TLA Interchange

Issue 3: Spring 2009 www.tla.ed.ac.uk/interchange

Are we short-changing our students? The use of preset criteria in assessment

Royce Sadler, Professor of Higher Education, Griffith University

Royce Sadler is Professor of Higher Education at the Griffith Institute for Higher Education, Griffith University in Brisbane. His writings on formative assessment have been enormously influential globally, particularly in relation to the importance of students coming to a grasp of what counts as high quality achievement if they are consistently to attain high standards in the work they produce. This article is an edited transcript of his keynote address at the TLA Colloquium on Assessment and High Quality Learning, held in June 2008 at the University of Edinburgh.

Let me begin by outlining what has become standard practice in a lot of higher education contexts in many countries – and also at school level. The basic sequence goes like this: We teach our students; we set them a task; we specify criteria at the same time (I'll raise later why we choose to do that); students produce their responses and submit them; we appraise these responses according to the criteria; and we (maybe) give them feedback. That's the end of what I call an assessment episode.

I want to contrast two ways of using criteria. They are, respectively, setting criteria before students begin work on responding to a task (I refer to these as preset), and initially not specifying any criteria at all (which now sounds like something from the assessment dark ages). Evaluation of any kind can't do without some levers, and those levers are what we call the criteria. My problem is with the requirement or convention that we specify them all in advance for the kinds of things that students produce for us in many, many contexts. There are a few contexts where you do need to specify criteria in advance, and they tend to be in the technologies. Here is an example from one particular area: it is where students have to construct an original solution to a difficult problem. The solution's quality is judged by whether the solution is a solution (at all), whether it is an efficient solution, and how robust that method of solving that problem would be for problems of a similar class. That's a very special context. I'm sure

CONTENTS

there are other contexts, but for the general sorts of assessment items and instruments that I've seen, the criteria that are set are basically those that people think make sense in terms of how the judgment should be made.

The idea of specifying criteria in advance is termed *analytic* and the opposite of that is *holistic* or *global*. Where did the whole idea for analytic criteria come from? I'm not totally sure about this, but I have been able to trace it back as far as 1920. Some of you will know the name of the psychologist Cyril Burt. He happened to be also a senior examinations officer in one of the London education authorities at the time. He collected and published many of his memoranda, which are largely on statistical issues. In one of those he wrote about the criteria that could be used for assessing writing at school level, but he didn't use the word analytic. The first use of that word seems to be by a person named Cast, who was one of Burt's doctoral students and wrote an article using the term in 1939. So it's about 70 years old at least.

The idea of using criteria did not, however, start with Burt. Back around 1750 there was an Anglo-Irish statesman and thinker named Edmund Burke. Some of you may know his work. He wrote a little treatise on aesthetics called *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful*. He was intrigued by the question of how we as humans identify those things that we call 'beautiful'. Do beautiful objects share some properties in common? If so, what are they? That's what he wanted to figure out. So he did the usual inductive thing. He got a whole lot of beautiful objects, and asked the question, 'What sorts of properties do these have in common that we respond to and label them beautiful?' He came up with seven or eight (depends whether you split one of them). I won't go through them all, but there were properties like these: things that are not too big or too small; things that tend to be smooth and rounded rather than angular; things that tend to be in soft colours rather than harsh colours, or if there's one harsh colour it is balanced by some softer colours. It's an enchanting little essay, and you can easily get the original text of it on the web. It's quite quaint to read as well.

Our interest is in whether Burke's criteria are sufficient to answer two important questions: (1) If I construct an object that has all of the listed properties, say at a high level, would it always and automatically be beautiful? And (2) are there some beautiful objects which don't measure up on the seven criteria? That's the real test of adequacy. It's two-pronged, and it's that sort of logic that I want to raise with you. If we apply those two tests to a lot of the criteria that we use in assessing student responses, we find that our set criteria fail the test. It turns out that we can recognise quality in a lot of student works before we identify the criteria that explain the quality. In Burke's case he identified beautiful objects and worked inductively from those to the criteria. He didn't go the other way round. I argue that if we wanted to test Burke's seven, we would need to do that as well.

