International Journal for the Scholarship of Teaching and Learning
http://www.georgiasouthern.edu/ijsotl
Vol. 5, No. 1 (January 2011)
ISSN 1931-4744 @ Georgia Southern University

The Unscholarly Use of Numbers in Our Assessment Practices: What Will Make Us Change?

Chris Rust
Oxford Brookes University
Oxford, England, UK
crust@brookes.ac.uk

Introduction

On being invited to write something "informative or evocative" I really couldn't resist the temptation to be a little provocative and return to a theme that I have touched on before (Rust, 2007), and is a particular *bête noir* of mine. Although I have to admit that, given I have no reason to believe that what I wrote previously has had any discernable impact, there is a slight sense of futility even as I start. In fact there is an obvious irony in spending time re-stating information about how stated information has been consistently ignored for years. Is presenting an argument for change in our practices, however logical and well supported by evidence, simply a futile act? And if so, what can bring about change? But that is perhaps another paper.

The *bête noir* to which I refer is the fact that clever, intelligent people – namely academics – can behave in such 'unclever' and unintelligent ways when it comes to numbers. Now I realise that we live in a worryingly innumerate world. Just one illustration of that is the fact that in the UK it seems that every six months or so there is a headline in the press that effectively says, "Shock! Horror! Report shows half our schools are below average!" And there is then a queue of politicians who line up to condemn this fact and argue for what should be done to change this situation. But none of them say, "What's the problem? That's as it should be – it's math!" So maybe I shouldn't be too hard on my fellow academics, but I just feel the faculty ought to be that little bit more numerate than society at large. But in fact, there seems to be a number blindness, a number ignorance in the academic community too, where the use of numbers actually seems to stop them thinking.

I refer to the ubiquitous use of numbers in the assessment of students' work. It does not seem to matter where you look around the world, academics seem to be entrenched in systems that (mis)use numbers to judge and record students' assessment.

Background

So what is wrong with linking numbers to academic judgements, I hear you ask? It has always been thus; these are systems that have stood the test of time, colleagues say; 99% of academics can't be wrong – or can they?

Well I believe they are very wrong, and for a number of reasons. Firstly, many of our assessment practices do things with numbers that are simply bad practice. And the incorrect ways in which numbers are managed are not even particularly complex: a first year statistics student would be failed for doing with numbers what happens in most of our assessment systems. Secondly, as well as being wrong, these grading practices actually block teachers from evaluating their students' learning; grading students using numbers

obscures what has been learnt, from both the student and faculty. Thirdly, there is good evidence that the use of numbers in assessment can skew learning in negative ways.

They are big claims, but they are not mine alone; plenty of other writers and researchers in this field have provided evidence and argument to support them. The galling thing for me is that despite this the use of numbers in our current assessment practices remains both unhelpful and unscholarly. They are the three major arguments behind why I believe we should stop doing a number of things; let us consider what those practices are, and the weight of evidence in more detail.

We Need to Stop Using Percentages Because They Do Not Tell Us Anything Useful

A lot of assessment is scored as a percentage mark. This has at least two major flaws. Firstly, if you scratch the surface, what does a score of 55%, for example (an average to low mark in the UK), actually mean? Two students can have the same score while having very different strengths and weaknesses. The number ignores and obscures this detail.

Secondly, when it comes to the marker's judgment, the fact that they chose not to give the work 54% would appear to mean that the marker can distinguish the quality of the work to a precision of one percentage point - a distinction of one hundreth. Actually, in the UK and some other countries it doesn't because in arts and humanities subjects at least, the full range is unlikely to be used.[1] The range is more likely to be between 35%-75%, but that still requires judgments with a precision to one fourtieth of difference. This level of fine-grained judgement would be impossible for just one criterion but these judgments are likely to be working with multiple criteria (see Elander & Hardman, 2002). In the US, I have been told that, in practice, it is now rare in many courses for a mark below 70 to be awarded, so that would be a 30 point scale, but the issues are the same.

Aggregation of criteria is a further problem. If one criterion is inadequately met can that be mitigated by another criterion being met well? And could a student go on failing to meet that criterion but always manage to pass because the aggregated mark is a pass? This could go at least some way to explaining the findings of Rump, amongst many others, that "*The research shows concurrently that students often show serious lack of understanding of fundamental concepts despite the ability to pass examinations" (*Rump et al. 1999).

We Need to Scrap Percentages Because They Are Not Scalable

A piece of work that scores 90% is not exactly one and a half times better than a piece of work that scores 60%. In which case the scale is not truly quantitative, so any standard arithmetic operations that treat those grades as if they are is illegitimate (Dalziel, 1998). But that is exactly how those numbers will be treated once they are entered into the university's system.

