The role of academic teacher judgments in assuring achievement standards

D. Royce Sadler, Griffith Institute for Higher Education, Griffith University, Brisbane, Australia
r.sadler@griffith.edu.au

Note: This document is essentially the presenter’s prompts, and will make most sense to those present when it was delivered. The substance is currently under further development for later publication as a journal article.

Broad context for this talk:

- Specifically, higher education, but the principles are not limited to that sector.
- Features of higher education: Academic flexibility; specific minimum requirements often exist for program accreditation (especially for professional registration and practice) but considerable flexibility remains. Generally, degree curriculum is not standardised, and neither institutions nor academics wish it to be.
- Concept of ‘peer review’ is well established and widely accepted as the primary tool for quality assurance in research funding and publication.
- Changes in past 20 years: Massification; widening participation rates; lowered minimum entry requirements (many students less well prepared); financial constraints (higher fees, higher costs, less government funding); increasing dependence (in UK, USA, Australia and other countries) on fees from international students; rising importance of international rankings of universities.
- Institutional concerns about reputations (relevant to marketing higher education).
- National concerns over academic standards; concerns about the integrity and comparability of course grades (i.e. without grade inflation) and, where relevant, concerns about degree classifications (1, 2a, 2b, 3). QAA and national equivalents.
- International concerns about (or interest in promoting):
  - Cross-border recognition of professional qualifications (substantive content of qualifications). Bologna, Tuning focus on these more than on actual standards.
  - Student mobility with credit (e.g. European HE Area; Bologna agenda).

Terminology used in this talk: Program: full suite of studies for a degree. Course (~ subject, unit, module, paper): a component of a degree. Grade (~ mark, score, point, percentage) as recorded for each course as a whole.

- Not part of this talk: degree classifications – which has a high profile in some countries (e.g., UK); low profile in others. If course grades could be trusted, degree classifications based on them would be less of an issue (except for tradition).

Narrow context and outline for this talk:

- Focus is: Academic achievement standards and how they may be assured in ways that are (1) theoretically sound; and (2) applicable to both undergraduate education generally (programs and courses) and advanced coursework programs.
- Reason for concentrating on academic achievement: “academic standards” as a term is used to cover a broad range: facilities, teaching quality, resources etc. The interest here is specifically in achievement (~ proficiency, competence, capability, attainment, performance, accomplishment – whatever term best suits particular disciplines and fields for
what students have learned during their degree programs and carry forward as personal capital for life and work.)

- Special focus is on **course grades** — not Grade Point Averages (GPAs) etc.

- **Traditional approaches** to ‘assuring’ course grades:
  1. **Uniform mark/score/point aggregate cut-offs** e.g. 85-100 points = A, 7, High Distinction (HD). Easy or tough marking? Memorisation versus higher-order academic outcomes? No problem: use same cut-offs for all courses for all aggregates of marks.
  2. **Control of the proportions** of different grades (grading on the curve) (e.g. top 5% get ‘Distinctions’). Purely relative; no attempt to address ‘absolute’ standards.

- **None of these has delivered to date. None can deliver even if dressed up. That is one reason the search for alternatives is in full swing.**

- **Current or proposed developments:**
  4. Various forms of ‘**codification**’, with or without supplementation by examples;
  5. **Standardised national assessments** of **graduate** (~ generic) **attributes** (~ skills, competencies, outcomes). Employer feedback influential; appeals to governments; allows rankings of institutions. Reductionist; outsources the issue of standards. Not here yet, but on the near horizon. Watch OECD’s AHELO.

- **Both codification and graduate testing are seriously problematic; the focus in this talk is on the first (codification).**
  - Part 1: Analysis of codification as a general philosophy and approach; why it cannot work to achieve what is being expected of it. This is the major part of the presentation.
  - Part 2: A way forward: direct dependence on devolved teacher/assessor/marker judgments; rank and file decision making; concept of ‘calibrated’ assessors (which in this situation are academics) based on peer consensus processes.

### Part 1: What is this thing called **codification**?

- Codification (of standards) – elaborated verbal descriptions or specifications of outcomes or mixes of outcomes to guide the award of marks, scores or (course) grades.

- A codification is **any formal representation of a body of knowledge by such means as words, diagrams or symbols**. It is designed to express knowledge in a compact material form so as to facilitate storage, transmission, application and analysis. Laws, regulations, policies, rules, instruction manuals, scientific papers.

- Variety of forms and terms used in relation to **higher education standards** at present:
  - Subject benchmark standards;
  - Discipline standards statements;
  - Thresholds standards;
  - Grade descriptors;
  - Rubrics; and
  - Marking guides.

- **SEE DISPLAYS 1A and 1B**

- These are intended to:
  - Be more descriptive and informative about actual outcomes than unadorned codes such as 7, A, HD;
  - Guide assessors in the award of marks or grades – more objectively;
• Provide more transparency to students and others; and
• Enable standards to be held broadly constant from year to year.