The whole question of identifying criteria and specifying them in advance for students has grown rapidly in significance and popularity throughout higher education, in at least as much of the English speaking world as I can find, especially in the last 15-20 years. In the US, it's become an industry. There are workshops and seminars, full-time consultants, websites, trading exchanges (where you can put up your own set of criteria and get three sets back for free). You can buy sets of criteria, or software programs to help you create your own, even a couple of open source software programs (free) for keeping track of your mental processes. As I say, the whole deal has become an industry. In Australia and the UK, many university assessment policies specify that students must be given the criteria when they are given the task. That way, they'll know how they're to be assessed.

Let me just rehearse with you the reasons that seem to be behind this movement. The first two have an ethical basis. The idea is that students have a right to know how the quality of their work is to be judged before they begin constructing responses. It's not fair to spring surprises on them afterwards, implicitly saying, "You submit your work, and in due course I'll tell you the criteria I used, maybe including some that you've never heard of before. I hope that won't surprise you." That's not regarded as good practice. The second reason is the belief that all student responses to the same task should be assessed according to the same criteria. That's also seen as an element of fairness. So you see there are ethical arguments for using preset criteria.

The third reason is that set criteria provide guidance to the students. If students know they will be assessed according to fixed criteria, they know they need to attend to those during the production of their work. The criteria have a kind of shaping role. The fourth reason is that the fixed criteria are said to add objectivity to the judgment. It is claimed they can take a lot – or all – of the subjectivity out of qualitative judgments. Also, they create an explanatory 'trace' – students can see how a judgment has been arrived at. This all seems logical and totally systematic.

The fifth reason I'll mention has to do with communication. Sets of criteria form a convenient and economical way to provide feedback to students. One of the books on the use of rubrics and criteria has the subtitle something like: 'How to Provide Feedback Efficiently'. The idea is that, if you specify criteria and rate responses according to them, you've told the

students half of what they need to know for feedback. It's convenient and efficient. The final reason is this: evidence from some research studies shows that the use of criteria actually leads to improved student work. It makes a difference to student learning.

These all sound like compelling reasons. That is probably why everybody seems to have adopted the scheme, and even called it best practice. As I said before, some universities have made it mandatory. Why do I have reservations about it? First, the whole idea doesn't have the strong research or theoretical background you might suppose. (Some of this comes back to the same kinds of short-comings as Edmund Burke's anaylsis.) Second, implementing it creates a lot of difficulties for lecturers that are often not admitted or discussed. I've often found this talking to academics. In my earlier experience, when I believed in preset criteria, I tried to use criteria systematically and quite meticulously. I found that it was not easy to do. In fact, it often provided great difficulty for me. On the surface of it, the process looked smooth and straightforward, but the practice of it was tough work. I couldn't really make it work either. So I started talking with other lecturers and found that my experience was mirrored in theirs.

I have some background knowledge in human judgmental processes. I haven't been a researcher, but I've read fairly widely on the topic. As I read and read, I found some explanations about why I wasn't feeling really comfortable with the process. In spite of all the good reasons for doing it – and I think those are good reasons that have to be attended to – I had an underlying disquiet. I've now traced back through some of my articles and found that I expressed hints of this disquiet as early as 1983. It's only in the last few years that I've become uncomfortable enough to do something about it. I decided that this issue really needs to get some analysis, especially when I found so many other people whose experiences were similar. Let me now make a number of observations, initially derived from my own experience, but which I now know other lecturers have found in theirs as well.

The first observation is that linear criterion-by-criterion judgments which are later compounded into an overall global judgment is not the way I, as an assessor, actually grade student work. Let's take an essay – but it could be lots of other things such as the solution to a complex mathematical problem, a programming issue, or a project. Whatever it is, the students produce some complex works – note I'm only talking about complex works here. What I find myself doing is running with two agendas at the same time. The first one is this: I notice things. I notice particular things. If it's a written piece, I might make comments in the margin. At the same time as I'm noticing things, I'm also trying to attend to the work as a whole, how it's all coming together. I can run these two agendas simultaneously, without any real effort or conflict. Most of the time I'm not even aware that I'm doing both of those together. When I get to the end of it all, I have a kind of feeling that this is really – hey, this is really – a neat piece of work. This student has certainly got it all together. Note that I arrived at the holistic judgment without paying explicit attention to criteria. That's pretty much the order I always seemed to follow. The next step is to express that overall judgment in terms of the preset criteria.

