Evaluating complex outcomes: The shortcomings of criteria-based schemes and what to do instead

D. Royce Sadler
Griffith Institute for Higher Education, Griffith University, Brisbane. r.sadler@griffith.edu.au

Presenter's prompts for a Keynote address
University of Alberta, Edmonton, April 2009.

Abstract. *This presentation comes in three parts. The first deals with the two basic models used for evaluating complex phenomena. One of these is the analytic, which employs multiple criteria specified in advance of the actual judgment. The other model is the holistic, which makes an overall judgment and subsequently unpacks it. Analytic approaches dominate in many areas of education, yet their theoretical backing is weak. They therefore often produce unsound judgments. The second part of the presentation takes this general argument and applies it to the evaluation of academic teaching, and conclude that a radical shift in approach is needed. The third part applies the argument to assessing and grading complex student responses to assessment tasks, and leads to a similar conclusion.*

- A common approach to evaluating complex phenomena, outcomes or behaviours is to analyse key properties, obtain separate judgments, measurements or 'readings' on them, and combine the results to arrive at an appraisal. For the purposes of this talk, these properties are called 'criteria'.
- This approach to complex appraisal has been termed 'analytic'; contrasts with 'holistic'. Original idea from Burt 1920 (in relation to marking only); term 'analytic' used by Cast 1939.
- The trend over the past two decades in particular has been to move steadily towards more and more specificity – even to the extent of atomism. The basic idea is that this allows for a more comprehensive characterisation picture of the phenomena (with all attributes that matter being taken into account) and leads to greater objectivity in appraisal.
- Stating these desired characteristics is what I mean by 'specifying criteria in advance'.
- This movement has been of rapidly growing significance in a host of areas of modern life, including higher education, especially over the past 15-20 years. Why?
    - **Ethics:** Faculty and students have a right to know how the quality of their work is to be judged *before* they start on their responses; it is not fair to spring surprises on them afterwards. It is also more defensible to use the same set of criteria for all phenomena in the same class (teaching or student works).
    - **Guidance:** Criteria provide a sort of blueprint for what's expected, allowing producers to shape their work accordingly, while it is being planned and under construction and also while it is being delivered (if relevant).
    - **Objectivity:** Fixed criteria (weighted if appropriate) take a lot of the subjectivity out of qualitative judgments. Allows a complicated judgment to be 'accounted for' logically and systematically.
    - **Communication:** Noting the levels on the criteria is a specific and economical (efficient) way to provide detailed feedback, because the framework has already been provided.
    - **Evidence:** Research studies shows it can make a difference to teaching development and student learning.

- Widespread use: Generally considered best practice. Some universities make it mandatory for evaluating teaching, courses, programs, and student learning.
- Strong momentum has developed: A whole industry has been created, especially in the USA: books, training programs and seminars, conference papers and journal articles, software, criteria banks. Now fully normalised in numerous higher education contexts, with many academics holding the view that appraisals made any other way are substandard or unfair.
- Given this climate and culture, why question it?
- Initial comments:
    - It does not have the sound theoretical or research backing one might have assumed.
    - Implementing it creates difficulties for assessors and faculty that are often neither admitted nor discussed.
- Evidence for my perspective: dozens of conversations with assessors, faculty and markers; my own experience as a HE teacher; my background in human judgment processes.

