Introduction to the process of moderating assessments

D. Royce Sadler, GIHE, Griffith University

Context and key issues

Student responses to assessment tasks provide the raw data from which inferences are made about student levels of academic achievement. In some areas of higher education, the responses are either correct or incorrect; there is little, if any, leeway. Multiple choice tests and quizzes of factual knowledge provide examples. In other areas, the assessment tasks require students to construct extended or complex responses, and there is no simple, correct answer. Extended responses are common when students attempt to solve complex mathematical, scientific or technological problems (wherever partial credit is allocated for partial solutions), construct written assignments and project reports, demonstrate professional skills or procedures, and produce creative works of various types. The relatively open form of these responses means that they typically differ from student to student. The quality of each student response has to be judged by a competent marker who makes a qualitative appraisal through drawing on personal professional expertise. Academics' own ideas of what makes a high quality student response to an assessment task (and therefore what constitutes high performance in a course) are influenced by (a) having made similar judgements in the past (if any), (b) making assumptions about – or perhaps knowing – how other markers would most likely judge the work, and (c) knowing the external expectations in the relevant discipline, profession or field.

Typically, subjective judgments made by different markers about the same pieces of student work differ from marker to marker, sometimes wildly. Some markers are characteristically generous, some are characteristically strict, and others may be inconsistent. Some markers are influenced by aspects other than the actual quality of the work, such as the amount of effort a student seems to have put into its production, their knowledge of students' previous performances, or their own personal tastes and preferences. Regardless of whether a marker shows a distinctive pattern, students have a right to expect that, within a reasonably small tolerance, identical or similar marks, grades or credit would be assigned to work of a given quality, irrespective of which marker makes the decision. Some processes that can assist in avoiding or minimising unevenness in marking are therefore called for. This is where the process of moderation comes in. The aim is to ensure that, as far as possible, similar academic standards are applied by different markers.

A common way of displaying differences between two assessors is to plot their scores for the same set of student works on a scatter diagram. This is simply a graph in which the X-axis denotes one of the marker's marks, and the Y-axis, the other marker's. The two marks for each student response are used as X and Y coordinates for a point on the diagram. If the markers are perfectly consistent, all the points will lie on a straight line. (Technically, this type of consistency means only that the markers' differentiate among student works in the same way. If one assessor marked consistently higher than the other marker by a fixed amount, the points would still lie along a line.) The typical situation is that the markers are not totally consistent, in which case the points will show some scatter. The amount of scatter reflects the level of consistency between the two sets of marks, and is referred to technically as 'inter-marker reliability'. Other ways of analysing inconsistency include calculating some simple statistics (such as averages and measures of the spread of marks). The task of moderation is to minimize discrepancies among assessors *before* students receive their marks. The English verb 'to moderate' dates from about 1400, and originally meant to regulate, or to abate excessiveness. However, nothing in the basic concept of moderation specifies the best way to do this; it is left open.

Academics who share the marking of large batches of student works can collaborate on how marks are allocated. This is the principle behind the approach known as *consensus moderation*. In its most common form, consensus moderation requires that all assessors mark the same sample of student responses with, or without, prior consultation among themselves. They then discuss the results of their marking in order to arrive at a common view about the grading 'standards' to be used for the whole student group. After the trial marking and conferring, the bulk of the marking may be carried out more or less independently, with only occasional cross-checks. Because this form of consensus moderation forms the basis of an approach to assuring academic achievement standards being developed at Griffith University, more about its significance is provided at the end of this paper. However, the term moderation is also applied to several other processes relevant to marking and grading, and three of these are outlined briefly below.

Other moderation models

Multiple marking. This approach also applies to student responses to a single assessment task within a course, but it does not depend on professional consensus. Two or more markers score all student responses. The separate scores are then subjected to statistical or mechanical

'moderation', which is simply a method of combining them. The simplest method is to average the scores from different markers, with no attempt made to arrive at inter-marker consensus. With three or more markers, a variation on this rule can be to first eliminate the most extreme score (if any) and then average the remainder. (This process is similar to that used in judging and scoring individual performances in certain competitive sports.) Statistical moderation can be – and usually is – carried out without scrutinising and discussing actual student responses. In some implementations, the specified calculations are implemented automatically on a mark spreadsheet as soon as the numbers from different markers are entered. In some UK institutions, double marking followed by averaging is standard practice for all extended student responses. Multiple marking is labour intensive (and therefore relatively expensive) for large course enrolments.

