Moderation, grading and calibration

D Royce Sadler, Griffith Institute for Higher Education

This paper is an edited transcript of a keynote address to the Good Practice in Assessment Symposium, 24 November 2009, Griffith University

Ensuring academic standards is high on the list of concerns for Australian universities. Most other countries have some sort of quality assurance system that deals directly with academic standards; Australia does not, except incidentally. Griffith University must, along with other universities in Australia come to terms with the issue of quality assurance specifically as it relates to academic achievement standards.

In broad conceptual terms, the big questions for Griffith University are: Where do we want to be in five years’ time in assuring our academic standards? How do we map a transition from where we are now to where we want to be (because that may involve significant change)? Why should we even attempt it?

Professor Kerrie-Lee Krause has already made a number of references to the context we are in, and I will begin my talk by giving you a few more specifics. When the Australian Universities Quality Agency, that’s AQUA, began its first round of quality audits, it decided to see how universities were performing in relation to the objectives and goals they set for themselves. That was the frame in which AUQA planned and carried out its work. AUQA is a company limited by guarantee, owned by the State and Federal Ministers of Education. The Ministers hand AUQA a brief from time to time. One of the significant changes to the original brief is along the following lines: It is one thing to assess how well universities are performing in relation to the goals and aspirations they set for their own performance, but what about the standards themselves? If universities can set their own goals and standards, how does anybody know whether they’re any good?

This is an important issue Australian universities have to deal with. How is a university to get a handle on it? One problem is that the Australian university sector does not have a quality assurance system (specifically in relation to its academic standards) that provides a level of scrutiny that approaches what applies in some other countries. How can we therefore hold our hands up internationally and say we are as good as anybody else? The fact is, we really don’t know. This makes us potentially vulnerable. We may well be, we may not be. The fact is, we don’t know, we don’t have the hard evidence. AUQA (on behalf of the Ministers) is saying it’s time we did know.

In some other countries, agencies such as AUQA are spearheading direct change. In Hong Kong for example, the Universities Grants Committee (which funds the universities) has determined that it will implement a certain model for assuring academic standards. Hong Kong is in the process of becoming a significant higher education hub not just for Asia.
but the world. You may not realise this, but most of the tuition in the eight Hong Kong universities is in English. One university teaches half in English and half in Chinese, but English is certainly the dominant language of instruction, and it will remain so. Hong Kong is gradually building up its proportion of international students as part of the goal of becoming the major hub in Asia. They’re view is that, if Hong Kong is to be the leading hub, it needs to be able to stand up, look people in the eye and say, ‘Our standards are up there with the best in the world’. The same is happening in Singapore. A few years ago, large numbers of students came to Australia from China, and of course many still do. But a lot of those who would have come to Australia a few years ago are now going to Hong Kong, which is seeing mainland China as a very lucrative market. The Hong Kong University Grants Committee has a whole branch of its operation dedicated to quality assuring their academic standards.

Similarly, the United Kingdom has its Quality Assurance Agency, the QAA, which has been engaged in this kind of exercise since at least 1999. They probably have a substantial lead on Australia, although it has not been exactly smooth sailing. If we now look at groupings of countries instead of individual nations, an OECD project called AHELO is well under way, as Kerrie-Lee has already mentioned. AHELO stands for Assessment of Higher Education Learning Outcomes. It is being driven from OECD headquarters. Some of you will no doubt have heard of the ‘Bologna Process’, which began back around 1960 when four countries said, ‘We want to get our act together to establish, and maintain, academic standards in our higher education institutions.’ The UK was one of those countries. Since then, the Bologna process has morphed somewhat, but has never lost sight of what it wants to achieve. Among other goals, EU countries want to ensure that people in Europe can go from country to country with qualifications that are recognised throughout the EU. This is so that people don’t have to be re-examined and re-certified when they cross a border. Consequently, mutual accreditation of qualifications is really important. The other goal is for students to be able to travel from country to country, study in one or more universities other than their home university, and have their achievements credited back.

One of the factors providing the Bologna Process with renewed impetus is the realisation in EU countries that the number of international students drawn out of Europe to study in the United States has been growing significantly year after year. Apparently, there has been a smaller flow of students from the USA to Europe. EU countries are now saying, ‘We are constantly losing students to the USA. We simply have to compete better.’ That is one of the reasons behind the EU’s recent upsurge of interest specifically in academic standards. Even though the USA had apparently become comfortable with their status as one of the primary destinations for international students and the large numbers they’d been attracting each year, the Bologna initiative has now concerned them. Unless they can compete, they stand to see their international student numbers reduce. You can see why the USA is taking the Bologna process very seriously and working through its implications. Some universities are doing what Melbourne University did – changing the whole structure
of their degree patterns. They are exploring the extent to which they should Bolognaise their curriculum. Underneath is a constant concern: Will our qualifications be recognised internationally? Do our grades and our transcripts stand up against world competition?

There have been a number of solutions proposed, and in my view none of them is perfect. Part of the reason they’re not perfect is universities and quality assurance agencies do not seem to be crystal clear on precisely what they’re trying to achieve. In particular, they are rarely explicit about what is meant by ‘academic standards’. Furthermore, they typically do not spell out the mechanisms, or causal links, between the processes they devise and implement and the desired end result or state. That might sound a very, very obvious thing to fix. However, whenever I ask people, ‘Can you explain to me the nature of the mechanism that connects this process with that outcome?’ I have often received ambiguous replies. I’m not confident what is being done will deliver on the desired outcome. Maybe it is all in too much of a flux at present.

