, 145). Sensitivity to the cues that a community of assessors regards as salient cannot, in general, be developed through personal experience of difference and variation alone, even if it is extensive. Clearly, this sets up a learning challenge for assessors. Bereiter and Scardamalia's (1993) work emphasised that, in grasping an underlying concept as shared by a community, verbalisation and discussion play a key enabling role. Without those, learning is typically slower and less certain.

Representation of degree of quality

The aim for assessors engaged in consensus moderation is to agree not only on what constitutes quality in a concrete setting but also on how a particular level of quality should be represented, typically by using a code such as a numeral or other symbol. Marking is therefore a mapping process. In the case of numerical representation, the scale is typically finely divided with 10, 20, 50 (or so) points or 100 in the case of 'percentage grading'. The numerals are generally treated and operated upon as if they were numbers, which is to assume they possess some of the properties of true measurements, namely equal sized units on a standardised interval scale. These properties are rarely, if ever, tested. One factor potentially contributing to non-linearity is that certain numerals may carry increased significance at particular parts of the scale. For instance, if barely passable work is coded as 10 on a 20-point scale, assigning a mark in the range 9–11 may be influenced by the possible consequences for a student who 'fails' that task.

Discussion

The reason for disagreements among assessors, when these occur, could be that the assessors hold different concepts of quality or, if there were agreement on that, they differ on how the symbols should be assigned. These two factors are separable in principle and could be explored empirically. The first part could be tested by asking assessors individually to make paired comparisons (Thurstone 1927) among a varied set of responses and then applying an algorithm such as that developed by Saaty (1977) to arrange assessed responses in order on an interval scale. If high agreement is reached on the positions but the marks allocated differ, the problem is due to differences in how the judgements are coded. In practice, these two operations normally flow seamlessly into one another and are not differentiated. However, in the larger context of assuring course grades, conceptual separation is necessary because their generalisations follow different directions and play distinct roles.

Conversations about actual student works, judgements made about their quality, and the grounds for those judgements identify meanings-in-use for the principal explanatory terms. Each summary judgement is tied to a concrete referent, the link being the specifically tailored rationale. Discussion provides a forum in which assessors establish a common vocabulary and set of meanings in relation to the mark to be awarded in that assessment event. To the extent that markers do not come to each grading event with a completely blank slate, they may well bring different ideas about how marks should be awarded. The moderation process is essentially a tuning exercise to reduce such differences.

To sum up, consensus moderation is carried out when multiple complex student works arise from a single assessment task. Qualitative judgements are necessary because the student responses are non-standardised. The underlying variable of interest is the 'quality' of a response, and the judgement is coded. Works judged of equivalent quality are assigned the same code. Because assessors judge differently, consensus moderation is used as a way of arriving at a shared understanding of the mapping to be applied by all assessors to improve consistency. However, being focused on student responses to a single assessment task, the consensus is localised in its scope. The main challenge ahead lies in generalising from the basic principles.

Moderation and peer review as procedural principles

Peer review is well established in higher education for evaluating research grant proposals and journal article manuscripts. Although not above criticism, it is widely and strongly defended as a significant academic quality assurance process when the objects under consideration are not standardised. In broad terms, peer review for grants and articles makes use of double-blind appraisal by reviewers, and independent and impartial chairing of the procedures by a research panel or journal editor. Labour demands are kept reasonable. Reviewers typically work independently, provide comments to the granting body or editor and make recommendations (approval, rejection or approval subject to certain conditions). Consultation may occur for problematic cases. Some aspects of peer review in the research domain have parallels with consensus moderation, and some key differences stand out. It is now shown how consensus moderation and existing peer review practices can contribute to a conceptualisation that is helpful in assuring academic achievement standards and course grades. As previously, the exploration is largely confined to conceptual and cognitive aspects.

