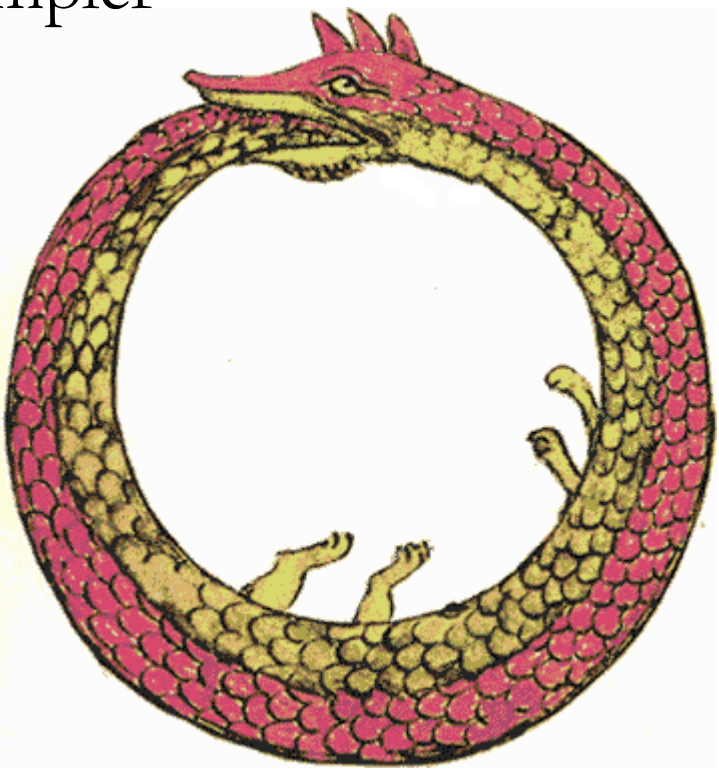
Sampler



OUROBOROS

Greg Ashman

This electronic edition was published in 2016

The cover image is of Ouroboros taken from an anonymous medieval Byzantine alchemy manuscript and is in the public domain

Chapter 7

RUBRICS

I remember watching the Sarajevo Winter Olympics as a child. I was eight-years-old and I could understand much of what was happening. The ski-jump was simple enough – whoever jumped the furthest won. Similarly, downhill skiing was a race against the clock. But then there was the figure-skating. Jane Torvill and Christopher Dean were famous at the time and I guess that pretty much the whole nation sat down to watch their performance to Ravel's Bolero.

I have to admit that figure-skating has never been a great passion of mine and so my attention focused on the scoring. They must have a complex, highly nuanced system in order to differentiate between a score of 6.0 or 5.9 out of six. How did they work it all out, I wondered? "No, they're just making it up," said my dad. It turns out that we were probably both right.

The problem with assessing a complex performance is that it is… er… complex. This is why I have suggested that, where we can, we should also try to assess the different components in isolation. Yet there will still be the need to review complex performances. Essays and other complex products will continue to matter and are the fruits of burgeoning expertise.

Royce Sadler is an Australian researcher who has thought a lot about this issue and I had the opportunity to see him talk in

2012 in Queensland. He displayed an assessment rubric and stated that the purpose of the rubric was to assess undergraduate writing. The audience was given a little time to consider this before Sadler delivered the coup de grace – the rubric was not for undergraduate writing at all, it had actually been produced to assess Grade 8 writing.

The point was well made. The rubric did not function on its own. We brought expectations to it and used these to interpret the statements in the rubric. What had seemed a perfectly reasonable approach for assessing undergraduate writing also seemed applicable to Grade 8 writing once we applied a different concept of quality. We were projecting on to the rubric an assessment that actually came from within.

A rubric is not like an engineering standard, Sadler explained. In such a standard, we can specify to the millimetre something quite explicit. Instead, a rubric will contain vague words such as 'coherence' and 'flow' and will have a graduation in performance that will move from 'demonstrating a good command of sentence structure' to 'demonstrating a sophisticated command of sentence structure' and so on. The danger is obvious; different teachers will interpret these statements differently.

The standard approach to addressing this issue is the process of 'moderation'; teachers will come together with samples of their students' work and ask other teachers to assess these samples. Some discussion may then ensue about the differences in the marks awarded. The hope is to work towards a common concept of quality within the team.

There are stronger and weaker forms of moderation. The weakest are those where class teachers mark their own papers and then choose a sample to take along to the moderation meeting. They may choose those papers that they feel the most confident about rather than the ones that need the most attention. If the second teacher already knows the mark awarded by the first teacher then this can create a 'framing effect'[1] – if the first teacher awards 8/10 then the second teacher is unlikely to award 4/10. Even when the second teacher fundamentally disagrees with the first, it is often left to the first teacher to decide whether to change the mark.

Strong personalities can dominate such meetings and the process can be emotive, with teachers feeling that they are an advocate for their students.

Now let us attach some stakes to the assessment. Perhaps a line-manager is analysing the data and teachers want to show that their students are making progress – low marks might place a teacher under scrutiny, particularly if they are lower than those given in previous assessments. Perhaps a teacher is keen to maintain a good relationship with his students and not be the bearer of bad news.

None of these biases need to be conscious. Nobody is necessarily being intentionally dishonest. Where there is wiggle-room and subjectivity, all of these factors may work to inflate marks and provide inaccurate assessments without anyone realising that this is what is happening. Everyone thinks things are going fine until an external assessment comes along and the wheels fall off the bus.

Equally, similar biases could depress marks for some sub-groups within the class. We have already seen that this could affect outcomes for students with special educational needs, behavioural difficulties, low socioeconomic status and so on.

