Consensus moderation for quality assurance of assessment: Overcoming the illusion of consensus.

Dr Duncan D. Nulty
Griffith Institute for Higher Education, Griffith University, Australia.

Note: This paper draws on interim findings of a $240,000 research funded project by Griffith University. Project title: "Developing consensus moderation practices to support comprehensive quality assurance of assessment standards". Ethical approval for this research has been granted under reference number GIH/08/11/HREC.

Theme: Learning
Pedagogic and learning debates about academic standards, practices and literacies in the context of increasing numbers of international, older and part-time students; e-Learning, distance and flexible educational trends and developments

Abstract (150 words)

Research at Griffith University is exploring the use of 'consensus moderation' (Sadler, 2009, 2010, 2011) as a process that can help to ensure consistent and appropriate academic standards when assessing the quality of students' learning outcomes. In principle, consensus moderation achieves agreement among markers about what comprises quality, and about the symbols (marks or grades) that are used to represent judgments about the quality *level* of students' work. The most directly observable result is marking consistency.

Unfortunately, the mere existence of *marking consistency* does not necessarily signify consensus about the judgments of quality of the students' work. Marking consistency can, and often is, achieved in other ways. Consequently, *the illusion of consensus* may be created.

This paper describes different ways in which consensus moderation processes may fail to achieve consensus. Understanding this is necessary for genuine quality assurance to be possible through appropriate policy and practice.

Background

Sadler (2011) described, in broad terms, an approach to establishing and maintaining academic standards, specifically suggesting that this would be advanced by: "… enlarging and repurposing current moderation methods and developing a system based on peer consensus." (p.99). Thus, Sadler proposes the use of 'consensus moderation'. Elsewhere, Nulty (under consideration) has explained that although most academics have a narrow understanding of what this term means (or do not know what this term means), there are many practices already in routine use that contribute to the formulation of consensus among academics, that help to ensure that the standards of judgment applied by them are appropriately 'calibrated'. Realising this is a significant advantage if Sadler's proposals are to be effected in a more systemic and rigorous manner than the *ad hoc* system currently prevailing.

However, along the way, a further problem needs to be addressed: specifically, that even if use of a range of consensus moderation activities can be *systemically* adopted, there remains a risk that these will largely serve to create *the illusion of consensus*. What follows outlines six ways in which this may happen. Importantly, all are problems for the operation of a 'consensus moderation' model, because they all undermine practices that groups will use when seeking consensus.

The illusion of consensus

The first, and most common, way in which an illusion of consensus may be attained is through what I call the '*impost of procedure*'. I will describe this one at some length, partly because it is most familiar, partly because it helps to explain the remaining five.

Currently, it is common practice for students grades to be determined using a procedure like the following. First, academics mark students' work submitted in response to assessment task specifications. These marks are combined (usually by simple addition). Then a grade is recommended (to an assessment board or similar). The grade recommended is based on the percentage total mark obtained, compared to percentage mark ranges associated with grades – so-called 'grade-cut-offs'. For example, the following extract shows that to be awarded the grade of "High Distinction" students need to accumulate 85% or more of the available marks.

High Distinction	85 – 100 %
Distinction	75 – 84%
Credit	65 – 74%
Pass	50 – 64%

Finally, an Assessment Board considers the recommended grades and makes it's final determination. In doing this, it is common for the distribution of grades to be scrutinised, and compared with the distributions from other cognately similar courses. Where inconsistency is found, explanations are sometimes sought (from the course convenor), and adjustments to the grade cut-offs sometimes made.

De-facto this is a norm-referencing activity, that does not involve any direct scrutiny of students' actual work, nor the marking of that work by the academic(s) responsible. The result of this intervention is that an 'aberrant' distribution of grades becomes somewhat more comparable with the distributions found in other courses – and the illusion of consensus in respect of marking standards is achieved. Formally, students grades may then be conferred by a high rank academic of the Faculty such as the faculty Dean.

Adjustments, as described above, are defended on the grounds that there 'must be something wrong' if students whose grades are otherwise consistent from course to course, are recommended grades that markedly differ. It is supposed that the assessment standards used in the offending course(s) must be inappropriate. Often this assertion is combined with the claim that the task specifications themselves meant that the level of difficulty associated with completing the tasks was not equivalent to that found in other courses. Significant variations in the quality of the teaching are not regarded as a viable explanation. It follows that adjusted grades are considered a more accurate representation of the level of academic achievement than the unadjusted grades.

While all this may be true, at no point in the process described is there any direct scrutiny of samples of students work, the marking of that work, the task specifications for that work, the marking criteria, or even the course context as represented formally by the 'course profile' documentation. There is correspondingly also no *comparison* of any of these variables with other courses. Only the recommended distribution of grades is looked at.

Thus, while such actions may be defended, it is possible that the adjusted grades conferred do not reflect the actual learning achievements of students – yet, an illusion of consensus suggests otherwise.

Other ways

First, is '*impost of seniority*'. Imagine a marking team of tutors led by a senior academic. The senior academic may, by impost of seniority, decree that the marking of any of the tutors be adjusted if it does not compare well with his or her marking. Again, no direct comparison of

marking or scrutiny of samples of students work may be involved: simply norm-referenced adjustment.

Second and third, are variations on the first. These are '*impost of authority*' and/or '*impost of expertise*'. These may come to bear in marking teams where the balance of power is less clearly defined than the first example, though, in practice all three may co-occur.

Fourth, is 'agreeing to disagree'. Here multiple markers fail to achieve consensus because they have failed to resolve a difference of opinion about what comprises quality in students work, and/or the relative merit of specific aspects of that work. The result is that each marker applies judgments that are different. Despite this, they may use a common marking guide, and may achieve comparable distributions of marks (and recommended grades). The problem is that while these markers appear to agree, they only agree to disagree: the marks they award are awarded for different reasons.

Fifth is 'conceding to the average'. Here a marking team also does not reach an agreement comprising a shared understanding of what comprises quality in students' work. But rather, they achieve acceptance that some form of compromise or averaging of positions will suffice instead. The problem with this is that the collective position adopted may not represent the views of any of the markers.

Conclusion

This paper has outlined six ways in which 'consensus moderation' processes may fail to achieve consensus, and yet yield data purporting to represent a consensus. For genuine consensus to be reached individuals need to concur, to be in harmony, to think as if of one mind. Achieving this involves people communicating openly together about actual examples of students' work, not in the abstract, and not dealing only with secondary data. It also requires a convergence of opinion, and emergence of shared understanding, not merely, for example acquiescence.

The challenge is therefore in two parts, first getting people together, and second obtaining genuine concurrence. The reality will, inevitably, fall short: genuine difference of opinion in tertiary education is frequently justified, and an overall judgment of students' performance based on the composite of judgments, may therefore represent, after all, the most valid representation of their academic achievement.

Postscript:

It's not enough to be able to herd cats, one also has to get agreement that 8 out of 10 preferred 'Whiskers'.

References

Sadler, D. R. (2009). Moderation, grading and calibration. Edited Keynote Address for the Good Practice in Assessment Symposium. Retrieved from http://www.griffith.edu.au/__data/assets/pdf_file/0017/211940/GPA-Symposium2009-Edited-Keynote-Address-FINAL.pdf

Sadler, D. R. (2010). Assuring Academic Achievement Standards at Griffith University: Discussion Document. Griffith University.

Sadler, D. R. (2011). Academic freedom, achievement standards and professional identity. *Quality in Higher Education, 17*(1), 85-100.