Research Quality Assessment: intended and unintended consequences

DAVID BRIDGES
Faculty of Education, University of Cambridge,
and University of East Anglia, United Kingdom

ABSTRACT Assessment, especially when linked to an infrastructure of rewards for successful performance, is a powerful driver of people’s behaviour. In the context of publications it is therefore important that assessment tools focus on what is genuinely constitutive of the quality of educational research (i.e. intrinsic characteristics of quality) rather than on short cuts, which may be easier to quantify but which are only related to quality in highly contingent circumstances (i.e. extrinsic characteristics such as citation indices). Drawing on the European Educational Research Quality Indicators project, this article identifies some of the perhaps unwanted consequences of different approaches to quality assessment.

Introduction
It is, I think, one of the few truths about education that approximates to the laws of physics: that assessment systems trump almost any other feature of educational policy or practice when it comes to driving behaviour, in particular when this assessment is linked to powerful rewards and punishments.[1]

Research quality assessment is a prime example of the operation of this law. After all, it functions to determine whether an individual scholar does or does not get a job or a promotion; it operates to determine whether the fruits of one’s academic labour get published in journals or heard in conferences; and it determines who does and who does not get funding for research projects. It shapes, in this way, what will count as research, and what will be recognised as legitimate and reliable knowledge. In the case of the United Kingdom’s Research Assessment Exercise (RAE), it determines directly what core funding for research goes to individual universities and indirectly what goes to individual departments and individual scholars. The ability of an academic to sustain an active research role at all, perhaps his/her very employment, depends on its outcome. No wonder, then, that individual scholars, university departments and higher education institutions are galvanised into behaviours that meet its incessant, but somewhat unpredictable, demands.

This being the case, we need to examine carefully the kind of impact on behaviour which different forms of research quality assessment have. Never mind the impact of research, what is the impact of different forms of research quality assessment?

One set of answers to this question was provided in a report commissioned by the British Educational Research Association (BERA) and the Universities’ Council for the Education of Teachers (UCET) in the United Kingdom on the impact of the 2008 RAE (Oancea, 2010), which is based on academics’ responses to the experience of the RAE and its assessment, and their financial and institutional consequences. I want to broaden the discussion a little at this stage by reflecting not just on the United Kingdom’s RAE, but also recent debates within and around a European Framework 7 project on which I have represented BERA: the European Educational Research...
Quality Indicators (EERQI) project. My aim is to identify some of the perhaps unwanted consequences of different approaches to quality assessment and to encourage caution and circumspection in a practice which, I shall argue, can seriously damage the health of the research community.

As a starting point, however, I will assume that the main purpose of research quality assessment is to promote high-quality work – whether by giving it a public airing in journals or conferences to higher-quality research outputs or providing funding for research centres which have demonstrated evidence of this quality. Even where this can be achieved effectively, it comes, of course, at some cost. Firstly, it risks disincentivising those who are not thus recognised (Oancea, 2010), even though they may have the potential to contribute something valuable in the longer term (the RAE, of course, went some way to address this problem as far as it affected early career researchers). Secondly, it creates within the sector a division between ‘research-intensive’ universities and those that are not (to give them a negative identity) or are defined in other terms – a division which is the focus of much recent higher education policy debate, but one which the RAE was, in a sense, consciously devised to create.

Thirdly, with no equal incentive to focus attention on teaching, it tends to create an imbalance in many universities in the priority and the rewards which are attached to research as compared with teaching. With this, and with the divisions between research-intensive and non-research-intensive universities, it raises questions about the very nature of a university and what it stands for. Fourthly, it creates similar divisions between departments in the same university and between staff in the same school – and Oancea (2010) provides a number of illustrations of this. What is doubly concerning is the evidence that many universities seem to have misinterpreted the unmoderated grades awarded by different panels and used them and subject league tables as a basis for their distribution of internal funding – which adds a dimension of unfairness to the hurt of selectivity.

Finally, it demands or, because of its significance, encourages a huge investment of time (and hence money) by universities preparing for the submissions and producing the actual submissions, and by university and HEFCE (Higher Education Funding Council for England) staff (many of them seconded from universities) in carrying out the assessment. The PA Consulting Group, which was commissioned by the HEFCE to carry out an audit of the burden of cost on higher education institutions, estimated the cost of the 2008 RAE as £47 million (PA Consulting Group, 2009).