Given these problems, I am going to suggest that such weak forms of moderation are actually a waste of time. We might be better deploying the meeting time for something else. However, we can probably do a little better if we remove some of the teacher choice in selecting papers and if we ask others to mark a paper without knowing the mark that the class teacher gave.

Continuing on this track, let's go all out and design the perfect moderation system. It is a thought experiment because it is unlikely to be possible in most schools. Nevertheless, it will highlight some key points.

Firstly, the writing assessment will be administered without the teachers present, all the papers collected by administrators and then typed-up. Names would be replaced by uniquely identifying numbers. The task would be such that students could not be identified from their responses. Each teacher would then be given a random sample of papers to assess. Of these, a random selection

of the identifying numbers would be made – again, not by the teachers – and it is these papers that must be brought along to the moderation meeting and discussed.

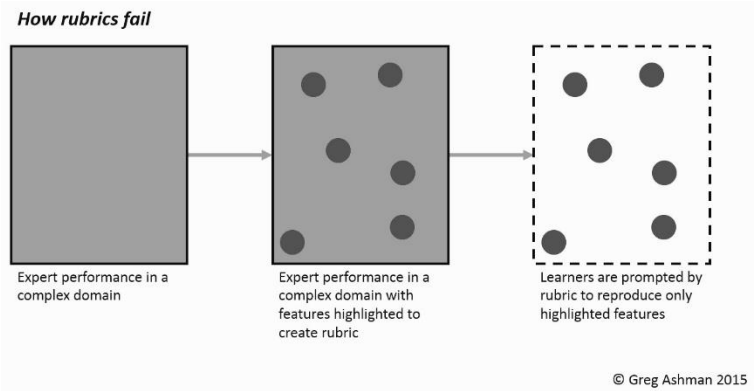
The teachers no longer know whose paper they have marked. They no longer know if it is a student in their own class or a different class. They can focus solely on the assessment. Have we fixed the problem? No.

The reason goes back to the very complexity of the task. However ambitious, a set of criteria written down in a rubric can only ever capture a sample of what an expert performance consists of. Sadler explains[2]:

> As a concrete example, consider a piece of written work such as an essay or term paper. Numerous lists and rubrics are publicly available to teachers, and commonly contain from 6 to 12 criteria. Regardless of which criteria are included in any particular list, it remains a sample. Behind it sits a much larger pool of latent criteria. One such collection, which was assembled from published lists, contained over 50 criteria. It was not exhaustive, and could have been extended further. Clearly, working with a manageable number of criteria has to involve selection, but at least for written works, any sample of reasonable size leaves out the majority. [references removed]

There is a danger here which is related to Goodhart's Law. Goodhart was an economist and his 'law' is often phrased that, "If a measure becomes a target, it ceases to be a good measure."

Imagine the following scenario: we have used our expert concept of quality or a sample of papers to derive a set criteria that distinguish between essays of differing levels of quality. These criteria are only a sample of all the aspects that we could notice but these are what we write into our assessment rubric. Teachers then view this rubric as a target and specifically focus on teaching students to demonstrate evidence of meeting these specific criteria, perhaps at the expense of other aspects of the performance.



So, even if we have a scrupulous system for moderation, we may still end up fooling ourselves. Sadler has a proposal for addressing this problem. Let us be open about the fact that some of the criteria that we use to form our judgements are not captured by the rubric[3] and let us bring these into the foreground as and when required:

> A… solution to the problem is to consider the universe of criteria as notionally partitioned into two subsets called for convenience manifest and latent criteria. Manifest criteria are those which are consciously attended to either while a work is being produced or while it is being assessed. Latent criteria are those in the background, triggered or activated as occasion demands by some (existential) property of the work that deviates from expectation. Whenever there is a serious violation of a latent criterion, the teacher invokes it, and it is added (at least temporarily) to the working set of manifest criteria. This is possible because competent teachers have a thorough knowledge of the full set of criteria, and the (unwritten) rules for using them. But it is precisely this type of knowledge which must be developed within the students if they are to be able to monitor their own performances with a reasonable degree of sophistication. [reference removed]

This is an interesting idea. Let's have a set of criteria but then invoke other, hidden criteria when required. We could have a criteria skeleton which we add different pieces of flesh to over time. Alternatively, we could have a constantly evolving set of criteria. As soon as we find that the essays are becoming a little formulaic, we can introduce a different criterion to mitigate this.

Another approach is to use relative judgements. When we measure on an absolute scale using a set of criteria, we introduce the possibility of all students scoring 9 or 10 out of 10, particularly if we have trained them well. However, what is really of instructional value are the differences between essays that score 10. What makes the best essays better than the next best essays? We won't even know there is a difference if they all score 10.

A way that we can do this is to force a comparison. We can lay the essays out on a large table and start to rank them using our expertise; our concept of quality. Once we have a rough ordering of the essays, we can start to ask: What makes these ones better than those ones?

In fact, this seems to be the most sensible way of deriving a set of criteria in the first place, rather than drawing-up a rubric in the abstract. And so we have the beginnings of a ouroboric cycle: Rank the essays, use this to derive a rubric, communicate the rubric to the next group of students or to the same group for the next essay, rank the essays again, find out what was not captured in the original rubric and is now making a difference between responses and disregard those criteria that are now less significant. And the fact that we all have to agree a *ranking* means that we cannot all simply satisfy ourselves by awarding homogenous scores. There is no point in trying to argue an essay up to a particular score because its place in the ranking will stay the same.

There is a danger that our group of teachers will go off and develop a different concept of quality to the community at large. So a useful check would be to intersperse essays of calibrated quality – perhaps from an external examination or scored by a different school or group of schools – in to the ranking. Schools could collaborate to make this work. There are even computer

programs now available that help teachers to rank essays by picking the better paper from a pair of papers.

References

1. Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
2. Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*,*34*(2), 159-179.
3. Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119-144.