In Australia recently, we’ve had the Bradbury Review of Higher Education which produced a substantial report. Two of the recommendations in that report deal directly with assessment and academic standards. Some months ago, AUQA produced and distributed a discussion paper about academic standards. I was a member of the Advisory Group that produced that paper. It turned out to be difficult even for the Advisory Group to nail down the exact nature of the central problem to be solved. I’ll come to that later. The Bradbury Report not only recommended that something be done about standards but also specified the approach that should be followed. I personally am not confident about the feasibility of the proposed solution strategy and whether it can actually deliver. Partly as a result of the Bradley review, the Government is setting up a new agency called the Tertiary Education Quality and Standards Authority, or TEQSA. Where AUQA will sit in relation to TEQSA is not yet clear. Maybe AUQA or its functions will be absorbed into TEQSA, or maybe it will have some separate role.

Parallel with those developments, the Australian Learning and Teaching Council (ALTC) is also seriously engaging with the standards issue, which is specifically listed in ALTC’s seven ‘Designated responsibilities’. Actually, standards have been on the agenda since the establishment of the ALTC’s forerunner, the Carrick Institute. Two of its responsibilities dealt with academic standards. I was asked about four years ago to produce a paper for the Carrick Board outlining a way to address academic achievement standards. It turned out that the Carrick Board at the time did not want to go as far as assuring achievement standards, and saw their agenda as more focused on standards of, or for, content and structure of qualifications, student entry, resources, teaching spaces, students with disabilities, equity attainment, and teaching quality. I assume they believed that if all the relevant factors were quality assured, the achievement standards could be vouched for. The Carrick Institute avoided a direct approach to assuring academic achievement standards. However, as I read the national and international contexts, academic achievement
standards is now the key issue. It is what worries a lot of people. Do the grades that are on students' transcripts actually mean what they say? The ALTC seems to be more inclined now to look at achievement standards. That is really healthy, and overdue.

Grade inflation in the USA is a significant topic of interest, research and comment. Speeches, academic articles, conferences, books and consultancies sustain a viable trade in determinations of problems and proposed solutions. Even Harvard and Princeton have been in the news, partly because of the proportion of what we’d call high honours that they have been awarding their students. The core question was: Were they deserved? That is more important than the actual numbers they’re giving. Were they deserved? Princeton at one point took a really hard line and decided they would grade strictly on the curve. Predictably, that upset many students and faculty. So it is a big issue in the USA. The fundamental issue is the extent to which grades truly represent the achievements they appear to.

Let me go back a bit, and say something about AUQA and Griffith University. AUQA audited Griffith in 2008. In its report were something like 20 recommendations. One (and only one) of them, Recommendation 7, had the word ‘Urgent’ beside it. The first half of this recommendation said: 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). Assuring academic achievement standards means that we can guarantee course grades, and could produce evidence about their veracity on demand. This is the external imperative that Griffith must respond to, but I want to suggest that there is also an internal imperative. We owe it to ourselves, and we owe it to our students and our future students, to be able to vouch for the quality of our work, and this includes the grades we award. What we are doing now is not best practice; we can do a lot better.

Our aim should be to see how we can do a lot better without doing three times as much work (or even 20% more!). The challenge is this: How do we do better, while doing the same or even less work? We have to be smarter in dealing with the grading issue. This development has to start by being clear on what we are doing. The transition from where we are now to where we’ve got to be is likely to be fairly demanding in terms of our thinking and our time. We should also design and plan so that the long term maintenance of the system we devise is not particularly taxing and is preferably one in which we can all feel some sense of pride.

Although it is easy to say that we should develop a sound and feasible system for doing assessments better, there is a curious anomaly, which I alluded to a few minutes ago. Both nationally and internationally, there’s a strong emphasis on developing a set of processes that should be implemented, yet a surprising lack of clarity about the precise nature of the end to be achieved. (I’ll return to that once again towards the end of my talk.) As to the level of actual teaching which we academics know best, whenever I ask the typical university lecturer, ‘Why exactly do you assess?’ the common reply goes like this, ‘Oh well, I mean, everybody does it. So yes, isn’t that just part of my job – something lecturers have to do?
They plan their courses, structure and deliver their teaching and then assess the students. That’s the basic setup.’ I ask, ‘Yes, but why do assessment?’ Answer: ‘It forces students to learn, and the grade is then put on the academic transcript.’ Suppose I probe a bit deeper: ‘What do the grades stand for? What are they supposed to represent?’ For me, the answer is simple: Each grade is intended to represent, in symbolic form, a student’s level of attainment in each course at the point of completing the course. It doesn’t matter whether the course is taught as an intensive or over a full semester, one year, or even longer. It doesn’t matter how the course is delivered, on campus or off campus, technology assisted or not. And it doesn’t matter which set of symbols is used. At Griffith University, we certify achievement levels by the terms High Distinction, Distinction, Credit, Pass and Fail. Other universities use A, B, C,…, or 7, 6, 5…. It goes without saying that such ‘grades’ recorded on student transcripts should mean what they say. The issue for us (as academics, schools, Groups and the University as a whole) is: How can we vouch for the validity of the grades? Do all grades awarded align closely with students’ actual achievement levels?