So, even if and when an exercise like the RAE works well (when, let us suppose, its assessment is valid and reliable), it does so at some cost to the higher education system and to individual researchers. The idea that the principle of research selectivity should be abandoned entirely seems to find little political favour at the moment, but my last observation about the time and cost consumed by the RAE has, of course, led (in the United Kingdom and elsewhere) to some interest in alternative mechanisms for quality assessment – and it is the risks attached to these that I want to focus on here.

Intrinsic and Extrinsic Indicators of Quality

In the EERQI project, we have come to attach significance to the distinction between what we refer to as ‘intrinsic’ and ‘extrinsic’ indicators of quality. Intrinsic indicators (though I would prefer to call them criteria) of the quality of a research text are those which are integral to the quality of that text; which are constitutive of that quality; and which are a condition of judging it to be of high quality. For example, if quality consists (among other things) in the coherence and consecutiveness of the argument, or in the validity of the methods employed, then evidence of coherence, consecutiveness or validity would be intrinsic indicators of the quality of the writing. The EERQI project followed the RAE in recognising rigour, originality and significance as three key generic indicators of research quality, and added style and integrity. Extrinsic indicators are those which do not inherently constitute elements of the quality of the piece, but which have a positive correlation with judgements based upon such elements. They predict (with a greater or lesser level of confidence) the quality which can independently be discerned in the text. We have talked about this in terms of a ‘probabilistic’ relation with quality. Thus, for example, it has sometimes been suggested that the institutional affiliation of an author – the fact that he or she is located in a
prestigious university – might bear a probabilistic relation with the quality of the text he or she has authored and could therefore act as a proxy for other more laborious forms of assessment (a suggestion, I should add, that some of us squashed very forcibly in the EERQI project).

There are at least two interconnected reasons why this distinction between intrinsic and extrinsic indicators might matter. First, the EERQI project is committed to developing assessment tools which contribute to the improvement of the quality of educational research. I have already argued that assessment, especially where it is linked to a whole infrastructure of rewards for successful performance, is a powerful driver of people’s behaviour. So it is important that we develop assessment tools which encourage behaviour that is focused on what is genuinely constitutive of the quality of educational research (i.e. the sort of intrinsic characteristics illustrated above), rather than related to it in only highly contingent circumstances (the extrinsic characteristics). It does nothing to improve quality if everyone seeks to get published in a small selection of journals identified as having a probabilistic relation with quality (a decision which, of course, may well become self-fulfilling in these circumstances), and it may even be damaging to the health of the subject. By contrast, if everyone seeks to demonstrate the validity of their research (because this is what is being assessed), this would be a behavioural effect which, by definition, we would welcome. Secondly, there is the caution provided by Goodhart’s law, which was derived originally from analysis of monetary theory and practice (Goodhart, 1983) and extended by Strathern (1997) to apply to audit in the British university system. In brief, it predicts that when something shifts from being a measure to a target, then it ceases to be a measure. The trouble is that what start off as perhaps empirically grounded (extrinsic) indicators of quality rapidly become targets that people seek to achieve – and this distorts behaviour in a way which invalidates the original evidence of an association or at least the grounds for believing that the extrinsic indicator has a probabilistic relationship with intrinsic features of quality. I shall illustrate this point more fully in what follows.

The EERQI Project and Its Search for Extrinsic Indicators

There were several different kinds of motivation for the EERQI project’s search for indicators of research quality. These included frustration with the kinds of indicators currently in use across Europe, including, in particular, existing citation indices, because they fail to acknowledge the vast majority of research from the continent which is published in languages other than English or journals other than those privileged by US-dominated systems. They also included: evidence of the unreliability of existing peer-review systems (few of which, it should be acknowledged, observe any of even the most rudimentary principles for achieving reliability in assessment); the time-consuming and expensive nature of peer-review systems (see my comments on the United Kingdom’s RAE); and excitement at the possibilities that new information and communication technologies and bibliometric systems might offer.

What the techie end of the EERQI project dearly wanted to achieve was a machine which could:

1. identify a text in any European language as an educational research text. This preliminary task itself proved dauntingly difficult even in English: established lexicographies could give a reasonable assessment of an article’s relevance to education (we have these now in seven European languages) but what else do you need to establish that it is a research paper?
2. identify semantic features of the text which might evidence intrinsic quality;
3. identify features of the text which might have a probabilistic relation to quality otherwise confirmed (for example, through a carefully constructed peer review); and
4. identify quantitative measures of quality. As Rey (2009, p. 1) has pointed out: ‘evaluation using figures always fascinates decision makers and laymen anxious to have references that are simple to understand and to handle’. What Gingras (2008) refers to as ‘the social law that any number beats no number’ underpins the international readiness to resort to any kind of numerical indicator of quality in preference to the arduous task of actually reading the text.