That first observation rolls into the second. When I try to account for my judgment in terms of the initial criteria, I find a strange thing happens: the judgment derived from my overall impression – my holistic judgment – often doesn't agree with my criterion-by-criterion assessment. I actually found that kind of disagreement fairly common. Shock, horror! Don't I know how to do it properly? I talk to other people and they find it the same. So what's going wrong here?

I now am convinced that it's important to realise that step by step judgments don't necessarily lead to the 'correct' overall judgment. Our portrayal to our students of how preset criteria work sends a message to them. As a method, it implies that this is the way we go about judging the quality of their work. And it's not. I've asked lots of lecturers if they read through each essay eight times (because there are eight criteria), and of course they say, 'No'. When it's pointed out, lecturers realise that they too attend to two different agendas at once. What we do sell to students, though, is that criterion-by-criterion is the standard way complex judgments are made. That is very misleading for them. Complex judgments are probably not made that way in many educational contexts, but they are in evaluating 'Car of the Year' for a motor magazine. Evaluate the finish, the interior, the design and all that sort of thing. Put them all together, with essentially arbitrary weightings, add them all up, and look at the result. That's how you find the Car of the Year. We don't need to argue about that here, because we're not looking at ratings for cars. We're trying to get at complex students works, which they produce and we evaluate.

My third observation is this. Sometimes I evaluate a student's piece of work, and find it exceptionally good. I understand what makes it of such high quality, but I can't find that criterion expressed anywhere among the set criteria. Do you ever find that? You say, this is excellent, but somehow it doesn't connect with the given criteria. What do we do then? What people generally seem to do is distribute that extra sort of property over the ones that are there. I know none of you would do that, and I sure know I wouldn't, but I tell you, I know some people who do. I talked to one lecturer and mentioned that sometimes this happens. He said, "It doesn't happen with me". Two weeks later he phoned to say, "That does

happen with me – often. That's exactly what I do! When you said it, it didn't register with me, but it happens all the time".

You see, there is a mismatch between our global judgments and what would be our criterion-by-criterion judgments. Sometimes the whole comes out to be more than the sum of the parts; sometimes the whole comes out to be less than the sum of its parts. If you do a Google search on 'more than the sum of its parts' and also on 'less than the sum of its parts', you'll find that 'more' comes out more often than 'less'; actually about twice as often. 'Less than the sum of its parts' doesn't, for some reason, hit the headlines as much. I remember reading a review of some software for a computer game. I'm not a great computer gamer, but I read the reviews of lots of things, just to observe the evaluative language. In this review, it said a particular game was technically sophisticated. It had every kind of trick in the book, in terms of its graphics and its speed and so on. In other words, it had all of the things that you would think made a good game. In one sense, the designers had got it all together. If you looked at the game through the lens of technical sophistication, it had plenty. If you had a look through the lens of the graphics, it was tops. If you looked at it through the lens of the situations it used, it delivered. But it didn't come together as a whole game. It fell short and was, in the reviewer's opinion, a lot less than the sum of its parts. The recommendation was it was not worth buying. That is pretty savage criticism, but the review was one that I thought was really important, because that's what can happen with our own students. If we're going to be truthful and portray to students the way complex judgments are really made, we have to be truthful all the way. We cannot partition our overall judgment, after the fact, into just the original criteria, which are the bits we think should satisfy them. That would be very poor feedback, and to the extent we rely on it, we give them disjointed or misleading feedback.

Here is another aspect. There are some discrepancies between holistic and analytic judgments which are almost impossible to nail down at all. You just know that a piece of work is excellent in quality, but you can't quite find the words to explain it. I don't know for sure what accounts for that, but what I do say is this: there are some works that seem to have a certain quality that is inherent in their wholeness. It doesn't seem to be manifest or attributable to the criteria when they are taken separately. In other words, it comes together better than the criteria would suggest. That's another part of the same phenomenon of the whole being more (or less) than the sum of its parts. A further problem is that the criteria often appear distinct and separable when you write them down, but when you come to apply them, they don't seem as distinct at all. Do you ever find that? Let me tell you about another phenomenon.

Do you sometimes write references for people where you're given, say, seven selection criteria? You read through the first one and you write a few sentences about the person; then you read the second criterion, and write a few sentences about that. When you get to the third one, you think, wait a minute; some of this has been covered already. Maybe I should extract those sentences and put them down here. There, that's the third one done. The fourth criterion: oh, what's going on here? I've sort of covered nearly all of that already. Maybe I'll just write, "Yes please", and move on to numbers five, six, and seven. Do you find that?

