Assuring Academic Achievement Standards at Griffith University

Second Edition, June 2010

D. Royce Sadler, Griffith Institute for Higher Education

The issue of assuring academic standards is high on the list of matters that Australian universities are currently being asked to consider. Part of the explanation is that most systems of checks and balances used in Australian (and many overseas) universities are inherently limited in what they can accomplish. For example, statistical or other procedures (such as controlling the proportions of the different grades awarded) applied to secondary data (specifically marks and grades) cannot get to the core issue of whether marks are assigned strictly in accordance with the levels of student achievement. Similarly, systems which rely essentially on following certain procedural protocols are not necessarily sufficient to assure that internationally competitive achievement standards are maintained. This is not an allegation that standards are too high or too low. It simply expresses the concern that universities in our higher education sector may be vulnerable unless they have comprehensive, soundly based quality assurance systems in place.

The context internationally

Developments for assuring academic standards are under way in various international contexts. Some of these are occurring in individual countries (such as the UK and Hong Kong) and are being spearheaded by their respective quality assurance agencies. Others cover groupings of countries, examples being the 'AHELO' project in OECD countries, and the Tuning/Bologna project in the EU. Schemes that rely on benchmarking and external examiners (or reviewers) provide partial solutions, but offer only limited direct access to academic standards by academics and students. Broad-spectrum graduate attribute tests, which are not yet commonly implemented, continue to be advocated as a way of achieving comparability across academic departments, programs and institutions, possibly through scaling results or ranking institutions.

The current context nationally

The recent Bradley Review of higher education in Australia provided a sweeping analysis of the sector, and contained numerous recommendations for change. Specifically on the issue of academic standards, it had this to say (p. xxii & p. 137):

> **Recommendation 23:** That the Australian Government commission and appropriately fund work on the development of new quality assurance arrangements for higher education as part of the new framework set out in Recommendation 19. This would involve:
> - A set of indicators and instruments to directly assess and compare learning outcomes; and
> - A set of formal statements of academic standards by discipline along with processes for applying those standards. *(Chapter 4.1).*

In 2009, the Australian Universities Quality Agency (AUQA) circulated a Discussion Paper on academic standards. This paper recommended that generic statements of standards be established in discipline and professional areas or fields. The Australian Learning and Teaching Council (ALTC) is also active on this front. Two of its seven 'Designated Responsibilities' are related directly to assessment and standards, namely:

- Liaison with the sector about options for articulating and monitoring academic standards; [and]
- Improvement of assessment practices throughout the sector, including investigation of the feasibility of a national portfolio assessment scheme (ALTC 2008, p.6).

Closer to home, assessment standards were specifically identified in the 2008 AUQA *Griffith University Audit Report* as an area needing attention, as set out in this extract from the Report:

> To ensure that the University and community can continue to have confidence in the standards being applied in the awarding of student grades and degrees at Griffith University, the Panel makes the following recommendation which is marked 'urgent' because of both its significance to the maintenance and enhancement of academic outcomes and standards, and for the currency of the activity involved.
>
> **Recommendation 7 (urgent)**
> AUQA recommends that more attention be paid by Griffith University to quality control aspects (such as moderation policy and procedures and the calibration of standards for the awarding of grades) in Stage 2 of the current Griffith University Assessment Project, and that Griffith Institute for Higher Education be more proactive in disseminating the good practice in assessment guidelines by providing a tailored and targeted academic support program. (p. 25)

The focus on academic achievement standards

The term *academic standards* can be interpreted broadly. It has been used in referring to curriculum content; program structure (for accreditation purposes); teaching and learning (in evaluating approaches and how they are implemented); and student learning outcomes. All of these are legitimate areas of application, and all are worthy of attention. However, the focus in this consultation paper is on academic *achievement* standards. Achievement is inferred from student responses to assessment tasks, and is often represented by marks, scores or other symbols. Ultimately, these appraisals lead to grades in academic courses, the grades being the main object of interest for quality assurance purposes. The AUQA report for Griffith University specifically targeted 'moderation policy and procedures and the calibration of standards for the awarding of grades'.

The main approach being explored elsewhere

In most universities, the starting point for assuring academic standards has been with statements of course objectives or Intended Learning Outcomes (ILOs). These are typically expressed in broad terms suitable for guiding teaching and learning activities, and are then expressed in finer and finer detail so they can apply to individual assessment tasks. The underlying premise is that clear and explicit statements of ILOs, criteria and standards will lead to consistency in marking and grading, and produce comparability across courses.