So, what might serve as a short cut to quality assessment and what are the possible damaging consequences – or what are referred to in some assessment contexts as their ‘washback effects’ (Alderson & Wall, 1993; Cheng et al, 2004)?
Short Cuts to Quality Assessment?

Place of Publication: the international journal

One option which has proved attractive both to French and Australian authorities [3] (for the purpose of institutional quality assessment) and many universities (for the purposes of assessing the quality of an individual scholar’s work) has been the use of ‘publication in an international journal’ as a proxy for quality. After all, this is a short cut to at least a level of peer review. Following the ‘violent criticism’ in France of the European Union-sponsored European Research Index for the Humanities (Rey, 2009, p. 5)[4], the French Agence de l’Enseignement Supérieur (AERES) drew up a list of research journals in five categories which, following some (and continuing) controversy, it claims indicates differences in kind rather than a simple hierarchy. However, since it distinguishes between international, regional, national and local journals, nobody is very convinced by the argument that this is non-hierarchical. In any case, a comparison of three rankings – the ERIH (European Research Index for the Humanities), AERES and a survey of the research community by Philippe Jeannin – by Olivier Rey (2009, p. 5) led him to conclude that ‘it is improbable that a list of scientific journals dealt with on a hierarchical basis upon which researchers in education can reach a consensus might exist’.

In a different context, the University of Ghent is one of many that use the location of publication as an indicator of the quality of staff publications in appointments and promotion (see Smeyers & Burbules, 2011). It awards 12 times the credit for a publication in an international journal as it gives for publication in a journal published in Dutch or Flemish. Paul Smeyers & Bas Levering wrote some time ago about the distortions and, indeed, injustices which result from the incentive or requirement placed on Dutch-speaking academics to publish in English in international journals. The English-language requirement, they argued, clearly puts non-native speakers at a disadvantage in seeking publication (albeit that, as Edwin Keiner observed at the recent European Educational Research Conference, ‘the real international language is broken English’). It also makes it less likely that the published research will impact on or be accessible to people who rely on or prefer their native language: ‘Would it not be very strange indeed that the very community who pays for the research in the universities can no longer expect to be informed of the outcomes, except in another language!’ (Smeyers & Levering, 2000, p. 77). The pressure to publish in English puts at risk national journals, four of which, Smeyers & Levering were reporting in 2000, were already struggling to survive.

Not only this, but, as Smeyers & Levering argue, research which is heavily and locally context bound is unlikely to be of interest to the ‘international’ research journals precisely because it requires the readership to take an interest in these local circumstances, characteristics and idiosyncrasies, and to engage with local practical and policy agendas. The risk is that precisely the sort of locally applied research which, they would argue, has the greatest potential value will be marginalised and even driven out of the educational agenda in favour of high-level research which addresses the issues which are of common interest to a wider international community, from which the tiresome demand to understand local idiosyncrasy has been removed (on the ‘international’ as a criterion of quality, see Bridges [2006]).

Citation Indices

From a continental European point of view, citation indices carry many of the same problems, since all those available rest very substantially on English-language publications and especially those from the USA. From a United Kingdom point of view, this is less of a problem, though the arbitrariness of the selection of journals for inclusion in the indices (again, they assume the validity of a very particular hierarchy for which no consistent or legitimate rationale can be provided) and the variations between indices thus irrationally constructed remain major limitations. The HEFCE has given what I think is a detailed and well-balanced consideration to the possibility of using citation as a basis of quality assessment in the next Research Excellence Framework (the successor to the RAE), but concluded that none of the existing indices are sufficiently robust across the system for it to recommend quality assessment on this basis (Higher Education Funding Council for England, 2009). It also recognises the diversity of the use and function of citation across
different disciplines. Philosophers of education, for example, use citation very differently from, for example, cognitive psychologists: they use it less frequently; they are more likely to refer to books and to sources that do not appear in contemporary citation indices (Aristotle might still head the list for philosophy of education if, for example, the Web of Science recognised Nicomachean Ethics); and they use fewer citations. Although it is possible to normalise some of these differences statistically, this possibility still leaves open some thorny issues about the granularity and reference point for such normalisation. But is the problem simply with the inadequacy of existing indices, or are there other problems related to the behaviour that the use of such indices might encourage?