We Must Stop Combining Scores As If They Were Just Numbers

The two arguments detailed above have already shown that numbers per se are not helpful or meaningful. Nevertheless, universities operate systems that treat the numbers as if they were simply numbers, rather than as symbolic representations of a range of different

[1] Don't ask why but this is another problem in itself, which we will return to.

judgements, which is what they are. As numbers, they can be added, combined and moved about. Yet as marks, representing assessment judgements, this makes no sense. For example, a mark for the report of a laboratory practical added to the mark from an exam. As marks, rather than numbers, they have assessed different things, they represent different types of learning outcomes; treating them simply as numbers obscures these underlying differences.

And even if we could legitimately treat marks just as numbers, the way university systems treat these numbers as equally valid, simplistically adding and aggregating them, ignoring differences in range and mean deviation makes the resulting outcomes statistically unsound anyway.

We Must Not Force Our Assessment to Fit a 'Normal' Distribution Curve

So-called 'normal' distribution is a pattern for the distribution of a set of data that follows a bell-shaped curve. The most extreme example of its use in university assessment that I have encountered is an internationally prestigious institution where all marks – be it those for a single piece of class work, or the degree classifications for the graduating cohort, have to be arithmetically forced to fit a 'norm' referenced curve. But how can this be justified in any way that has scholarly credibility? As Bloom pointed out, "*The normal curve is a distribution most appropriate to chance and random activity. Education is a purposeful activity and we seek to have students learn what we would teach. Therefore, if we are effective, the distribution of grades will be anything but a normal curve. In fact, a normal curve is evidence of our failure to teach" (Bloom et al, 1971).*

And what effect is such a system likely to have on the students? They are surely not going to see learning as a cooperative activity. In such a regime, any student helping any fellow student would have to accept that through doing so they might in fact be helping them to beat them to one of those few higher grades. In which case, it seems to me, you can probably kiss goodbye to any notion of collaborative learning or group-work. I don't think education should be a cut-throat competitive race where the devil takes the hindmost.

And even where there is not a formal requirement to apply a 'norm' curve, there may be informal pressures, some in the marker's own head, to try and make the results fit into a 'norm' curve.

We Must Stop Assessment Judgements From Being Distorted by Erroneous Other Factors

As if the problems with the use of numbers as the representation of assessment judgements about student learning already discussed weren't bad enough, there is an additional problem in many systems caused by the awarding of what Sadler (2009) calls transactional and bestowed credit. This is where some of the marks given are not related to learning at all but are being used to control student behaviour - marks given for attendance or taken away for handing work in late, for example. When these marks are added to marks given for evidence of learning, it just makes the resulting combination mark even more meaningless.

A friend and colleague of mine, who gained his first degree in the US, tells me that if he could have met deadlines he would have graduated with honours. As it was, he scraped a 3.0 because so many papers were docked for being late in his first 2 years. Never mind that in his last 2 years he was up around 3.5 – 3.6 and got (most) work in on time. His teachers

actually told him that marking was a game that students were required to play if they wanted high marks. Well if we are to be scholarly about our assessment practice, I suggest there is no room for views like these; marking must not be seen as a game.

We Must Stop Assessment Judgements From Being Distorted by Assessment Task, Subject Discipline and Institutional Rules

In the UK we know that students are more likely to score highly on coursework (essays, reports, presentations, etc.) than in examinations, and also in more numerate disciplines, such as mathematics and engineering, compared with disciplines such as history or sociology (Yorke, 1997; Yorke et al, 2000, Bridges et al, 2002). But the systems those marks are fed into treat them as if they are all the same. Despite this widely documented difference, university assessment systems treat all numbers as equivalent as they collect, aggregate and process them.

We hear much in the UK (and possibly even more in the US) about grade inflation and the ever rising percentage of students gaining the top degree classifications, but there doesn't seem to be a similar concern that graduating students of mathematics are four times more likely to get a first (the top grade) compared to students in history. I believe this is solely because of the systemic unfairness in the ways that marks are used to derive the degree classification. You can score 100% on a math task; in the UK, good work in history, as we have already said, will be lucky to score 75%. But there will be no concession to this difference in the way they are treated by the university's system; arithmetically, each of these results will treated in exactly the same way. And given that in the UK now over 60% of advertised graduate entry jobs do not require the degree to be in a specific discipline, and it is the quality of the degree achieved, as demonstrated by the classification that will significantly influence whether the applying student gets to interview, this systemic unfairness has very real consequences.

We also know in the UK that given the same results, the idiosyncrasies of different institutions' algorithms can make up to a degree classification difference (e.g. Armstrong et al, 1998) resulting in similar unfairness.

We Must Stop Assessment Practices Having a Negative Effect on Learning

Thus far, I have concentrated on the fairness and reliability arguments, and to my mind these alone provide sufficient grounds for change. However, there are also powerful arguments regarding the effect the use of numbers has on student learning. A depressing number of research findings indicate a declining use of deep and contextual approaches to study as students' progress through their degree programmes (e.g. Watkins & Hattie, 1985; Gow & Kember, 1990; McKay & Kember, 1997; Richardson, 2000; Zhang & Watkins, 2001). And we know that "*…students become more interested in the mark and less interested in the subject over the course of their studies"* (Newstead, 2002*).*

There is also strong evidence from work done in schools in the UK that where work is returned to pupils with feedback, but without a mark the students are far more likely to take note of that feedback and subsequently produce better future work. This is when compared with when work is returned with a mark attached. When there is a mark attached they are far more likely not to even read the feedback and they are less likely to subsequently improve (Black & Wiliam, 1998). And in the US, through the battery of MCQ tests that make up the Scholastic Aptitude Tests (SATs) that students have to take in order

to get into university, it could be argued that students are groomed to focus on ranking and numerical grading from a very early age.