What we have been doing at Griffith University for many years actually falls short of delivering (quality) assured course grades. (Some countries use the term ‘secured’ rather than assured.) Current practice does provide some checks and balances, but only in a limited sort of way. We need to, and can, do better. That’s what I intend to focus on in this talk: what assuring academic achievement standards consists of, and how we can go about assuring them. By the way, whenever I say ‘standards’ in this talk, I will be meaning academic achievement standards. I don’t mean teaching standards, resource standards, or standards for student entry. There may be one or two points where I use ‘standards’ with a slightly different meaning, but the context will make that clear.

For many people, the way ahead is not self evident and the various reviews and literature don’t generally have satisfactory answers either. It’s still a developing area, but one that I have been working on myself for quite a few years. In your pack you should have a copy of my article called Grade integrity and the representation of academic achievement. [Added for readers not present at the Symposium: This document is available on request.] This is my best shot (so far) in spelling out what it is we’re on about. It’s a fairly long paper, but it distils my thinking as it’s gone on over the years. It narrows down what I believe to be the nub of the problem, and sets out a way forward. (I have to warn you that the article is a denser read than the transcript of this talk will be, but it’s where I’ve put down formally what I understand to be the big picture.)

But before launching into the deep specifically on grades, let me take a few moments to outline something fairly new on the horizon – Australian Higher Education Graduation Statements. Think of these as compilations of information that are complementary to official academic transcripts, substantially constructed by the students, and as records of their achievements in a range of fields, not just their academic work. In other places, these statements go by other names: Records of Achievement, Graduation Portfolios, Diploma
Supplements and Progression Files. The people behind the Bologna Process are strongly promoting the idea. They want employers to be able to look at comprehensive documentation of each student’s academic and life experience, interpret it, and get to know a lot more about their job applicants. Whether universities should take responsibility for the folios of information is a debated point. Regardless of whether they take off, I am quite sure that the need for the academic transcript, which is the only formal certification apart from the degree testamur itself, will never go away. It will always be a central part.

There’s also no chance that our idea of ‘courses’ will go away either. Courses are the chunks that make up degree programs. Also called modules, units, subjects or papers – depending on which country you’re from – they’re here to stay. In one or two countries still, students don’t sit for any assessments that are recorded until the end of their degrees. Then they do their ‘comprehensives’ or ‘finals’, which could be spread over three weeks or so. When they pass their finals, their degrees are classified and awarded. That is becoming more and more unusual in today’s world. The USA started off the idea of courses that are mostly one semester long. We in Australia started adopting the idea back in the 1980s and it’s pretty well universal in Australia now. In the UK they’re busy ‘modularising’ their degree programs. That trend is progressing in Europe as well, partly because they are trying to foster mobility of students throughout the EU. Clearly, two things look certain to remain fixtures. One is the transcript and one is the idea of courses. Now let’s get back to course grades.

Our obligation is to make sure the grades that go on an academic transcript are as true to label as we can make them, so this is a matter of truth-in-labelling as much as anything. What do I think of this idea of comprehensive supplementary documentation? In some of my previous roles, I’ve been on the receiving end of multiple applications for jobs. One we advertised had 60 applicants. I don’t know how I would process 60 voluminous folios, or what I’d make of the information in them. I suspect what I’d do is dig out the formal transcripts, and use those to do a quick sorting first. The first pass might bring the number down to 8 or 10. I’d scrutinise them and then short list, probably to about half. Do you see what I mean? The sheer mechanics of dealing with literally a carton load of folios is overwhelming. On the other hand if you only had one applicant you could well be very interested in looking carefully at the full works. I am definitely not saying they’re a useless idea; I’m saying the practicality of it might turn out to be less than a lot of people hope for.

I need to make it clear that from now on I’ll be talking about summative assessments only, not formative assessments. I have strong interests in both areas, but I believe that we can make summative assessments work well for students. I’m very interested in that as a side bar.

There are three propositions that I want to champion today. The first one is this: Grades should correspond strictly with the actual achievement in all courses. That is, there’s no systematic grade inflation, and no grade deflation. Furthermore, there is no cavalier
marking either. Each grade should strictly represent the level of achievement of a student in a course. If that sounds dead simple to you it’s not in practice, and I’ll indicate a little later why that is so.

The second proposition is that grades should indicate the level of an individual learner’s performance in absolute terms, not the person’s performance relative to other members of the class. That’s been Griffith’s policy since Griffith was founded in 1971. Anybody been here since this University was founded? No, I haven’t either. All of us have come later, but that principle has been here at Griffith University since the very beginning, as Griffith’s policy. We want to make grades represent exactly what they say they represent. If we could do this successfully, it would make a student’s rank (in comparison with others in the cohort) irrelevant.

The third proposition is that grades in different courses should be comparable. It turns out that comparability is a really interesting topic. What does comparability mean? Does it mean that comparable things have to be the same? The answer is no. We’re not talking about uniformity or identicality. We’re talking about comparability. Things can be comparable but different. When they’re ‘comparable’ they’re comparable in relation to some criterion. We make comparability judgements all the time. Let me give you a quick example. When people are buying houses they go to look in the area in which they’d like to live and see what’s around within their budget. Suppose you are a buyer. You might find one house with a large imported Italian fountain and a beautiful established garden, and be really taken by the place. You think it is absolutely wonderful. You look at another house. It doesn’t have the fountain but it sure has a great view. You can see Mt Coo-tha and contemplate the TV towers, if that’s your thing. And another house you look at is on a larger piece of land which is less accessible, but there’s a permanent spring on it, just perfect for feeding into a mini-lake.