In the abstract, the criteria are separate, independent of one another. But in concrete application, the meanings overlap and run together. Consider for a minute this table in front of me. If you measure its length and its breadth and its height, you'll get three separate measurements – but they're three separate measurements. If we say, "Let's double the length of this table", we know what we're doing. We're keeping the others constant. We can't reproduce that sort of action with the kind of criteria we deal with because the criteria are nothing like physical dimensions. They are concepts (or constructs) that we carry around in our brains. When we try to unpack concepts, we find that the borders between related concepts are not sharp. That's not a problem in everyday life! It might be a problem if you were trying to be totally analytic in your whole life, but none of us lives our life that way. We live our lives with concepts that make meaning in various contexts. Nearly every word that we use has many meanings. It's only when we string them together in a context that we understand. We could take those same words, put them in a different order in a different context and they would mean something entirely different – I don't mean the opposite, just something different. This business of criteria similarly merging into one another crops up more often than we admit.

The final observation I make is this: Different lecturers use different criterion sets, even for the same genre of work. Did you realize that? How amazing! The truth is that you can find whole digests of criteria. If you have a look at, say, criteria for written work, you'll find literally hundreds of lists. My question is this: What principle governs prioritizing or privileging one list over the others? Is there some sort of underlying philosophical or practical reason that says this list is better than that one? What happens if you pool lists? (But we don't give students 50 criteria; we give them six, eight or ten, not large numbers.) At one stage I collected criteria for written work; I stopped when I got to about 60. I referred to them in one of my articles, but after I published that article, I kept collecting. I think I got to 95. Then I thought, "This is going to go over the century, so I'll call it quits." I wasn't sure what I would do with them all anyway. I had established my point.

When I laid out the 60 criteria and gave them to lecturers who assess students' written work, I asked, "Do any or all of these matter?" Mostly, these teachers read through and said, "They all matter, or could matter in particular cases." If we're going to select from such a large pool, which shall we select? It's more complicated than that, too. They don't only all matter (at least, potentially), but they're all somehow connected as well. Some are nested within others. Some, if you take certain meanings of them, are almost opposite to others. Some are really hard to get hold of. One that appears on lots of lists that is hard to get hold of is 'flair'. When I ask people to nail down what they mean by flair, they typically don't know what to say. What they do say is something like this: "I know when something stands out and sparkles, and I say to myself, 'It's got flair'." I say, "Right, that's really explained it brilliantly!" But it's true that flair occurs in all sorts of contexts. Flair can be evident in a clinical interview, in a seminar presentation, in a video production. There's something special about it. We use this omnibus word 'flair' as if it explains something. But what it explains is not something that necessarily means the same thing in different contexts. In any case, we might all respond to things slightly differently, or we may not be comfortable using the word flair at all. That's another problem.

Here's yet another one: Criteria interact. When we put preset criteria down we are more or less telling the students that these things act independently, and when we somehow compile them, whether we add up some scores on them or whether we compile them in our brains, we are really saying that the whole is exactly equal to the sum of the parts. In practice there turn out to be some co-occurrences of features that matter more than the individual occurrences that happen separately. When a certain two things occur together, that's special. When three things in particular occur together, that's very special. If we take just six criteria, there are six basic criteria. If we allow two-way interactions, that is in pairs – a with b, a with c, a with d, a with e, …, then b with c…, and so on, – there are fifteen of those. Then there are three-way interactions. Maybe our brains can respond to more than two interactions at a time, say three-ways and four-ways, I don't know. But if we allow interactions of all orders, with only six criteria, there are about fifty different combinations that we could employ. Now which of those actually makes a difference when we're looking at a holistic judgment? I have no idea. There's a whole body of theory and experimental investigation that tries to unpack how people's assessments of complex objects can be unpacked and modelled in terms of the criteria. That technical approach has been applied to appraising apples, eggs, swine, corn, and stockbroking options – everything under the sun, including beautiful objects and student essays. There are experiments to show that in the stockbroker's mind, the stockbroker often believes that they are making a judgment about stocks to recommend on the basis of certain criteria. However, when explored experimentally, their holistic judgment does not agree with the ones they think they're attending to. Yet sometimes their judgments are consistently right. They must be responding to something that's more complicated than simply a linear combination of performance on specified criteria.