A recent study in Sweden (Dahlgren et al, 2009) has also shown that opting for a grading system (as opposed to using numbers/percentages) may not be any better. The study found that as soon as any form of grading is introduced into the assessment judgement, when compared with just a simple pass/fail system, students were both far more likely to take a surface approach, and far less likely (33% cf 73%) to see the assessment task as a further opportunity for learning .

SoTLA

Although I realize that university education is extremely resistant to change, what I find most difficult to understand is that, given the power of the arguments and the weight of this research evidence, which has been known for some time, these practices nevertheless still continue. This is why we need, as I have argued before, to explicitly extend the Scholarship of Teaching and Learning movement, and its agenda, to include the Scholarship of Assessment. And the first step would be to change the name; **SoTL needs to become SoTLA**. This would start to address the 'out-of-sight out-of-mind' situation that I believe we are in currently. What's in a name is important. Having assessment in the name would act as a constant reminder of the central importance of assessment in the teaching and learning process, and the need to develop the scholarship of assessment within the academy. And only when we have a critical mass of faculty who understand the scholarship of assessment do I think we have any chance of bringing about significant change and a stop to our unscholarly practices.

References

Armstrong, M., Clarkson, P. and Noble, M. (1998), *Modularity and credit frameworks: the NUCCAT survey and 1998 conference report,* Newcastle-upon-Tyne, Northern Universities Consortiuum for Credit Accumulation and Transfer.

Black, P. & Wiliam, D. (1998) *Inside the Black Box: Raising standards through classroom assessment*, London: nfer Nelson

B. Bloom, J.T. Hastings, and G.F. Madaus (1971) *Handbook on formative and summative evaluation of student learning*, New York: McGraw-Hill

Bridges, P., Cooper, A., Evanson, P., Haines, C., Jenkins, D., Scurry, D., Woolf, H. and Yorke, M (2002), 'Coursework marks high examination marks low: discuss', *Assessment and Evaluation in Higher Education*, 27 (1) pp 35-48.

Dalziel, J (1998) 'Using marks to assess student performance: some problems and alternatives', *Assessment and Evaluation in Higher Education* 23 (4) pp 351-366

Dahlgren, L.O., Fejes, A., Abrandt-Dahlgren, M. and Trowald, N. (2009) Grading systems, features of assessment and students approaches to learning, Teaching in Higher Education, 14 (2) pp 185-194

Elander, J. & Hardman, D. (2002) An application of judgment analysis to examination marking in psychology, *British Journal of Psychology*, 93, pp 303-328

Gow, L. and Kember, D. (1990), Does higher education promote independent learning? *Higher Education*, 19, pp. 307-22

McKay, J. and Kember, D (1997), "Spoon feeding leads to regurgitation: a better diet can result in more digestible learning outcomes", *Higher Education Research & Development*, 16 (1) pp 55 – 67

Newstead, S. (2002) "Examining the examiners: why are we so bad at assessing students?" *Psychology Learning and Teaching*, 2 (2) pp 70-75

Richardson, J.T.E. (2000) *Researching student learning: approaches to studying in campus-based and distance education*, Buckingham: Open University Press

Rump, C.. Jakobsen, A. & Clemmensen, T. (1999) "Improving conceptual understanding using qualitative tests", in Rust, C. (Ed) *Improving student learning 6; improving student learning outcomes*, Oxford: Oxford Centre for Staff and Learning Development

Rust, C. (2007) "Towards a scholarship of assessment" *Assessment and Evaluation in Higher Education,* 32 (2) pp 229-237

Sadler (2009), 'Fidelity as a precondition for integrity in grading academic achievement', *Assessment and Evaluation in Higher Education*

Watkins, D., and Hattie, J. (1985) 'A longitudinal study of the approaches to learning of Australian tertiary students', *Human Learning*, 4, pp. 127-41.

Yorke, M. (1997) "Module mark distribution in eight subject areas and some issues they raise", in N. Jackson (Ed), *Modular higher education in the UK*, London: Higher Education Quality Council, pp 105-107

Yorke, M., Bridges, P and Woolf, H. (2000), 'Mark distributions and marking practices in UK higher education; some challenging issues', *Active Learning in Higher Education,* 1 (1) pp. 7-27.

Zhang, L. F. and Watkins, D. (2001) 'Cognitive development and student approaches to learning: an investigation of Perry's theory with Chinese and US university students', *Higher Education*, 41, pp. 236-261.