Despite its popularity, this approach is problematic on several counts, of which five are mentioned here: (a) the causal chain by which statements of objectives or outcomes can be translated into consistent academic achievement standards has not been rigorously articulated; (b) the progressive mappings from outcome statements to grade descriptors may not proceed smoothly, because outcome statements and grade descriptors serve inherently different purposes; (c) regardless of the level of detail of any statements produced, they almost invariably refer to at least some abstract concepts (such as 'analytic reasoning', 'creativity', and 'independent thinking') and include qualifiers, modifiers and hedge words (such as 'high level',

'some', 'effective', and 'appropriate') which require interpretation before they can be applied, the problem being that different academics put their own interpretations on such words; (d) the approach makes no direct attempt to address comparability of grades across courses; and (e) the method has not yet been shown empirically to deliver assured standards, consistency in grades awarded, or comparability across courses.

The current approach at Griffith University

Griffith's approach to assuring academic achievement standards is in two stages.

Stage 1: Academics score or mark student responses, and from these data propose grades for students according to their own professional judgments. Stage 2: The relevant Assessment Board or Panel reviews the distributions of proposed grades from all courses. Any distribution that differs ‘too much’ from institutional or faculty conventions is discussed with the Course Convenor with a view to adjustment. Stage 2 is simple to apply, and produces ‘results’ with a minimum of fuss except where course enrolments are small or different cohorts are demonstrably unequal in overall ability or diversity. For the latter cases, variations may be permitted.

What does the Stage 2 adjustment set out to achieve? First, it aims to ‘correct for’ grade proportions that are too skewed towards the high end (undeserved high grades, that is, grade inflation) or towards the low end (not enough high grades, or with too many students failing or barely passing). These are assumed to reflect ‘standards’ that are set too low or two high respectively. Second, by making the grade distributions from different courses broadly similar, the aim is to achieve grade comparability across courses. In reality, the strategy fails to achieve either aim. Here are three of the reasons.

First, the method is concerned with achievement only in a relative not an absolute sense. Firm expectations about acceptable proportions of grades mean that student grades are to some extent influenced by performances of other students in the same cohort, yet individual students have no control over the membership of the cohort or the cohort's achievements. Controlling grade proportions is therefore inherently unfair. In effect, it constitutes a form of norm referencing, which is both counter to University policy, and outmoded as a quality assurance strategy.

Second, Assessment Boards typically operate with proposed grades and their frequency distributions, but do not access the primary evidence of student achievement. As a result, the approach is structurally blind to: the quality of course assessment plans and assessment tasks; the range of student activities and work for which marks are awarded and counted as 'achievement'; and the quality of teaching, learning and actual achievement.

Third, the approach is not capable of guaranteeing that the standards used for grading student work remain the same from year to year. It resets the grading parameters not only for each year’s cohort but also for each course cohort. This makes it all but impossible to detect any long-term drifts in standards.

Outline of the solution proposed for Griffith University

Consensus moderation is generally accepted as both appropriate and necessary whenever multiple assessors are involved in marking student responses to a single assessment task. The proposed strategy is to regularise and refine the existing practice of consensus moderation, and extend its scope, first to cover achievement over a whole course, and then comparability of achievement across cognate courses. Where courses are taught and assessed by a single

academic, a colleague would necessarily be involved in the moderation processes. Throughout, the process would work directly from primary data, by which is meant actual student works, not marks or scores. This has always provided the hard evidence for achievement. If an assessment task does not require students to either produce artefactual outputs or leave some other type of physical record, a secondary artefact such as an audio or video recording would need to be created. This may be necessary for musical, dramatic, clinical or similar performances. The current default way of indirectly 'assuring' appropriate grades by controlling the proportions of grades would give way to an approach which addresses the issue directly – at the site of grading decisions. This is where the assurance of academic standards has to be dealt with. If implemented with care, the alternative process would reduce the likelihood that the grade distribution problem would arise at all. This is not only sound in theory but also consistent with the AUQA recommendation.

A common method of moderating the marks awarded by different assessors in a course is for them to all trial-mark the same sample of student responses to a given assessment task. They then compare the marks they tentatively allocate, engage in focussed discussion and come to agreement on appropriate academic achievement 'standards' to which they agree to mark the remainder of the student work. The latter is often carried out by the markers acting independently, with perhaps some cross-checking of another sample of marks after the event, and with discussion of difficult cases. This model is referred to in this paper as *consensus moderation*. Observe that it provides a concrete environment in which to work, not an abstract environment of words, descriptors and statements. It taps directly into the primary evidence of student achievement. Furthermore, it is imperative to engage in consensus moderation whenever casual staff or contract markers contribute to marking student work.