All in all, preset criteria have grave limitations. The question then arises – how do we as teachers, and perhaps policy makers, respond? In the back of our minds, we also know that holistic judgments can run aground when people simply assess by 'gut feeling'. There are all sorts of things that interfere with the ability to make consistent judgments, as when we mark a whole set of papers sequentially. Here is an example. A lot of us have to assess lots of student work in a short time. Come exam period we've got a whole lot to do. If we assess six pieces that are mediocre, and the next one comes in at slightly better, the generally tendency is to rate the later one more highly than it really deserves. It's a contrast effect that occurs partly because the earlier works started setting a base line. Similarly, after a whole series of quite good ones is followed by one that is a bit mediocre, it gets marked down a bit. These kinds of serial patterns are well documented. There are other kinds of patterns where people respond to irrelevant cues. In the literature on reliability and validity of holistic assessments, there's a fairly damning list of things which can occur. But, and this is important, nearly all of the experiments undertaken use assessors who have never been calibrated against one another; they have never been trained. That raises the issue of whether the phenomena are general and completely unavoidable, or whether through training people, we could do a lot better.

Constantly in my mind are the students. We run a risk of disadvantaging our students in lots of ways if we don't understand what we are doing. I believe the status of holistic judgments has to be raised upwards. If we do that then we have to avoid all the traps that are documented in the literature, some of which I've talked about. If we now return to the original reasons for setting criteria in advance, and the ethical reasons in particular, students should know how their work is to be judged before they do it. How would I get over that one? Well, I get over it in a way that is different from giving the students the criteria in advance. That is the wrong way to go because it is dysfunctional. It can actually damage the student's ability to become good at assessment.

This is my solution: I induct students into the principles of making holistic judgments. When I'm a lecturer I might see 150 works from a class in a semester. As I see that 150, two things happen that are given to me

on a plate. One of them is that I see a full range of quality. The second thing is that, for any given level of quality, I see a range of expressions of that quality. In other words I am fully aware that an A-level student doesn't have to do things in a fixed way. There are many ways of producing A-level work. There are many ways of producing B-level work, and C-level work and so on. I see those all the time. That's part and parcel of seeing 150 every semester. What do my students see? If they're unlucky they see one, their own. If they're lucky they see their own, plus those of one or two friends or colleagues or perhaps a few model answers on the website. That is nowhere near a sufficient basis for students to develop the ability to make high quality judgments. The range of overall quality is limited, and the range of expressions for a given quality level is insufficient.

I've been talking about quality, because I think quality is really the key issue when we're trying to assess students' work. What is quality? If you try to define it, you'll probably find that your definitions collapse. You've probably heard this over and over again: "I don't know how to define quality, but I know it when I see it". If you do a web search for "but I know it when I see it", and do another with "recognise" substituted for "know", you'll find lots of other concepts that exhibit this characteristic: democracy, honour, public interest, pornography, faith, style. My interest is just in quality. My responsibility as a university teacher is to teach students what quality looks like, so they can be appreciative of, and 'experience', the same sorts of quality that I see routinely when I'm grading.

How can that come about? I've already hinted at one tactic: they have to see exemplary material, or exemplars. These can't be ones that I as the lecturer construct, because they are always created by an expert, and are therefore not authentic. This is where certain types of guided peer and self-assessment become absolutely critical to inducting students into that knowledge base about quality. It's the quality of works of the same kind that they themselves are being asked to produce that is critical; it's as simple as that.

Earlier on, I said that part of the rationale for giving students the criteria in advance is so that the way their work is assessed is with the surprise element largely taken out. Taking uncertainty away is the fair thing to do, and is therefore an ethical imperative. But we need a solution from a different direction. Let's agree that it's not fair to spring surprises, especially in something as important as assessment. My solution is to try to bring students over to 'my' side of the appraisal desk, and start to see whole works – and many of them – as if through my eyes. Can we start them seeing the quality range? Can we start them seeing how the same quality can be expressed in very different ways? Those are what I see all the time. Can I as a teacher organize for that? Suddenly you get worried and ask me, "Hey, wait a minute. What about criteria? Where do they come in?"