***Use of an external moderator*.** This involves appointing a suitably qualified person as an arms-length moderator for assessments of student achievement in taught courses. Such external *examiners* or *reviewers*, as they are also called, are usually senior academics from other institutions and the system is common throughout the UK higher education sector. External moderators typically scrutinise samples of student works from different courses in a degree program. They then provide professional opinions about the academic achievement standards being applied. External moderators do not usually check on inter-marker consistency. The same basic procedure is used in examining research higher degree theses (where, at Griffith University, the markers are called Examiners, and the broker is called the Chair of Examiners). It is also used in evaluating grant applications and manuscripts submitted to refereed journals. In the latter case the journal editor often acts as the moderator. In these two situations, numerical scores may or may not be used. With external reviewing, live interactions among the assessors or between the moderator and the assessors are relatively rare, and this limits the achievable levels of shared vocabulary and knowledge about standards.

***Reviewing grade distributions*.** This is the review of provisional grade distributions from different courses. Whereas ordinary moderation usually takes place wholly within courses, the review of grade distributions is an attempt to arrive at comparability of grades across courses. It approaches the inter-examiner consistency issue from a broader course-level perspective, and the process is substantially removed from primary evidence of achievement. The usual practice is for a review panel to scan distributions of recommended grades from different courses, looking particularly for those distributions that appear to be in some way unusual or aberrant. At Griffith University, this role has been carried out by Faculty, School or Department Assessment Boards or Panels. Grade distributions that are cause for concern are investigated

with a view to reshaping them until they are ‘acceptable’. The limits of what is acceptable and unacceptable are mostly not clearly defined. The reshaping may take place through rescaling an entire distribution of grades by systematically adjusting all scores so as to give the distribution a ‘more acceptable’ average and spread (mean and standard deviation). Alternatively, individual grade boundaries may be adjusted, essentially by hand and eye. The underlying intention is usually either to limit failure rates (in order to improve retention rates) or to control the proportions of high grades (on the principle that limiting the supply in the face of steady demand maintains their value). As a general economic principle, the latter is often sufficient to maintain ‘value’, even though that value is widely practiced, market-related and has no substantive backing in terms of actual levels of student achievement or performance. Reviewing grade distributions is eminently doable, and always results in a broadly predictable outcome. It has been the default method at Griffith and many other universities largely because workable alternatives have not been available. However, it is flawed on several grounds, of which three are outlined (Sadler, 2009). First, the method is concerned with achievement only in a relative rather than an absolute sense, and this makes it difficult to pin down academic durable achievement standards. It resets the grading parameters not only for each year’s cohort but also for each course cohort. Second, adjustments are made to distributions of grades without re-examination of the primary evidence of student achievement. Among other things, this makes it unresponsive to the quality of teaching and the types of ‘components’ which are included in the course assessment plan under the heading of ‘achievement’. Finally, it is not capable of guaranteeing the integrity of grades from year to year, or detecting long-term drifts in academic standards.

Consensus moderation revisited

The process of trial marking followed by discussions among assessors (outlined earlier) lies at the heart of the approach being introduced at Griffith University to assure academic achievement standards. In principle, the basic principles need to apply at three key levels: (a) appraising student responses to a single assessment task, (b) reviewing the evidence from a number of assessment tasks in a particular course in order to ensure that the grade awarded is based on agreed academic standards and so properly represents a student's level of academic achievement, and (c) moderation to achieve comparability of grades across courses. Although details of these extensions latter lie outside the scope of this paper, peer review can replace the review of grade distributions as a means of quality assurance. In all three situations, the assessors operate in close and direct contact with the primary evidence of achievement –

student responses to assessment tasks. During their discussions, assessors scrutinize actual student works as an essential part of the process of coming to consensus decision. Provided the processes required are carried out with integrity (specifically, with full professionalism in the group dynamics, no collusion among assessors, and no procedural or other shortcuts), consensus-seeking moderation involves conversations about particular grading judgments. It has the potential to improve both insight and practice among academics, and lead to a shared vocabulary (particularly the meanings of various criteria and how they are used) as judgments are explained. Ideally, it results in collective understandings about, and a mutual respect for, academic achievement standards. At all levels, engaging in consensus moderation is geared towards improving consistency in marking and grading, and protecting the integrity of the course grades which are recorded on student academic transcripts. The direct professional interactions that are characteristic of the different levels of consensus moderation not only identify common ground for assessment and grading but also are clearly a collegial activity which is consistent with fundamental academic values.

Reference

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, **34**, 807-826.