Peer review of grades in a single course

Two concepts involved in consensus moderation are important in assuring course grades – an underlying variable of interest and codes to represent levels of performance. The underlying variable is *achievement* attained by the end of the course, as determined by inference from primary evidence, namely, student responses on all summative assessment tasks or observations of relevant behaviours in specified settings. Potential sources include artefacts or other types of physical products, or secondary records such as audio or video recordings for musical, dramatic, laboratory, clinical and similar performances. Evidence drawn from multiple sources has to be integrated in some way. The full achievement continuum is, in many contexts, partitioned into relatively few bands or segments, the relevant code or label for each being alphabetical (A, B, C, etc.), numerical (7, 6, 5, etc.) or verbal (distinction, merit and pass). Each judgement is made by looking directly at the evidence through eyes that accommodate breadth, depth and quality across task types, assimilating and balancing it all. The conceptual shift from 'quality of a single production' to 'achievement based on differentiated evidence' is substantial and, to many assessors, unfamiliar.

In everyday use, an 'achievement' or 'attainment' is a significant performance status which is valued and in many cases has not previously been reached. Achieving or attaining denotes bringing to fruition or to a successful end, especially through effort, skill, practice and perseverance (Sadler 2010). Broadly speaking, academic achievement consists of acquired knowledge and capabilities for performing advanced types of tasks independently, on demand, and consistently well in some area of specialisation. This interpretation of achievement does not admit as legitimate evidence any contribution for effort, participation and completion of practice exercises, valuable though these may be for learning. It does not take into account any non-achievement penalties or rewards (such as for late submission or improvement, respectively). Given that the object of interest is the level of achievement attained specifically by the end of the unit of study, evidence obtained during the learning period may not adequately reflect a student's final level of performance (Sadler 2010). The break points between grade bands cannot be defined with absolute precision but are, in the words of Bruner, Goodnow, and Austin (1956, 29), 'fuzzy transition zones' which require fine professional judgements for borderline cases. The actual positions of the break points are essentially arbitrary, but once settled, fix the standards to be applied.

Relying as it does on making direct linkage between the evidence and the grade, this portrayal of the grading process has made no reference to marks, rubrics, scoring systems or other guides to decision-making. It rules out any concessions or allowances for variations in student characteristics, entry qualifications, teaching conditions or resourcing levels. The latter are, of course, critically important inputs to teaching and learning, and they ordinarily do affect achievement levels, but they do not constitute achievement itself. This exclusive focus on evidence of achievement and the integrity of course grades feeds into the way in which grades are ordinarily interpreted and used inside and outside higher education institutions.

Determining grades that are commensurate with actual achievement levels requires evidence of high quality. Inadequacies in primary data make it difficult to distinguish evidence of poor achievement from poor evidence of achievement. Task design and task specifications (with no reliance on oral elaborations) are therefore

necessary inputs to the grading process, but not to grades themselves. They are not in the same category as the potentially influencing elements mentioned in the previous paragraph. The ideal would be that they are subjected to close scrutiny in their own right. Unless the basic data are known to be responses to high-quality assessment tasks, the integrity of grades is undecidable.

Assuming passage through that filter, the process starts with 'bare' student works with no information about the identity of students, no cuing about how previous judgements were reached, and no data on the spread of performance represented in samples of student works. Any such interpositions are likely to interfere with the ability of observer-markers to perceive what is actually there. As Abercrombie's (1969) research showed, observers operating with prior guidance 'tend to see what they are expected to see whether it is there or not' (99). Student works exhibit or express a certain level of achievement, so an assessor's initial idea of what constitutes achievement in a particular course, which is held as an abstraction, takes on definite form through exposure to real instances. This is consistent with two principles. The first is that the judgements should be made in as direct and absolute a way as possible. The second is that standards should be formulated in a way which allows them to be applied to particular assessment tasks at will.

Standards-referenced grading and intersubjectivity

Comprehensive dictionaries list upwards of 25 meanings for the word 'standard'. Two or more of these meanings are often used in the same discussion about educational standards, the participants being unaware of the different shades of meaning. The intention here and in the remainder of this article is to use the meaning set out in Sadler (1987) which, with a slight rewording, is as follows:

> Standard: A definite degree of academic achievement established by authority, custom, or consensus and used as a fixed reference point for reporting a student's level of attainment.