During the moderation process, members of the teaching team come to agreement on two fronts, which they typically do not distinguish in their own minds, and do not need to. First, they come to a shared understanding about what constitutes *quality* in the context. Recognition of quality is a fundamental evaluative act, which can be substantially (but not necessarily totally) decomposed and explained afterwards. Many academics in many teaching contexts find it difficult to define something as elusive as 'quality', but are nevertheless capable of recognising it when they see it. Part of this ability involves recognising works from different students as of 'comparable' quality when the works themselves are different. Second, members of the team come to consensus on how various points along the continuum of quality are to be mapped to the symbols in use (such as marks) so the latter can be assigned meaningfully and consistently. The overall objective of this exercise is for assessors to assign the same symbols for the same quality of work. In effect, they set 'standards' by consensus for student responses to a particular assessment task. Because this process is the starting point for further developments in the assurance of grades, moderation of marks for a single assessment task within one course is labelled in this document as **Level 1**. A stage which logically precedes Level 1 is outlined below.

The quality of the evidence. Consensus moderation mostly deals with marking standards for student responses to a single assessment task. However, this is but one element of the wider agenda. The first of the additional aspects has to do with the course assessment plan and the design and specifications for each assessment task. As alluded to earlier, an assessment plan may include a number of elements which are intended to count towards the course grade but are fundamentally irrelevant to judging the depth and extent of academic *achievement*. For grading standards to have academic integrity, all elements that contribute to the course grade must

provide evidence of achievement. Marks for effort, credit for compliant behaviour such as attendance or participation (except in special circumstances), or penalties for breaches of protocol do not constitute evidence of achievement. Course assessment plans therefore need to come under scrutiny on this front.

Just as important to the quality of evidence is the design of summative assessment tasks, the adequacy of task specifications, and how the tasks are communicated to students. Key aspects are the extent to which the tasks call for responses that require students to demonstrate achievement of the course objectives, especially those that require complex or sophisticated knowledge and skills (the so-called higher-order aims of the course). These tend to be neglected because assessment tasks that tap into them are not always easy to develop. Actual task descriptions as presented to students form the obvious starting point for appraising adequacy. The question to be asked is: If taken literally, do the specifications point students in the desired direction? Other shortcomings may become apparent only during the course of appraising student responses. Design, specifications and communication are all important to get right. The aim is not to tell students how they should go about doing the task; it is to describe and explain the nature of the required end 'product'. Because assessment tasks are intended to assess students' levels of achievement, students should not have to guess what the lecturer had in mind when the assessment item was constructed or selected.

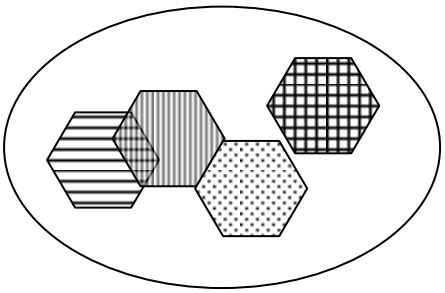
A typical example of defective specifications is when the task description does not actually require students to do much more than regurgitate memorised facts or other information, or complete some task in a routine way. If the lecturer expects significant operations (such as analysis, extrapolation or critique) to be performed on the information, then that needs to be made clear. Otherwise students are left to make assumptions as best they can. Some 'successful' students will, of course, know or can guess correctly what to do. Others will take the words literally, or read their own meanings into them, and thereby fail to take the opportunity to demonstrate their achievements. In the example given, a mere reproduction or compilation of information would very likely be penalised as a shallow response, although some of those students would have been capable of more performing complex operations on the information.

This naturally implies that a step needs to be taken before the main Level 1 consensus moderation is carried out, namely, peer review of assessment plans and task specifications. To formalise the situation described so far, consensus moderation of academic judgments and associated marks is the most commonly practiced level, and serves as the foundation for calibrating assessors. The precursor stage that focuses on course assessment plans and tasks is equally important, and is referred to in this document as **Level 0**.

Level 2: Peer review of grading in a course. The consensus moderation process also needs to be scaled up to higher levels in order to provide a systemic quality assurance solution. Level 2, the next level up, covers grading based on student responses to several assessment tasks in the same course (that is, achievement evidence from different sources). The action required is similar in structure to that in Level 1, but wider in scope. It draws on the same fundamental intellectual processes involved in consensus moderation. At this level, academics again seek to arrive at consensus on two fronts: (a) the constitutive nature of *achievement over a whole course*, and (b) the location of appropriate division points along the achievement continuum which represent the boundaries between different course grades, and by implication, the standards being applied. The evidence fused together at Level 2 consists of all responses to all summative assessment tasks *from a sample of students*. The objective of the exercise remains to reach consensus. For each student in the sample, the assessors scrutinise all the evidence of

achievement. The sample of students should therefore be large enough for the assessors to become 'calibrated' against one another. The key question for the members of the review panel is: Does the total evidence for the work produced by a particular student correspond with a grade of HD, D, C, P or F?