Nationally and internationally, codification of academic standards is widely believed to be the way to go.
• Huge investment in national projects. Considerable momentum;

Codification is incapable of delivering what is needed.
• This is a strong assertion, and there are several good reasons. Here is the main one.
• The terms used in the codification of standards do not possess concrete referents (as do industrial standards).
• SEE DISPLAY 2
• Elements are both qualitative (aspects, characteristics, properties, ‘criteria’) and quantitative (how much: minimum required, or maximum allowable, to qualify).
• These are general characteristics of industrial, technical, product and commercial standards.
• The characteristics are separable and independently measurable on standard scales with standardised units (and so are 100% intrinsically valid measurements).
• Failure to measure up on one property amounts to failure overall; no compensation (trading off) of performance on different criteria is permitted.
• Testing apparatus can be constructed at will, any time, any place.
• These are not comments about the appropriateness of a standard that is set and to be applied. It is about the composition and structure of technical standards.
• RETURN TO DISPLAYS 1A and 1B AGAIN (Either will do.)
• Observe the mix of the qualitative (criteria etc) and ‘quantitative’ aspects, the latter signified by qualifiers, modifiers and possibly hedge words.
• Qualitative elements (criteria) are conceptually and contextually interpreted.
• Cannot regard the characteristics as intrinsically separable; may be distinct in the abstract but overlap when applied. This is a common phenomenon.
• Not measurable on standard scales; no standardised units; ‘measurement’ is assessment ‘instrument’ dependent; invariably involve multiple judgments.
• Grading/marking is making categorical decisions along a continuum. Not GO/NOGO. Cut-points are judgmental; intersubjective agreement is negotiated (which is OK) BUT once ‘standards’ are set they must become stipulative.
• The idea of ‘personal’ standards is inimical to assured grades. (Any concept of academic freedom that extends to individuated academic standards is untenable and needs to be challenged.)
• The terms in Displays 1A and 1B (aspects and amounts) can be interpreted to fit a wide range within any set of concrete objects up for appraisal. This is basically because the terms are relative rather than definitive. In blunt terms – and in practice – the descriptions can be interpreted so as to fit or ‘cover’ almost any collection of student responses to an assessment item, including sets from different levels of education. There are, intrinsically, no ‘absolutes’.
• Descriptions are by necessity open in nature; open with respect to the number of criteria; open with how levels are mixed. It is impossible for codifications in educational assessment to match industrial standards.
• Trade-offs (compensations) are not only possible but inevitable. Education does not deal with standardised products coming out of a manufacturing process. But the idea of quality still counts.

• Student works which would be judged of equivalent quality could well differ in respects other than those mentioned in the description – and often do.

• Option 1: Become more specific.
  o Consequences: (a) Applies to fewer cases; (b) Harder to get consensus.

• Option 2: Supplement the descriptions with exemplars. This could help, but requires interpretation.
  o If the description is based on a set of real cases, it will go for the most common criteria – the ‘highest common factor’ (HCF). This necessarily means it will be unable to take into the judgment some aspects or properties that justify a higher or lower assessment than rigid application of the prescribed (HCF) criteria.
  o If the description is devised in the armchair or by a committee, it will most likely be more general. This means that the relation between a particular exemplar and the codification will require more interpretation and elaboration.

• Option 3: Prioritise direct judgments of achievement made according to the (raw) evidence of achievement, namely the collection of all student works that contribute to summative assessment, with NO HOLDS BARRED.

Part 2: Moderation scaled up

• Moderation as commonly implemented to improve inter-scorer consistency for appraising student responses to a single assessment has to be generalised to whole-course grades; this represents a sizeable and uncommonly made jump, but is necessary for course grades to be properly assured.

• The big question: Using as primary evidence all student works submitted for summative assessment, what level of achievement (course grade) is warranted?

• This bypasses the accumulation of marks, and any codifications that exist. How can a system which is not based on explicit descriptions function?

• A clue has already been given already in the use of the phrase “judged of equivalent quality”. People can (correctly) judge the quality of things without prespecifying the criteria that will be used.

• Prespecifying the criteria makes untenable assumptions about which properties will matter in particular instances, and thwarts (even pre-empts) a high quality determination. It frames or constrains the judgment.
  o We should prioritise judgments of quality, and then provide justifications for those judgments. This is a requirement of professional decisions. Without proper justifications, they are likely to include elements of opinion or taste.
  o Each justification (explanation) for a judgment will be in words which apply specifically to that instantiation, so the ‘fit’ between the item itself and whatever properties substantiate the judgment of quality is good.
  o The criteria enter the arena only at the point of justification: the judge needs to invoke whatever criteria are salient to the judgment – and only those.
  o Not employing uniform criteria for all cases helps to get at the underlying standards, which are abstract, and not formally ‘defined’ by codification.