A set of graduated standards provides the framework for expressing a judgement as a course grade. Particular features of this interpretation are that: each standard has some solidarity about it; the reference levels, once set, are treated as fixed; standards are not simply 'out there' waiting to be discovered but are set through human agency; and student performance is evaluated and reported in terms of the standards. It would therefore not be correct to say 'Standards are rising' to mean that actual levels of student performance are getting progressively higher. The standards do not rise or fall – but they may be set or reset as a deliberate act, as and when necessary. (The interpretation above leaves open the method or methods by which standards are set.) Treating standards as fixed reference levels implies that they are stored in some way so they can be accessed, referred to and used in other contexts and at other times. In the language of copyright law, they require some 'material form', a term which refers to any mode of information storage that is sufficiently permanent or stable for it to be identifiable, perceived, reproduced and communicated. In passing, observe that the process of moderation does not ordinarily make assumptions about explicit, fixed standards as defined above. The main aim there is to achieve consistency among markers for a localised event. However, there will, of necessity, be some implicit 'standards' being applied.

In principle, standards should facilitate judgements which are in keeping with judgements of similar objects (specifically, student evidence of achievement) made by competent assessors whenever required. In some higher education contexts, academics have a formally protected right to grade according to their own preferences, and so set their own 'standards'. Such 'standards' are essentially private, potentially idiosyncratic and do not comply with the interpretation here, hence the use of quote marks. The possibility of assuring course grades then vanishes. Ideally, standards are properly thought through, set by consensus, adequately externalised and held as shared knowledge among academics in a discipline, field or profession.

Judgements which are integrative, holistic and made without formal decision templates or procedures are commonly labelled in a somewhat derogatory way as 'subjective' as if to suggest they are based on little more than unsubstantiated opinion or personal taste. That line of thinking does subjective judgements a grave disservice. Many professionals constantly rely on so-called subjective judgements that are not, and sometimes cannot be, verified by independent objective means such as a standard laboratory test. Subjective judgements can be soundly based, consistently trustworthy and similar to those made by comparably qualified and experienced professionals. They can also be poorly based, erratic and unreliable. Furthermore, in some circumstances quite different judgements may be equally appropriate for different purposes.

When presented with a collection of a diverse range of phenomena or objects, members operating within a guild of like-purposed professionals should in principle be able to make the same judgements within a fairly small margin of error. Such judgements would be accepted as 'true' beyond each judge's personally constructed decision space (that is, the space available only to a particular judge), provided the parameters for the shared decision space are set and accepted collegially. The meaning and significance of evidence are shared, as is what is deemed to count as evidence. In short, given the same stimuli, the people making the judgements should react or respond similarly and judge similarly. The existing term that is closest in meaning to this state of affairs is 'intersubjectivity', a term used in phenomenology, psychology, philosophy and several other fields (with appropriately nuanced meanings).

Intersubjectivity is distinct from interscorer reliability, in that not only are similar judgements made but also the grounds for those judgements are shared as well. Consistency on its own can be potentially achieved without that. Intersubjectivity is also distinct from objectivity if by objectivity is meant an unarguable fact, such as 'one water molecule contains two hydrogen atoms and one oxygen atom'. As Scriven (1972) pointed out, the quality of a judgement made by a single assessor is not automatically suspect and deserving of dismissal merely because it has been made without collaboration and without the help of instrumentation. Academics as professionals who consistently arrive at sound judgements are effectively 'calibrated' against their competent peers and also, in professional contexts, against any relevant socially constructed external norms. As with other competent professionals, they subscribe to and support the idea of deprivatised standards and know how to apply them properly (Sadler 2011).

Assuring grades across course boundaries

Commensurability of grades within a course is one thing. However, for the reasons listed in the Introduction, more is generally aspired to. The full assurance of

achievement grades requires that they also be comparable across course boundaries, and this presents a challenge to both conceptualisation and practice. The problem is similar in principle to that faced by external examiners who carry out retrospective reviews of grades and the evidence for them, and also by some professional accrediting agencies that scrutinise actual student works in their quality assurance processes. But is it necessary for comparability to be established across all courses in an institution? This section is about establishing a meaning for substantive comparability with just a few comments about review approaches.

The word 'comparable' has two distinct meanings. The first, which is written here as compare-able, means 'able to be compared, admitting of comparison with others'. This follows directly from the etymology, the emphasis being on possibility or 'able to'. The second meaning is 'equal or equivalent to', which is written here as com-prable. The compare-ability of two things is prospective, and a logical prerequisite for any determination of com-prability, which is both retrospective and descriptive. For objects that clearly belong to the same class, their compare-ability is rarely given a second thought. Performances of a set piece of music by different pianists are compare-able, even if one is a professional concert pianist and the other a student in performance piano.