Seven Observations

1. The linear criteria-by-criteria judgments which are compounded to produce an overall appraisal are NOT the way faculty proceed in making complex judgments. The usual practice is to run with dual agendas, often deciding on the respective contributions of criteria *after* the appraisal. The implicit portrayal that the step-wise sequence is optimal sells colleagues and students a false message.
2. Limiting appraisals to fixed criteria means that non-standard criteria either are not, or cannot be, attended to.
3. Discrepancies occur between the holistic and analytic judgments. The whole may be more, or less, than the sum of its 'parts'. (Here, 'parts' means properties or criteria.) The tendency in practice is to adjust the separate reportings (on criteria) to agree with the holistic, not the other way round. Appraisers don't necessarily tell colleagues or the students that they do this, or why, or how. (In fact, they studiously avoid it!) Another word for it is fudging. Fudging is a serious and misleading omission, a hidden element of the analytic procedure.
4. Some discrepancies between holistic and analytic are difficult to account for (at all).
    - A complex phenomenon may possess an 'indefinable' quality that is inherent in its wholeness, but is not attributable to any of the stated criteria in particular.
    - Another phenomenon may 'stand out' on the basis of some particular quality (criterion) that is identifiable, but that criterion is not on the set list. What does one do? Anything one can do breaches the implicit contract between the assessors and the assessees that is established by promulgating a set of criteria.
5. The criteria often appear distinct and separable in the abstract, but when applied, merge into one another. This phenomenon also occurs in a lot of contexts that involve complex judgments (other than education).
6. Different lecturers use **different criterion sets** for works in the same genre. Putting all the criteria together would make a big pool. From this pool, each lecturer's set may be regarded as a (non-random) sample. But what principle governs the choice of one selection over another? In what sense are they 'equivalent'? This is never explicitly tested or explored. At the receiving end are the students, who are presented with set after set for different tasks or modules, and are supposed to make sense of it all. (Is going for a course-based set the answer? Not if the observations above have any validity.)
7. **Criteria interact**. Co-occurrence of highs on several criteria may count for more than the levels on separate criteria, but the linear/independent approach, if strictly applied, undermines this possibility. Illustrative data: With six criteria, there are six first-order effects, 15 possible two-way interactions and 40 third and higher order interactions (61 in all). The preset criteria

limit is only 6. How can we be sure none of the interactions matter or have had an influence on our global appraisal?

Application

- The development so far can be applied to many aspects of decision making about human performance, including selection for employment. I choose two that come directly from the context of teaching
- Where do we go from here? In brief:
    - Make judgments at the largest scale possible. Let all the others hang out. Salience rules.
    - Review the status of holistic [or configurational] judgments, (upwards!).
    - Avoid the traps documented in the literature about global judgments.
    - Induct students into the principles of making holistic judgments, and engage them in activities that encourage an appreciation of a 'work as a whole'. How does it come together ***as a totality***?
    - Turn this exercise into a vehicle for faculty development or teaching content and skills, not as an add-on.

Evaluating teaching: The 7 Macro Questions

Data sourced from academic peers

1. How thorough and **up to date** is the course, with the most appropriate (disciplinary, professional) material included?

2. Does the lecturer really, **really know their stuff**? [See third column in attachment.]

3. How well have the **students achieved**? [Provide peer-reviewed first-order evidence from an arms length sample of student responses to assessment tasks that are warranted to represent the students' own work.]

4. How **commensurate are the grades** awarded with the levels of achievement demonstrated? [Provide concrete first-order evidence, with particular attention to the levels of Pass and High Distinction.]

Data sourced from students

5. How much of their learning do **top-quartile students attribute to the lecturer** (regardless of approach taken)? [One questionnaire item?; better, a focus group discussion]

6. On a 1-10 scale, how highly do students rate their **experiences as learners** in the course? We want students to enjoy learning and get great satisfaction out of it. If we don't, they will be put off learning, and there's enough of that around already. [? One questionnaire item – ask them? All students?]

7. What is the nature of the **most serious concerns students have expressed** about the teacher (as a person, or about their effectiveness or professionalism as a teacher) with respect to their experiences (as persons or learners) in this course? No questionnaire can cover all conceivable concerns. [Open responses to course feedback sheets; records in the Dean's office]

- Parker Palmer:

- '**Good teaching cannot be reduced to technique**; good teaching comes from the identity and integrity of the teacher.'
- 'In every story I have heard, good teachers share one trait: **a strong sense of personal identity infuses their work**…. Bad teachers distance themselves from the subject they are teaching, and from their students. Good **teachers join self and subject and students** in the fabric of life.' (p. 10-11)
- '**Technique is what we use until the real teacher arrives**, and this book is about helping that teacher show up' (p. 5), and
- 'We teach who we are.'

- Michael Scriven's 2 papers (see References). (Try McMaster University).
- John Woolcock: Teaching students to 'think microbiologically'
- Reductionism, Taylorisation of knowledge; holistic; role of evidence (a) ***for*** reductionist approaches, and (b) ***in*** reductionist approaches. The 'measurement' of 'variables' – a meta-issue that itself is not an empirical 'fact' or discovery but a decision (to adopt that paradigm)

Evaluating student learning

- See the presenter's two articles listed in the References. The journal article is the full theoretical argument, as developed in the context of the assessment of student learning only. The book chapter contains a briefer and less rigorous account of what's in the article together with how I approached teaching in such a way as to induct students into the mysteries and practice of high-quality appraisal of student responses to assessment tasks.
- Recognise that in the real world after graduation, constructing a prior set of criteria (or reaching for a rubric) each time a complex work has to be appraised is unlikely to be an available option.