These three properties are quite different from one another. Let’s suppose they all have the floor space and facilities you need, and carry about the same asking price. Furthermore, you regard each as representing fair value and you could afford any one of them. In those senses they are comparable properties. That said, your decision comes down to personal preferences and priorities. It turns out that, for you the buyer, they are not comparable with respect to amenity value. You are completely taken by the fountain – especially because you didn’t expect to find it. You place such a high personal ‘value’ on it that it tips the balance for you. We make comparability judgements all the time; we compare things that are not identical by somehow balancing off the various aspects. In the example, they’re all dwellings and they’re all on pieces of land in the right area and at affordable prices. But they’re all different.

If you’ve got children, let me ask if you treat your children exactly the same way? Often not, I suspect. My sons received different presents at Christmas, nearly every year. The main aim for my family has always been to treat them all fairly, in ways that each would also
regard as fair. This takes into account who they are, what their interests and needs are at the time. What we as parents tried to do was to always be fair. I found the same when I was working as a Dean, to decide according to what’s fair. Here’s an example. We had three Heads of Schools in the Faculty, and after years of special pleading, instituted a flat funding arrangement which was worked out according to student load. The deal was that any Head who believed a case for special funding existed for their School could present their case at a meeting of the three Heads and me as Dean. As a group, we committed ourselves to being fair to one another, and acknowledged that our judgements would be open to valid argument and refinement. We undertook to objectively analyse any special-case argument and then decide what would be fair to the School concerned, and also best for the Faculty as a whole. We did not want any School to be unable to do its best work within the total resources available. Fairness was to be the overriding consideration. Flat funding would apply unless there was clear evidence to the contrary as to why that would be inappropriate. Strangely enough, there were no takers, and the Schools (which were quite different in their academic responsibilities and student loads) decided to live within the flat funding model. The issue we are facing in this talk is: How is it possible to think about the grades awarded in different courses being comparable? I agree that there are practical limitations to how far we can go with this concept, but I want you to bear the idea in mind, because it is crucial.

Let me run through those three propositions again. The first is that grades awarded should correspond strictly with the actual level of student achievement, and not be influenced by non-achievement variables such as effort, persistence, or participation. The second is that grades should indicate the level of an individual learner’s performance in absolute terms, not in comparison with the performances of other members of the class. The third proposition is that grades in different courses should be comparable. Mostly you won’t find the desired end point of assessment and grading practices articulated in these terms, but it’s the way I believe we as a University should move. It is my attempt at clarifying our end goal, which I believe is consistent with the grading principle specified when Griffith University was founded.

How to bring this about in practice is a separate consideration. It must be not only conceptually sound but also practical and efficient to implement. We can’t just dream up some grand plan that will be hugely costly to build and maintain. That’s ridiculous – no one’s going to do it. If we can be committed to the principles I’ve just outlined, we can make considerable progress. To some extent it is uncharted territory. We might find ourselves up a few blind alleys along the way, and we will no doubt make a few mistakes. But we can work at our own pace so we can change the way we think about and practice assessment and grading.

Let me outline briefly the system Griffith has in place now. What I’ll say is not at all a criticism of Griffith over the years; the same approach applies to most other universities in
Australia. Basically, this is what we do. In single-lecturer courses, markers go ahead and mark student work; they enter the marks in the spread sheet; and then they award the grades to their students. Actually, that’s not quite correct. Academics don’t award final grades in a course, they propose grades. We are advised not to tell the students their grades until they have been approved by the relevant Assessment Board. (I know there are both Assessment Panels and Assessment Boards, but I will use ‘Boards’ to cover both). Proposed grades go to Assessment Boards, which look them over and in most cases approve them. However, for some sets of proposed grades, a Board says, ‘Wait a minute. These look a bit whacky. How could you have a class of 57 students with 56 of them getting High Distinctions and the other one getting a Distinction? That seems extraordinary.’ The Board is charged with the responsibility of doing something about it. Note that all they have looked at, typically, is the distribution of proposed grades. Under the Rules, a Board must consult with the convenor to get advice and if necessary, take steps to sort it out. Most Boards would want the grades to be distributed without such high proportions at the top end. In other words, the Board feels that the first proposition (grades must match achievement) is really unlikely to be true. That situation applies for results that are top heavy. I’ve been in Academic Committee meetings where the bottom end has been the concern, and the subject of long discussion. Here’s an example: For some courses in a particular Faculty, the failure rates might have come in at over 40%. That was regarded as an unacceptable failure rate. So what’s going on? Maybe these students have been graded too harshly. This two-stage process of proposing grades and then scrutinising them at a higher level is what’s used in most other universities that I’ve been able to talk to. Many have it written in their policy.

Let me just make a few more comments about that. The Assessment Board proposes changes, or pushes for changes, if it seems that a distribution of grades is outside tolerable limits. The University doesn’t set tolerable limits, so the Boards decide what they think is too aberrant to let go. Generally in such a situation, a Board trusts the rank order of students but doesn’t trust the proposed grade allocation. Having been a Chair of an Assessment Board I know this is what we did. I have talked to quite a few other Chairs of Assessment Boards, and most say they do the same. It’s has not been really clear to date how we could move on from there.