For other objects that may appear different superficially, they may be compare-able with respect to some higher-order criterion. Two radically different journal articles may both, for instance, provide important contributions to knowledge, meet the requirements for good scholarly writing and be equally relevant to a particular journal's aims. Publishability would then function as the higher-order criterion, and the two articles could be then judged as com-prable. Similarly, the examination of different doctoral dissertations typically makes appeal to such higher-order criteria as: comprehensive and detailed knowledge of the field; mastery of appropriate research methodology; originality and significance of contribution to knowledge; rigour in reasoning and inference; scholarliness of thesis structure and presentation; and worthiness for publication.

In the context of student performance in different courses in different fields, the ways in which the courses may be compare-able need to be identified and stated explicitly. Of particular value in proceeding along this path are higher-order criteria. Among the aspirations for academic learning, and hence among the potential criteria for academic achievement are: critical analysis; problem solving; locating, evaluating and using relevant information; effective communication; respect for evidence; and originality, initiative and creativity. Some properties function as constitutive criteria in particular disciplines and professions, examples being safe practice in the health fields and artistry in certain performing and visual arts. In some scientific and technological fields, correctness, robustness and efficiency of solution strategies are constitutive of quality, and could allow very different objects to be compared. On a slightly different plane sit sophistication, complexity and rigour. All these types of elements are broad indicators which should need no explanation or defence; they are openly promoted as some of the defining elements of higher education.

Consider 'critical analysis' as a representative example from the list above. As a higher-order criterion, it is used across a range of fields because something of that label is both identifiable and highly valued. This does not imply that it has essentially the same interpretation, implying similar structure and cognitive demand. High-level criteria enjoy wide endorsement largely because of their generality (House 1977):

> Although the label … [critical analysis] … is compact and convenient, it applies to a rich and generalised idea whose power lies in its ability to transcend particular cases. In any concrete situation, a meaning appropriate to the context has to be generated. There is no logical necessity for meanings to be expressed in identical terms in different contexts. (Sadler 1985, 290)

Recent empirical research on this topic has revealed wide differences in interpretation of specified attributes in different fields, and even within different subdomains of the same field (Jones 2009). Critical analysis expresses itself differently in music, in information technology, in history and in construction engineering – and at different academic levels within each of them. Within construction engineering, the interpretation may also depend on the purpose for which the critical analysis is required.

Although criteria naturally take different forms in different contexts, conceptual overlaps are common and, for determinations of compare-ability, non-trivial. Higher-order criteria are critical for determinations of the compare-ability of achievement in what are designated here as 'cognate' courses. These are courses located within a particular field that have substantive similarity in concepts or subject matter, but not necessarily in course objectives and structure. Courses in a cognate cluster may share some common ground (curriculum overlap); be conceptually 'contiguous' (a curriculum sequence); and/or simply be in the same field and bear a generic similarity to one another. That cluster may share membership and so be connected with other clusters forming a wider network in a disciplinary or professional field. In discussions within cognate courses, various shades of interpretation are likely to emerge, in time giving rise to a shared vocabulary that fosters professional communication and understanding. For reaching at least an approximate consensus on what a term such as critical analysis is to mean, such discussions offer more scope than would be possible otherwise. Part of the reason is that deliberations about initial apparent differences in cognate fields are more likely to sharpen the language of discourse until the essence of critical analysis becomes clear. Another part of the reason is that courses chosen at random may well be too disparate for conversations to begin.

It is not necessary that those involved in the review process have actually taught the courses, but they must be able to ask quality assurance questions of an appropriate type and depth and engage intelligently in subsequent discussions. They should be able to make judgements about (or at least query) the demands made on students in terms of appropriateness for the context of the course, and whether assessment task specifications could, if taken literally, be satisfied by only lower-order productions. Members should be able to judge the sophistication of student responses and the grades which best fit them. The process and the consequences run parallel to those that occur in consensus moderation described earlier, but this time the overriding consideration is the integrity of grades.

In that the higher-order outcomes provide both the rationale and the main tool for engaging in compare-ability deliberations, they are, educationally and professionally, as legitimate as specific knowledge in a given field. Whether the subject matter content is viewed as a vehicle for developing higher-order outcomes, or the focus on higher-order outcomes is viewed as instrumental in the development of advanced subject matter proficiency, is immaterial. They amount to much the same thing in practice.