References

Palmer, P. J. (2007). *The Courage to Teach* Wiley

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education*, **34**, 159-179.

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*. Ch. 4, pp. 45-63. Dordrecht: Springer.

Scriven, M. (1988). The validity of student ratings. *Instructional Evaluation* **12**(2), 5-18.

Scriven, M. (1989). The design and use of forms for the student evaluation of teaching. *Instructional Evaluation*, **10**(1), 1-13.

Woolcock, J. B. (2006). *Within the hollowed halls of learning: Reflections on 37 years of university teaching.* Brisbane: CopyRight Publishing.

r.sadler@griffith.edu.au

EDMUND BURKE (1729–1797)

Irish statesman, author, orator, political theorist, and philosopher.

A Philosophical Enquiry into the Origin of our Ideas of the Sublime and Beautiful
1757

BEAUTIFUL OBJECTS SMALL

SECTION XVIII

RECAPITULATION

ON the whole, the qualities of beauty, as they are merely sensible qualities, are the following: First, to be comparatively small. Secondly, to be smooth. Thirdly, to have a variety in the direction of the parts; but, fourthly, to have those parts not angular, but melted, as it were, into each other. Fifthly, to be of a delicate frame, without any remarkable appearance of strength. Sixthly, to have its colours clear and bright, but not very strong and glaring. Seventhly, or if it should have any glaring colour, to have it diversified with others. These are, I believe, the properties on which beauty depends; properties that operate by nature, and are less liable to be altered by caprice, or confounded by a diversity of tastes, than any other.

Types of 'Scholarship'

General definition
Scholarship is an essentially cognitive activity that is manifest as both an attitude of mind and a practical commitment to rigour. Rigour is characterised by: an appropriate balance between personal engagement and intellectual detachment; logical reasoning; disciplined and precise thinking; adequate grounds for opinions; evidence for conclusions; and freedom from bias.

'Discovered' knowledge What 'knowledge' means: externalised, stored. **Aim:** To explore and understand the nature of some external reality; discovery of what is 'out there'; search for objective truth. **Key approaches:** Empirical research; collecting and analysing data about the reality being investigated; formalised qualitative & quantitative methods. **Indicative** disciplines**:** Mathematics, sciences, some areas of the social sciences.	**'Developed' or 'created' knowledge** What 'knowledge' means: externalised, stored. **Aim:** To create new knowledge primarily by identifying ideas and their interrelationships through **Key approach:** Creative insight; careful, systematic thinking and reasoning. **Indicative disciplines:** Creative arts, humanities, some aspects of social sciences, IT and sciences.	**Personal knowledge** What it means 'to *know*'. What people know. **Aim:** To attain intellectual mastery over a substantial body of knowledge as a privately held capital resource. **Key approach:** Learning for oneself through intensive interrogation of existing explicit and tacit knowledge; interaction with people and the external environment; and both analytic and holistic reflection on it all.
Discovery Conceptualisation Empirical research Construction of formal proofs; Demonstration of a result Confirmation or disconfirmation of an hypothesis; Validation or replication of earlier findings.	**Distillation** Analysis, Critique Synthesis, Extension Integration Reorganisation, reformulation Logical reasoning Pedagogical orientation Thought experiments Theorising	**Knowing** Extensive, comprehensive (even encyclopaedic) Deep, interconnected, integrated Brain-resident Strategically accessible Communicable, especially orally Malleable, tailorable to demand
Outputs Externalised *and exhibited in material form* Journal articles Books Conference papers Research monographs	**Outputs** *Externalised & exhibited in material form* Books Textbooks Reference books Monographs Articles Multimedia Significantly original Significantly marketable Peer reviewed or Commercially published	**Outputs** Valid answers to significant and complex disciplinary or professional questions Informed opinion, advice Grounded position statements Wise judgments Policy formulation & input Real-time discussion, questioning, conversation and interaction