In reality, many academics, most I suppose, make sure that their grades are going to comply with what’s tolerable. Otherwise they’ll get the call, ‘Excuse me, I think we need to talk about this.’ The safe path is to make them comply. The Assessment Board may say, ‘Oh a bit high, (or a bit low),’ but then let them go through because otherwise they’d be there forever making fine adjustments. There are certain deadlines that have to be respected too. The practice of looking at the distribution list and pushing it to comply is sometimes called ‘grading on the curve’. However, it’s not really a ‘normal’ curve that’s important, or any other curve shape for that matter. The real issue is the relative proportions of the grades and the apparent need to control them. Where does that thinking come from? Well, it
comes from market thinking. If we have some valuable commodity that is in constant demand and we restrict its supply, generally speaking that will make it hold its value. That market mentality means that if we give everybody a High Distinction, we devalue that grade level. If we fail nearly everybody, a pass takes on great value. We actually control the value that’s placed on it by controlling the proportion awarded. It’s a very simple logic.

That type of method does work, at least superficially. It is always easy to do, and it always delivers the goods. By making everybody conform, with everybody thinking the grades are comparable, by definition. Controlling proportions puts the marks from the tough and easy markers onto the same scale. It gives grades similar distribution and therefore similar ‘values’. Of course, none of this could happen unless all parts of the University use the same scale (from high distinction to fail) and the same nomenclature. The joint practices of controlling grade proportions and using a common scale seem to force comparability of sorts, so we all sit down and say, ‘Job done’. We think no more about it until next semester.

The two fundamental requirements are these: protecting the value of grades, and using a common scale. If we can do the first in a logically sound way, the problem would be substantially (but not completely) solved. Before explaining an alternative, I want to indicate why the proportional control approach is thoroughly unsound. Why are those procedures questionable? There are seven shortcomings I can identify.

1. The Assessment Boards always operate with secondary data. They might have the marks spreadsheets in front of them and they usually have the grade distributions. They all consist of secondary data. There is no hard evidence of the achievements of students which can be discussed. All the Boards have are records of people’s judgements about those achievements, which is why it’s always secondary data. If an Assessment Board were to say, ‘Let us call for the actual data’, sometimes they get the evidence they need, directly. I had to do that once as Dean. I recalled all the student work for a whole cohort of students in a course and I had them marked independently. This resulted from an allegation that one lecturer was giving all of his students high distinctions and distinctions because he didn’t want to mark the work seriously, and he also didn’t want any students to complain about his teaching or the marking. Recalling the work was the only way I could see to get hard evidence to evaluate the allegation. The reason the allegation was made was that some students who were really excellent students felt that their own grades had been devalued by giving everyone high grades. They said, ‘This is not fair to us. These other students are getting an undeserved bonus.’ The story short is that the academic in question left of his own volition within a year. I was greatly relieved because I don’t know how I would’ve handled any discipline problems. The lecturer concerned felt discovered, and that was enough to shame the person into going. It was a very difficult exercise to do but it showed that manipulation had occurred. I think of him as a rogue grader,
meaning he did not do what he was supposed to do, despite the fact that the University had trusted him to do it. It was interesting that the problem came to light through a considerable number of students expressing extreme dissatisfaction, even though the results had been ratified by the Assessment Board.

2. Institutional grading policies often allow for aberrant grade distributions to be approved if there’s a strong case for them. Here’s my comment on this one. Academics aren’t bad at making strong cases. Half of them construct arguments for a living and can be pretty persuasive at convincing an Assessment Board, often in the absence of hard data. It can be pretty hard for an Assessment Board to gainsay what they are told, because there is no simple way of investigating unless work that had given back to the student is recalled.

3. Controlling proportions produce grades that reflect performance relative to that of the class cohort. This can be unfair and unethical. Students have no control over who else is enrolled in the same course. Their grades depend on how other people perform, not solely on their own academic performance.

4. Relative grading is blind to actual achievement in any absolute sense. It can’t detect, or actively conceals, these aspects: the quality of course assessment plans and assessment tasks (there are some woeful tasks out there); the quality of teaching, learning and assessment; and the types of components that are built into the grade (this comes back to what constitutes achievement). I’ll just make a further little point on the last one of these. For me, attendance is not achievement. I say that should not count for anything in any course grade. If you want your students to attend all the time, make your teaching interesting, make attendance a condition for getting a grade at all, or give them all a pizza or ice-cream voucher. Don’t give marks for anything other than student achievement, or you will simply inflate grades for no good reason. There are a whole lot of things other than attendance to which the same principle applies. Overall, when you control grade proportions, you don’t have any idea what’s been included in the grading process unless you explicitly dig for it.

5. Relative grading resets the grading parameters every time the course is taught, because you have different groups of students. You reset, you reset, you reset. It could be that your actual grading standards are going down because the students you’re getting are not as good as they used to be, or there’s a flood of them and there used to be small numbers. All sorts of things, like the actual levels of achievement may be going down on the average, but the grade distribution stays the same. Controlling proportions can’t pick up any long term drift in actual achievements.

6. Course convenors can blame the Assessment Board for any changes, and say to students, ‘Well, we don’t know what happens up there at the Board. You know that we do the right by you students. We give you a high mark but then the Assessment
Board knocks them back to suit their own agenda. It’s all very arbitrary, but we lecturers are powerless. ‘Am I exaggerating?

7. Controlling proportions is contrary to Griffith’s policy. Historically, the Assessment Policy has always stated that relative grading is not appropriate, but I don’t know of any Assessment Board that adheres to this. We should either change the Assessment Policy and take it out, or acknowledge that it exists as a last resort strategy. Today, I want to propose a different way of assuring academic achievement standards.