The intention in separating compare-ability from com-prability in this article has been to make several points. First, compare-ability across a number of cognate

courses, each of which has its own objectives and characteristic operational criteria, can be established through using higher-order attributes to link them together in a critically important way, without compromising the unique contribution each course makes to an academic programme. That said, to the extent that the criteria for judgements of achievement are similar in different courses, connections at the higher level may be easier to make. Second, focusing on higher-order attributes and criteria is consistent with the generalised goals of higher education as a social enterprise, and an important aspect on which it is judged externally. Finally, the separation has served an explanatory purpose which legitimates cross-course peer review as a scholarly activity. Com-prability does not need to be analysed in similar depth because most of the hard work has already been done and furthermore, the two technically separable interpretations of comparability raised earlier in this section can be operationalised as a single process.

Assuming that the review engages a panel of suitably qualified peers, the following information would be relevant: the course title; a brief description of the course subject matter content; the year level of the course; and the place of the course in the academic programme (e.g. whether it is essentially a service subject or constitutes fundamental knowledge for development in a discipline or profession). The point of peer review is for the panel, in its collective knowledge and experience, to assure that the grades are awarded, so far as it is possible to tell, in ways that are strictly commensurate with the student's level of achievement, and that the grades carry approximately the same value from course to course, with no course being intrinsically simple or easy and none unreasonably difficult.

Aligning the standards in one higher education institution with those in another and with those expected by relevant accreditation agencies, discipline associations, professional bodies or employers, is the ultimate goal (Sadler 2011). This is the arena in which comparability across courses has to be played out in full. If pursued successfully, with light but adequate sampling of works from students late in their degree programs in different institutions, it would enable statements to be made about academic achievement standards with hard evidence to back them up. Institutions which form consortia for that purpose, or opt into an existing scheme, and collaborate with others in pursuit of assured standards would then know with a reasonable level of confidence that the grades entered on their student transcripts represent with substantial accuracy the levels of achievement their students reach. At the same time it would leave institutions free to follow whatever internal political, policy and procedural agendas they devise. Putting the emphasis on assuring course grades and academic standards at the end point may provide a sound return on investment.

Containment of workload through calibration

Whereas moderation relevant for a single assessment task is repeated for subsequent tasks, the ultimate objective is the development of 'calibrated' academics. The ordinary concept of calibration is straightforward. Periodic calibration of physical instruments (such as weighing machines) is common practice. A high-quality weighing scale can usually be relied upon to produce, time after time, accurate measurements of weighed objects but it cannot be assumed that accurate readings will be produced indefinitely. A public weighing machine is tested against standardised weights at scheduled recalibration intervals and whenever the accuracy of its

readings is questioned. The same principle applies in the context of grading. Through engagement with certain calibration procedures, assessors become able to tune their judgement-making ability. Professionally calibrated assessors would accept responsibility for grading against agreed achievement standards, participate in periodic (but not continuous) checking and recalibration, perform the bulk of the decision-making independently, and consistently produce grades with the desired properties without the need for third-party confirmation or adjustment.

The calibration process, if it were to be effective, initially could be expected to require time and care to establish, but may take place at times separated from the normal deadlines of marking student work. Ideally, this would be for all academics; a more realistic prospect is for this capability to become sufficiently devolved and distributed among a sizeable body of academics to enable them to function as local custodians of standards knowledge, informing and guiding the decisions of colleagues, including short-term and part-time teachers. Such an aim could, if it were achieved, stimulate concomitant changes in teaching, learning and assessment practice, and changes to institutional priorities in the deployment of teaching resources. Assured achievement standards and course grades may then lead to higher levels of student achievement without imposing standardisation of curriculum, teaching or assessment.

Conclusion

Given the degree to which years of practice have accustomed academics to their current ways of doing things, changing the patterns of both thinking and practice would not be rapid or simple. Such changes never are. A significant part of the change could come about through replacing the narrow concept of 'moderation', which implies the resolution of differences, with the broader concept of 'calibration' of academics. Agreed and deeply internalised standards logically call for periodic review to ensure they remain relevant and up to date. The goal is for academics to be confident in their own informed and calibrated judgements, and able to trust their colleagues' abilities to make routine appraisals of student works with an appropriate degree of detachment and self-regulation. Furthermore, the way in which academic achievement standards are assured needs to be transparent to colleagues, students, quality assurance agencies and the wider society. The pursuit of assured grades and academic standards could, if successful, have far-reaching implications for teachers, graduates and higher education institutions. This article is an attempt to provide background thinking for such a quest.