First I want to state up front that you can never have any decent explanation for a particular judgment unless you invoke criteria. But the criteria you need to invoke may well be different for different works according to how salient the various criteria are. Wittgenstein noticed that we do not normally go around 'noticing things' that are not remarkable. Do you know what colour the carpet is here in this room, without looking now? You've seen it and walked on it already, and if you look down you'll know the colour. But until now, you may not have taken any notice, because there's no need to take notice. There are lots and lot of things that we do not notice unless they stand out.

Suppose we give students fixed sets of criteria. The same sets are to be applied to all responses from all the students, yet we know from experience that this same set may not be sufficient to 'cover' or explain all our judgments properly. Some of those criteria will address things which are ordinary and don't deserve any comment – at the same time we force ourselves to neglect things that do deserve comment. So apart from exemplars what do we need? We need to give students experience in making a variety of judgments, and coming up with reasons or justifications for those judgments. Every such explanation needs to invoke criteria. You say, "Well, why not give just them our criteria?" This is not the best idea, for good reason: We cannot assume that the meanings we attach to the words that we use, including the criteria, will be understood by the students.

I had an interesting conversation with a three year old once. In Australia there used to be, and maybe still are, two TV programmes specially geared for preschoolers. They are broadcast before or just after breakfast. One is called Sesame Street – which you know (I saw an episode on Edinburgh TV a few days ago). The other was called Playschool, which is very different, and produced by our equivalent of the BBC. I asked this three year old, "Which do you like best, Playschool or Sesame Street?" He thought for a little while (I'll save time by skipping the reflective pauses), and said, "I don't know". I said, "Tomorrow, if Mummy said you could watch only one of those, which one would it be?" "Playschool". "OK, so why do you like Playschool better?" "I don't know. I just do."

I pressed further. "Is there something that happens in Playschool that doesn't happen in Sesame Street? Or maybe something that happens in Sesame Street that you don't like? There must be something like that." He thought for a bit longer and then said, "The stories

are longer in Playschool." I said, "There are plenty of stories in Sesame Street." "Yes, but they're all in bits. You know when Mr. Music plays his music and the windows come, and then they open, and when they open, there's Teresa there, and she's got a book with pictures, and she's going to read to us? I like to sit down and listen to the story." (That story would probably run for five minutes.) He liked the longer story, and Sesame Street doesn't have a segment like that. We know there's an entirely different philosophy between the two programmes, but here's a three year old making a holistic judgment and then identifying criteria for judging between two programmes. In my terms, they have to do with length and continuity. I put to you that these are not nonsense criteria. They're perfectly sensible and no doubt accurate as well.

What I like to do with my students is this. I give them a piece of work from another student, just like the one they have produced themselves, and ask this unstructured question, "How good is the one you have now? Then tell me why." Some university colleagues say to me, "You've got to give them something to start with, some criteria!" I say not. Try it out with your students; see what happens. A student might say to me, "I'm not sure about this piece. I don't know; but somehow it's just not right." "So, why isn't it right?" "I don't know, it's sort of, in pieces." I say, "What do you mean, 'in pieces'?" "Well, everything's on the topic, but it's not linked together. In fact, it's like three small essays". I would use the word coherence for that. Without using that term, my students have responded to the work in front of them and given a valid explanation, but without using 'my' term.

The point is this: If I'd used the word 'coherence' at the front end, they wouldn't necessarily know what it means as a word. They may know what it means in another context, but they may not realize how coherence manifests in the present reality. Initially, it's better for them to recognize the substance of the criteria and use their own terms in referring to it, because it's the recognition that matters. That's 'noticing'. A little later on, we can attach a name that we share together. Down the track, all we have to do is say, "It's not coherent is it?" They know, and I know, what we're talking about. So over several runs at teaching in this way, we help students build up a repertoire of criteria from which we draw, as need be, to explain a judgment about quality.

Why do I think that's so important? Because when students are constructing works themselves, they need to attend to two agendas. One is the detail of what they're doing at the time, and the other is how the work's coming together as a whole. Does that sound like the dual agendas that I use when I'm trying to assess their work? They've got to attend to both of those. One of the ways we sell students short is not to engage them in making holistic judgments of the kind that we make ourselves. They need to make holistic judgments, and then come up with reasons. Those reasons are the raw materials for a discourse about criteria. There must be that discourse about criteria.