An early step would be to eliminate non-achievements from being part of summative assessment. It’s become accepted practice across Australia, and in the USA and the UK, to include a great range of non-achievements as components of course grades. (I’ve written a whole paper on that if you’re interested in following that up.) [Added: This is the Fidelity article listed below. It is a companion paper to the Grade Integrity article, and is available on request.] More and more non-achievements have been added to grades over the past 35 years, and the practice is now part of the normal assessment culture. Those things would’ve been thought inadmissible in the past, but somehow there’s been an incremental creep to embrace them. I remember when I was Chairing an Assessment Board there was a lecturer who proposed high grades for all her students. When we chatted with her, in a non-confrontationist way, we found that according to the course assessment plan students could collect marks for completing a series of tutorial exercises that required producing practical things that ‘worked’. If they worked, you got the marks. Students who completed all their tutorial sheets got the marks. The quality of their work was hardly an issue. We found that when you added all those things that students did and got credit for, she’d created a platform of 40% of the total marks. Student’s marks for real achievements were built on top of that platform. Consequently, most students got high marks. This brings us right back to assessment planning. We need to look at it closely.

If we could agree to completely disallow non-achievements, we could start on the real task. The next steps are to focus on awarding grades that are strictly commensurate with actual levels of achievement, work towards comparable grades across courses and programs, and fix our standards so they are stable over time. Here’s a summary of the approach I think we should take. It involves identifying standards for student work. How do we set academic achievement standards intelligently? What would it mean to create standards that are somehow comparable across courses? I’ll rephrase that last bit this way: ‘comparable across cognate courses’, that is, courses with some substantive similarity to them. How can we make the standards we apply visible to our students? (I say this because many students do not know what high quality work looks like. They really don’t!)

Two weeks ago I made a guest appearance for a colleague in a business course on just this issue. The course lecturer asked me if I’d illustrate my approach to making standards visible in the session, because she was unsure how to go about it. Let me tell you briefly what I did. We had asked all of the students to bring one of their earlier pieces of work to
the evening session, without any comments on it. These scripts were de-identified, and each student got to read four or five of their student colleagues’ works in the period we had available. I asked them to answer the following four questions about each work they read:

1. Is the assignment on target? Does it address the assessment task as specified?
2. How good is the assignment? Where would you rate it on a 1-10 scale?
3. How do you justify your assessment of its quality? Explain, in your own words as best you can, why you rated it as you did; and
4. What would you advise your student colleague to do to make it better?

What happened? As students looked closely at other students’ works, they got an insight into what we see all the time. When we mark the work of, say, 60 students, we get a constant refreshment of what we think quality consists of, in very concrete terms. We also have confirmed for us that for any particular quality there is a range of ways of achieving that level of quality. They’re not always the same. I call each of those an ‘expression’ of that quality level. Observing these two aspects all the time is so much part of what we do that we’re not even aware of it, but that’s what happens. When do students get this opportunity? They usually have just one piece of work – their own – with a bit of feedback if they’re lucky. Just based on that, how could they possibly experience the rich access we enjoy about what quality looks like? The answer is, they really can’t. So when we distributed student works around, some students were just gobsmacked. Those that had done fairly poor work saw some good stuff, really good stuff. Those that had done good stuff got a surprise to see how poor some of the other submissions were. And those that did mediocre work saw other mediocre work (and the variety of ways to do mediocre stuff) along with some good and some weak stuff. Everybody learned something.

A little while after that, the lecturer talked to the students and asked how they thought the session went. One of them said ‘The greatest thing that happened to me was when I saw other people’s work and what high quality work actually looked like.’ One student even got a bit emotional as he said ‘Why didn’t we know this before? I could have done better in all my courses if I had known this.’ The point I want to make is this: giving students access to the concepts of quality and standards is something we can do in our courses without enormous trouble. We can do it through our tutorials if we have them. We can do it through our lab sessions if we have them. We can re-structure our current time and activities to give our students access to the standards we use for marking and grading. That’s critically important. We owe it to them. The ideal is that we apply them consistently in our grading, and check periodically that the quality assurance system that we have is functioning well. Every quality assurance system needs to have quality assurance of the quality assurance.

Here’s another curious thing. In the various higher education reviews here and overseas, the scope of what is meant by ‘academic standards’ isn’t usually stated nor is their essential purpose. Who’s in favour of good academic standards? Everybody is. When I ask, as I have a habit of doing, ‘What exactly do you mean by the term ‘academic standards’?’ I
have found it difficult to get nice, crisp answers. Yet it’s supposed to be what we’re trying to do! Some agencies are responsible for curriculum standards, qualifications frameworks and the like. Different accrediting agencies (such as those for medicine, engineering, physiotherapy and so on) have minimum requirements for degree content and structures. Even they often (but not universally) pass over the critical area of academic achievement standards. These are about what students actually achieve. The term that I use (and have used consistently since 1987 when I published my first article on standards) is standards-referenced assessment. What we need to is find ways of ‘capturing’ the standards we want to use, so we can compare students’ work with those standards. Setting them appropriately is not a simple task, but using them once they are set should be relatively simple. Suppose we set them appropriately – not too high, not too low. After doing that, the absolute numbers of students, and the proportions of students, who are awarded grades at the various levels becomes irrelevant. What this process means is that each grade represents a particular level of competence, knowledge or skill. That is the crux of the matter.