Notes on contributor

D. Royce Sadler is a senior assessment scholar at the Teaching and Educational Development Institute, The University of Queensland, Brisbane, Australia and Professor Emeritus of Higher Education, Griffith University, Brisbane. He has published papers on formative assessment, achievement standards, grade integrity and summative assessment.

References

Abercrombie, M.L.J. 1969. *The anatomy of judgement: An investigation into the processes of perception and reasoning*. Harmondsworth: Penguin.

Bereiter, C., and M. Scardamalia. 1993. *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago, IL: Open Court.

Brown, R. 2010. The current brouhaha about standards in England. *Quality in Higher Education* 16, no. 2: 129–37.
Bruner, J.S., J.J. Goodnow, and G.A. Austin. 1956. *A study of thinking*. New York, NY: Wiley.
Dewey, J. 1939. *Theory of valuation*. International encyclopedia of unified science vol. 2, no. 4. Chicago, IL: University of Chicago Press.
Dreyfus, H.L., and S.E. Dreyfus. 1986. *Mind over machine: The power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.
Dreyfus, H.L., and S.E. Dreyfus. 2005. Peripheral vision: Expertise in real world contexts. *Organization Studies* 26, no. 5: 779–92.
Elton, L. 2004. A challenge to established assessment practice. *Higher Education Quarterly* 58, no. 1: 43–62.
Hartog, P., and E.C. Rhodes. 1935. *An examination of examinations*. London: Macmillan.
House, E.R. 1977. *The logic of evaluative argument.* CSE Monograph Series in Evaluation, 7. Los Angeles, CA: Center for the Study of Evaluation, University of California.
Jones, A. 2009. Rediscliplining generic attributes: The disciplinary context in focus. *Studies in Higher Education* 34, no. 1: 85–100.
Linn, R.L. 1993. Linking results of distinct assessments. *Applied Measurement in Education* 6, no. 1: 83–102.
Marton, F., and S.A. Booth. 1997. *Learning and awareness*. Mahwah, NJ: Lawrence Erlbaum Associates.
Oppenheim, A.N., M. Jahoda, and R.L. James. 1967. Assumptions underlying the use of university examinations. *Higher Education Quarterly* 21, no. 3: 341–51.
Saaty, T.L. 1977. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology* 15, no. 3: 234–81.
Sadler, D.R. 1985. The origins and functions of evaluative criteria. *Educational Theory* 35, no. 3: 285–97.
Sadler, D.R. 1987. Specifying and promulgating achievement standards. *Oxford Review of Education* 13, no. 2: 191–209.
Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, no. 2: 119–44.
Sadler, D.R. 2009a. Grade integrity and the representation of academic achievement. *Studies in Higher Education* 34, no. 7: 807–26.
Sadler, D.R. 2009b. Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education* 34, no. 2: 159–79.
Sadler, D.R. 2010. Fidelity as a precondition for integrity in grading academic achievement. *Assessment and Evaluation in Higher Education* 35, no. 6: 727–43.
Sadler, D.R. 2011. Academic freedom, achievement standards and professional identity. *Quality in Higher Education* 17, no. 1: 103–18.
Scriven, M. 1972. Objectivity and subjectivity in educational research. In *Philosophical redirection of educational research* (71st NSSE Yearbook), ed. L.G. Thomas, 94–142. Chicago, IL: National Society for the Study of Education.
Starch, D., and E.C. Elliot. 1912. Reliability of the grading of high school work in English. *School Review* 20, no. 7: 442–57.
Thurstone, L.L. 1927. A law of comparative judgement. *Psychological Review* 34, no. 4: 273–86.
Urmson, J.O. 1950. On grading. *Mind* 59, no. 234: 145–69.
Wittgenstein, L. [1967] 1974. *Philosophical investigations.* Trans. G.E.M. Anscombe (Original work 3rd ed. published 1967). Oxford: Basil Blackwell.