When we can recognize quality when we see it, make judgments that are not too far off the mark, and invoke criteria that we can use in discourse, we've got a fairly good grasp of what the concept of quality means in the context. That's what I want my students to have – quite generalisable knowledge. That is why I am not short-changing them if I induct them into it. It comes at the goal of not springing surprises from a different, and more educative, direction. Suppose when they go into the workforce the employer says, "Have a look at this report. Tell me what you think of it, and whether it's got implications for our firm." Can you imagine students saying, "Ah, mmm. I don't know where to start. Can you give me some criteria? I can't do it without criteria".

I've had colleagues phone me up and say, "I'm going to introduce seminars for the first time in my module. Can you send me a rubric or some criteria for it?" I say, "I think we need to have a talk about this". What the lecturer's looking for is a collection of half a dozen criteria that assess this or that characteristic. Some of these will be obvious and always important. We want for example, a person to be audible and not to babble. The number of such criteria is not very large. But many of them don't manage to get to the heart of the quality of the seminar presentation; they have to do with the actual presentation skills. These are important, don't get me wrong. They are very, very important, but the real substance is: Does this seminar presentation interest people? Do they understand the significance of the content? Do they learn from it? Does it grip them? Those aspects really matter, and they're deeper than most checklists that focus on: Can you hear them?. Are they wandering around and disturbing you? Are they fiddling with something? Coughing all the time? No, no, none of those. They have no annoying mannerisms, the presentation has some structure to it, and you can hear them – right! But that's not all we're after. Has it engaged us? Has it got our minds going? Are we learning from this? Has it come together as a whole? And to cap it all, different seminar presentations may be excellent, or dismally poor, for their own reasons. Those aspects are noticed, and are worthy of comment. The criteria that match them must, however, be drawn from that larger pool of criteria that is technically always available but never needs to be invoked in its entirety.

Unless we induct our students into recognizing that idea of quality, we sell our students short. I am giving you two references for today's talk. One is to an article called 'Indeterminacy and the use of Preset Criteria for Assessment and Grading'. It's the full argument, and

will be published early in 2009. The second reference is to a book chapter. It's complementary to the article, and is about 8,000 words long. The first half is basically a condensed version of what I have talked about today. The second half is how I try to induct students into this idea of making holistic judgments, complete with rationales for those judgments – and to press them to be realistic and honest about those judgments. Students have said to me that they've never ever come across this sort of thing before, because all their other assessments during their degrees had been using preset criteria. One of their first concerns was, "Why aren't you giving us criteria?" I said, "Well, I want to teach you about quality, that's part of my job, although you might feel the way I do this is threatening to start with." Basically I got the students to create short works, 300 words each, and then through various means of coded labelling and switching them around from one another, I asked very simple questions like, "How good is it?" The students started to say, "Well how would I know?" I said, "Have a good look, and think. Do you respond to anything; do you react to anything? Is it all coming together?"

Once, to my surprise, a student came up and whispered to me, "You know, this one I'm looking at isn't very good." "Oh? Why isn't it?" "Ummm, well the person hasn't actually done what you asked them to do." I thought, "Oh what a surprise! I've never heard that before! Much!" For the first time ever, they had recognised a problem. How many times have I written, "Question Not Answered" on exam papers since I started teaching in 1965? Thousands of times! I would have worn out rubber stamp after rubber stamp if I'd had the stamps to do it, because it happens so often. Yet somehow the students couldn't make the connection between what I want them to do and what they delivered. The feedback I had previously been so anxious to give them just didn't work effectively. I expect you can get the general idea.

Much more than we give them credit for, students can recognize, or easily learn to recognize, both big picture quality and individual features that matter. They can decompose judgments and provide reasons for them. That's the platform we should start from, not from setting up criteria which we know neither they nor we could stick to meticulously. You can now see, I hope, how the use of preset criteria short-changes the students. Thank you for listening.

Sadler, D. R. (2008). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education.*

URL: http://dx.doi.org/10.1080/02602930801956059

Sadler, D. R. (2009). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed) *Assessment, learning and judgement in higher education*. Dordrecht: Springer.

This article is an edited transcript from the TLA Centre's Colloquium on Assessment and High-Quality Learning which was held on 16 June 2008 in Edinburgh. You can listen to this talk and talks by David Boud, Dai Hounsell and David Nicol by downloading the podcasts from the TLA website. Details of forthcoming Colloquia are also available from this site.