How is it possible to specify standards? Let me first talk to you about standards that can be codified. By codified I mean expressed in words, however you structure the words. A lot of work has been done on discipline-based descriptors. I need to explain why word-based processes cannot deliver. This is not an empirical matter, it’s a logical matter. For this I now reach over for one of my props, these spectacles. There we are – you actually do look a lot better through these! These are not sunglasses by the way, or at least, not just sunnies. You can’t see from where you are, but embossed on the side of these specs is this code: AS/NZS1337. That means that these are safety specs. Over here I have the official standards for these safety specs. There are 43 pages in the description of the standards.

Any manufacturer who wants to make specs to satisfy this Australian standard has to conform to what’s in this set of codifications. It doesn’t mean that every pair of specs is tested. What it means is that this design produced by this manufacturer reliably produces safety specs which meet or exceed the minimum requirements set out in here. Standards Australia has developed hundreds and hundreds of standards covering all sorts of products and processes, a lot of them to do with safety, health, component compatibility and product performance. Any manufacturer who wants to make safety specs would be a fool not to get hold of these standards and test their stuff out thoroughly before submitting a random sample for official testing.

As I read through these, one thing came out as startlingly apparent: every requirement refers to something that is standardised already, or if it’s not standardised it tells you how to create the test rig so that the object can be tested. Take impact resistance, for instance. For this, you need a particular testing apparatus, but you may not be able to get one off the shelf. So you get a lathe worker to make an object of a particular material and with clearly specified dimensions. You also need a special tube, also fully specified. You drop the object
down the tube from a specified height on to the safety lens. If it penetrates less than a specified (accurately measurable) amount, the lens satisfies the penetration standard. Everything in this 43 page document refers to standardised objects, commodities or conditions like these.

In our context, where we are concerned with academic achievement standards, we can write all the specifications we like but unfortunately none of the things the words refer to will be fully standardised objects, commodities or conditions. In other words, when we talk about academic achievement standards we do not have these fixed points. There’s a whole body of research on the limitations of codified knowledge, but it is constantly overlooked by education systems, primary, secondary and tertiary. The standards we need are abstract, and cannot be specified in words or rubrics. It is logically impossible.

So how can we do it? Well we have to work with fixed objects that match the words. The most accessible source of objects is the body of real student works. For particular ones of those, we write words so that each description matches closely a particular object. The descriptions draw attention to the aspects that have contributed to the qualitative judgement, but the actual standards they exemplify remain as abstract entities, mostly in our heads. Our challenge is to find workable ways to come to a consensus on them, without tying ourselves in knots. That’s the basics of it.

A few weeks ago I had a look at a set of grade descriptions from different universities. I picked out some of the hedge words and qualifiers contained in their ‘definitions’ of standards. Here are some of them: excellent, strong, thorough, highly, deep, inaccurate, shoddy, scant, some, considerable, muddled, barely and a great deal. Would we agree on all of those? How would we know? Something muddled for me might not be so to Lynda or Ian or Ray or Kerrie. Something that seems excellent to me might signify to other competent colleagues that my horizons are pretty low. Frankly, we don’t know much about what those words mean. They are not sufficient to ‘carry’ the standards. Words can and often do mean different things amongst lecturers. We need a shared concrete reality to give them substance. The simplest such basis consists of student works.

One of the key ideas to talk about today is moderation. It’s something that we probably all have had at least some experience with, and it provides the key for getting a grip on achievement standards. If we don’t practise marking moderation in our normal academic work, at least we probably know colleagues who do and we understand the basic idea. For our grades to be comparable we have to get beyond any idea of ‘my standards’. What are my standards? Think through this with me. When we approach a marking exercise we already have some ideas of what we think high quality is. That gets given existential form by the work we scrutinise. We look at it and may recognise (and even say to ourselves), ‘This is really good. In fact, this is so good, it’s better than I could do’. Have you ever had that experience? It’s great fun isn’t it?
When we recognise quality, it’s not if we have a prior formula and apply it, maybe over and over. For many of the works students produce, we make qualitative judgements. We don’t follow some rules or specifications, tote it all up, and use that to ‘discover’ the work’s quality. That holds in many professions and disciplines, but there are some exceptions. I’ll give you a quick example. In some parts of numerical mathematics, you may set an assessment task for students requiring them to create an algorithm for solving a particular class of problem. You also require them to write the code for a computer program that will implement the algorithm and do the calculations. In this situation, it may be difficult to make an overall judgement about the solution strategy through trying to ‘recognise’ its overall quality. You may in such cases ‘define’ a quality computer program as one which can actually solve the problems (it is effective), is also efficient, and robust. For the purpose of illustration, robust means that the program can reliably deliver correct solutions not only for straightforward problems but also for ones that are known to be so tricky as to be almost intractable. We then define quality in terms of effectiveness, efficiency and robustness. You can have only a weak idea of the quality of the computer program until you measure those three properties and combine them together.

I won’t labour these types of situations. Our interest here is in the many, many circumstances in which we can (and regularly do) recognise quality when we see it, before we define it, and before we frame our justification for the qualitative judgement. (I wouldn’t go so far as to say that this characterises the majority of assessment judgements, because I don’t really know, but these conditions hold for a great many academics from a wide variety of fields.) We have to find ways of sharing our standards so they become interpersonal and not private standards. The fundamental processes of consensus moderation can be reconfigured and repurposed to serve a bigger end. It is fortunate that these basic principles are ones we generally feel comfortable with already.