• The ‘standards’ to which we as professionals respond are in our heads. They are abstract, but no less real for that.
• The residual problem is that we all have our own ideas about standards. We need to be calibrated against one another, and against the expectations of the wider world, so we can (later, independently) judge consistently and professionally.

• I refer here to rank and file academics, who constantly make judgments about student achievement. That is the very site to which we must attend.

• This process breaks with tradition, but is based on the fact that what makes for ‘quality’ in complex works, or collections of evidence, cannot be captured by any strict definition. However, competent appraisers can consistently identify quality when they see it. And also provide valid grounds for their judgments.

• Outsourcing the responsibility for assuring academic achievement standards to external quality assurance or testing agencies, however well intentioned, removes the responsibility from academics for working on and through the calibration issue. It disempowers academics, and fails to address the core problem of assuring course grades. Third-party interpretations of grades and transcripts would then continue to reflect either existing institutional reputations (which may be ill-deserved) or, if standardised testing were mandated, institutional rankings based on generic test results.

• A lot more needs to – and can – be said about these types of judgments, but that has to be for another place, another day.

Thank you for your attention.

References


### Display 1A

**Example of a rubric**

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Level A</th>
<th>Level B</th>
<th>Level C</th>
<th>Level D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding</td>
<td>Authoritative command over all issues; high level of integration.</td>
<td>Solid level on most major issues; only moderate levels for others.</td>
<td>Generally superficial or partial; significant gaps on some key topics.</td>
<td>Limited and patchy across course material.</td>
</tr>
<tr>
<td>Analysis</td>
<td>In-depth analysis; aspects identified and interrelated; originality evident.</td>
<td>Showed some insight, but overly derivative; key connections missed.</td>
<td>Adequate in parts; many weak or unsupported generalizations.</td>
<td>Almost entirely descriptive; little or no evidence of insight.</td>
</tr>
<tr>
<td>Writing and</td>
<td>Exemplary in all respects, with no or few lapses.</td>
<td>Technically competent, with minimal errors; but lacking in flair.</td>
<td>Acceptable; imprecision common; may be a tendency to verbosity.</td>
<td>Poor or sloppy expression; technical aspects deficient.</td>
</tr>
<tr>
<td>Presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Display 1B

**Examples of standards statements adapted from several universities**

- Strong evidence of highly original thought; excellent analytical and critical abilities as well as a thorough grasp of the topic from background reading and analysis; also demonstrates exemplary organizational, rhetorical and presentational skills.
- Demonstrates basic understanding and comprehension of most of the topics; acceptable mastery; some evidence of critical thinking; information presented mostly derivative; leans towards reproducing material with little evidence of original thought; some sweeping generalisations.
- Basic aspects are addressed and researched to some extent; demonstrates mainly description, showing basic understanding of the topic but no application; little evidence of analysis but no clear and logical argument relating to the subject; draws primarily upon course materials.
- Demonstrates evidence of being able to assemble the bare minimum of information, poorly digested and not very well organized in presentation. Inaccurate, poorly argued, redeemed only by evidence of some acquaintance with the subject.
2.6 MEDIUM IMPACT RESISTANCE

2.6.2 Performance criteria

When tested in accordance with Appendix M, an ocular shall be considered to have failed—

(a) if it cracks through its entire thickness into two or more pieces;

(b) if more than 5 mg of the ocular material becomes detached from a part of the ocular surface remote from the surface struck by the ball;

(c) if the ball passes through the ocular; or

(d) if contact is made with either eye of the test headform by the ball, frame, ocular or any part or fragment of these.

Extract from Appendix M (Method for the determination of Medium Impact Resistance)

M4 PROCEDURE

...  
(c) Project a nominally 6 mm diameter steel ball at a velocity of 45 +1.5, –0 m/s or a 6.35 mm diameter steel ball at a velocity of 40 +1.5, –0 m/s onto each of the following impact sites, repeating Steps (a), (b) and (e) for each impact site:

(i) At the reference point of each ocular as given in Clause 2.4.1.

(ii) At normal to the surface of the frame above the ocular, within 20 mm of the mid-line (as given in Figure 2.4) of the frame.

(iii) At 90° to straight ahead along the ocular through the centres of the front of both eyes of the headform.

(iv) A welding eye and face protector shall additionally be tested at the midpoint of each side of the cover plate within 5 mm of the edge of the filter retaining assembly.

NOTE: Unfitted oculars need only be tested at site (i).

(d) In case of dispute, the 6.00 mm diameter ball shall be used.

...  

* This extract is limited to standards for oculars of medium impact resistance. The full 109-page Standards document includes a large number of other specifications. The section on impact resistance contains detailed specifications of the testing equipment and process. Among other things, these cover the mass of the 6 mm steel ball, propulsion system, projectile guidance, ricochet control, support for the specimen to be tested, minimum accuracy for time measurement, point at which projectile velocity is measured, test equipment setup, data recording devices and reporting requirements.