Level 3: Comparability of grades across courses again retains the fundamental intellectual processes employed in Levels 1 and 2, but takes another crucial step towards generalisation, again through peer review. (This principle is foundational to determinations of quality in various fields of academic work.) The 'object' of scrutiny now consists of the achievements of students *in different courses*, and the overall aim is to achieve *comparability in grading* across courses. In the attached figure, each hexagon represents one course. Two of the courses share some common ground; two are conceptually 'contiguous', and one sits in the same field and bears similarity to the others. For convenience, these four courses are referred to as 'cognate'. Again, the participants bring to the discussion table actual student works that contribute to the summative grade, together with their annotations setting out the rationales for the proposed grades. This is the arena in which comparability across courses has to be ironed out.



When Levels 0 - 3 are put together, it is possible to reach a conclusion about the comparability of academic achievement 'standards', together with documented grounds for the professional judgments made. Although these processes form the nub of peer-reviewed quality assurance, the 'standards' so determined so far still remain confined to a single institution.

Level 4: External calibration involves aligning the standards in the teaching institution with those employed in other institutions, and with those expected by relevant accreditation agencies, discipline associations, professional bodies or employers. This step is to establish and protect the reputation of the teaching institution. In contexts where experienced academic staff already have first-hand, primary knowledge of the academic achievement standards that prevail elsewhere as a result of close involvement with accreditation processes or employers, that brings a substantial benefit to the calibration purpose, provided that the fundamental data (student works) are interrogated directly. Inspection of grade distributions, verbal descriptions of learning outcomes, or institutional self-reports of their quality assurance processes do not provide first-order evidence of the integrity of the grades.

Locking down standards. One aspect remains to be addressed – the comparability of academic standards across time. Standards need to be stable from one course offering or year to the next offering or year, so that they do not drift around (even incrementally) unless a specific decision has been made to intervene and reset them. What is required is some way of 'capturing' agreed-upon achievement standards in an enduring form so they can be locked down and made available for future use and reference. They need to be useable in deciding grades and facilitative in resolving appeals against grading decisions. They also need to be accessible by academics and students on demand. The most straightforward approach is to create a repository of annotated student works, possibly the same ones used at Levels 2 and 3 of peer review, with their grading decisions and the rationales for those respective judgments. The core material to be archived is, therefore, actual student works or productions that have been graded. Each

collection of student work should be accompanied by a clear statement as to why the evidence shows that the grade awarded is strictly commensurate with the student's achievement.

The challenge

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 will take time and effort. Such changes always do. However, once the basic processes are in place, the labour intensiveness should decrease substantially. A significant part of this would come about through replacing the narrow concept of 'moderation', which implies the resolution of differences, by the broader concept of 'calibration' of academics. Agreed and deeply internalised standards naturally call for periodic reviews, but the goal is for academics to be confident in their own informed and calibrated judgments, and able to trust their colleagues' ability to make routine appraisals of student works with an appropriate degree of detachment and self-regulation. As academics develop their expertise in calibrated grading, appreciate the institutional imperative for quality assurance processes, and become comfortable with a system of periodic checks and balances to ensure the system is operating as it should, they may welcome and value academic achievement standards that are owned by the academic community as a collective rather than by individual academics. In any case, the way in which academic achievement standards are assured needs to be transparent to colleagues, students, quality assurance agencies and the wider society. The fundamental ideas and processes are clearly consistent with long-standing academic values. The proposed system would be building on those foundations to ensure confidence in both professional judgments and the quality assurance system.

References

Australian Learning and Teaching Council. (2008). *Annual Report 2007-08*. Chippendale NSW: ALTC

Australian Government. (2008). *Review of Australian Higher Education – Final Report*. D. Bradley (Chair). Canberra: Department of Education, Employment and Workplace Relations.

Australian Universities Quality Agency. (2008). *Report of an Audit of Griffith University*. Melbourne: AUQA.

The main changes made for this second edition are the addition of Level 0 (peer review of assessment plans and tasks) and Level 4 (external calibration). Additionally, many minor editorial adjustments have been made throughout.

[Filename: AssureAcadStndrds Consultation Paper (Second Edition, Sadler) D14 Create date: 14-Jul-10]