I want to introduce what I will call ‘Level 1 moderation’, which is the one we are probably most familiar with. At this level we’re typically talking about a single course, a single assessment task, a large number of student responses, and multiple markers. We want to bring about consistency among markers so that the mark a particular student receives does not depend too much on which marker assesses the work. There are several variations on method, but a common one goes like this. The course convener selects a sample of student works; all the markers mark them independently; they all meet together to see how their marks compare; and they discuss the reasons for their marking decisions, backwards and forwards until the marking standards are clarified. The goal is to tune each marker against all the others, so they all sing the same song in the same key. That is called consensus moderation. After that, the markers mark their shares of the balance of the works, often without further tuning but with provision for discussion about any particularly difficult cases. There may be only a handful of them. I used the word ‘tuning’ just a moment ago, but I now suggest we change that to calibration.
It is possible to identify two cognitive processes that are occurring while markers are engaged in the calibration process. These happen so naturally that participants aren’t necessarily focally aware of them as separate processes. First, the markers come to agreement on what quality consists of. Second, they come to agreement on mark allocations. Putting that another way, they come to agreement on the correspondence between the quality of a particular performance and the mark to be allocated. They agree on how the quality continuum should be chopped up for the marks. The reason for breaking the overall process into two aspects is to develop a basis for generalisation to other levels. In Level 1 consensus moderation, the ‘object’ of scrutiny consists of the student responses to a single assessment task or item, and the underlying variable of interest is their quality. The markers’ agreements are thereby limited to that domain. Before moving on, note that consensus moderation is a form of peer review, which we regard as a legitimate, necessary and valuable academic activity. We practise it in the contexts of research grant applications, manuscripts sent to journals for publication, and the collaborative and collegial work we do as committee members. We regard this as normal.

Let’s now escalate the process to what I will call Level 2. The object of scrutiny changes. Instead of student responses to a single assessment task, it expands to include all student works that provide the evidence for the (summative) course grade, again for a sample of students in the course. This time the ‘object’ consists of all the works completed by each student for different tasks, and the underlying variable of interest (the one to be moderated) is the level of achievement which the evidence signals for each student, and the grade that corresponds to that achievement: high distinction, distinction, credit, pass or fail. The quality of student responses to all summative assessment tasks is taken on board. The same basic principle of consensus moderation is applied, but the scope of the works is broadened. This takes us into some new territory, but should not be particularly hard to do. Again, the aim is for the lecturers to calibrate their judgements.

What if there is only one lecturer-marker in a course? That would require colleagues in cognate courses to participate, and that points the way to the Level 3, which covers comparability across cognate courses. You might immediately say, ‘I know nothing about other courses.’ I suggest that you probably do. Have you only taught one course in your whole life? No. Have you ever been given a new course to teach? Yes. Have you taught a number of new courses in your career? Yes. So where did you get your ideas of standards from so you could begin to mark student work? You carried them with you as generalised abstractions into the new course. They were then refined in practice. Lecturers who contribute to the same program are likely to find there are two or more colleagues able to discuss the standards of student work. We don’t start with a blank sheet. There are other aspects which support this view, but time is short. When I talk about ‘cognate’ courses, I mean courses that have some kind of overlap, or follow on, or are substantively near one another. In your pack for this session you’ll find the diagram [reproduced below] to illustrate
these relations. When academics get together around cognate courses, they can explore how achievement manifests itself in different courses. They can have sensible conversations.

This idea of calibration is quite an important one. Most of you probably realise that most measuring devices have to be calibrated, many of them quite regularly. What I am holding here happens to be a 2 ounce weight. (I collect weights for fun.) If you could look closely at it, you would see a little insert at the bottom into which a crown has been stamped to show that this particular weight has been calibrated and approved for use. Around the button are stamped about 20 consecutive dates. Each year during the currency of this weight, a person from the government’s weights and measures division would come to the user’s shop to check that it still weighed two ounces, and that the owner hadn’t drilled a hole in the bottom. If it passed, it would be date stamped for that year, and so it went on. This happens to be a pre-decimal weight, which explains the ounces. The date sequence stopped when Australia went metric back in the 60s. The point is that it was regularly calibrated. These days, digital scales at the supermarkets and fuel pumps at the service station are regularly checked and certified. Between checks, it is assumed that honest operators will be measuring accurately. Much the same philosophy can carry over into our grading contexts. If we can calibrate ourselves against competent peers, and the system works openly and with integrity, it would not be necessary to go into the whole exercise of moderating all student works every semester, because we as markers are the instruments being calibrated.

That will have to do, I am afraid. I ask you simply to take this sketchy outline on board for the present, with one or two final observations. Working towards consensus moderation at different levels focuses specifically on the sites where grading decisions are made, and on the academics who make them, using real evidence. This is a far cry from massaging grade distributions into shape in an effort to control the proportions of different grades, and so achieve a superficial form of comparability. Further developing and refining these processes is new territory for us all. We should not be afraid of it. We can do a lot more than we think. We’ve all already done similar things in other contexts. We’ve brought standards from other places and times and applied them to new courses. The idea of setting standards, sharing standards, holding to them, and facilitating student access to them can be a collegial exercise, not one that’s confrontationist and likely to embarrass. We need to take
possession of the agenda. We have an opportunity to engage with these issues in ways that are congenial, collaborative, and compatible with academic values. If we ignore the opportunity, we may find that a system is imposed on us by an external agency which we find alien to our fundamental values and labour intensive to operate. I leave that thought with you.

Thank you all so much for listening.

References
Sadler, D. R. (Online 2009). Fidelity as a precondition for integrity in grading academic achievement. Assessment and Evaluation in Higher Education. DOI: 10.1080/02602930902977756
URL: http://dx.doi.org/10.1080/02602930902977756