Again, we are faced with the problem which is attached to the use of any such extrinsic indicator of quality (if, indeed, it is an indicator of quality): that it encourages behaviour designed to maximise the impact on the extrinsic indicator rather than maximising the quality of the research writing. Thus, we already observe publishers holding briefing meetings for journal editors to tell them how they can maximise the citation counts of articles published in their journal (it helps, for example, to publish all issues of your journal as early as possible in the year to give it the maximum opportunity to achieve citation); authors being encouraged to cite as frequently as possible articles published in the same journal (to improve the journal’s impact factor and hence gain more credit for authors published in the journal); university research managers running staff workshops on how to get cited; and academics collaborating in so-called ‘citation clubs’ to promote each other’s work.

Note that none of this activity is actually focused on the improvement of the quality of the research – which is apparently a very roundabout and uncertain way of improving your citation count! As Adler & Harzing (2009, p. 74) have argued: ‘Rather than genuinely fostering relevant knowledge, the emphasis on ranking seems to be driven by a desire to identify winners and losers in a game of academic prestige’.

One of the other approaches being trialled by the EERQI project attempts to focus on machine-discriminable features of research text. So, it is proposed to distinguish texts which feature: (1) a synopsis or abstract; (2) references or a bibliography; and/or (3) footnotes – though this seems to me to be an indication of the desperation of the EERQI research team rather than a serious contribution to research quality assessment. Apart from the fact that there are overwhelming difficulties in relying on such indicators (many journals now prohibit footnotes; references for chapters in books are commonly collated into a single list at the end; and neither books nor book chapters tend to have synopses), even the presence of all these features would indicate little more than that the piece has the style of a research article, and would tell you nothing about its quality. Such spurious indicators are hardly worth taking more seriously than Jim Dillon’s (1981) piece, published in the American Psychologist, which demonstrated conclusively a high correlation between quality, as judged independently, and the presence of a colon in the title of an article. Dillon vowed that he would thereafter always use two colons in his titles in order to guarantee their superior quality! If the last example is ridiculous, it illustrates, nevertheless, the futile distractions of attempts at assessment that stray too far from intrinsic features of quality and cautions against putting much faith in correlations.

Impact

In the United Kingdom, the central focus of current debate about how to judge research quality is on the proposal that assessment should include an assessment of the impact of a particular piece of research or a programme of research or the work of a whole unit of assessment. The HEFCE consultation document is careful to distinguish between assessment of quality and assessment of impact, arguing that impact needs to be assessed in terms of ‘economic and social benefits that are built upon research of the highest quality’ (HEFCE, 2009, para. 53a; original emphasis). More widely, this distinction becomes blurred and it is difficult to believe that in practice impact would become regarded as a criterion of quality rather than an entirely independent variable.
The debate has already indicated what are plausibly going to be the consequences of adopting impact as a criterion contributing to the overall grade in research quality assessment. Most obviously, it illustrates another consequence of almost any approach to research quality assessment: that the criteria employed almost always privilege some kinds of research over others. In the kind of ‘evidence-based educational policy’ advocated by the What Works Clearinghouse in the USA (and followed by some disciples in the United Kingdom), the randomised controlled experiment becomes elevated as the ‘gold standard’ for educational research, exhibiting the quality which uniquely should command respect among policy makers – and this view about what constitutes quality in educational research serves to exclude many forms of enquiry which might otherwise be equally deserving of credit and trust (see Bridges et al, 2009). On a different scale, there have been recurrent debates among the programme organisers for the British and European Education Research Association conferences, and no doubt other settings, about the criteria against which conference papers are evaluated. An assessment which calls, for example, for a score for (1) a clear statement of the research methods employed, (2) evidence of ethical consideration for research participants and/or (3) a clear statement of the results might well leave many submissions rooted in the humanities, and the theoretical and critical literature with no score in these categories. To take a third example, the use of impact as a criterion of quality privileges research which is near policy and practice rather than research which is more abstract, more deeply engaged with theory, and more fundamental, perhaps. It may privilege large-scale policy-orientated research rather than small-scale action research focused on a local setting (if, for example, scale of impact is one measure). Generic and apparently non-exclusive criteria such as ‘rigour’ only work insofar as they leave open the interpretation of the criterion in terms appropriate to the different genres of research under scrutiny (and insofar as an assessor can correctly identify the genre and hence the criteria appropriate to its assessment).

Conclusion

Let me conclude by summarising some of the risks and damaging consequences to which different approaches to research quality assessment expose us. These include:

- demoralisation and divisiveness in higher education institutions (illustrated in various forms in Alis Oancea’s [2010] report);
- distortion of the research agenda (in the example I gave, to meet the requirement for international publication or citation in US-dominated citation indices);
- distortion of the world of academic publishing by the use of artificial hierarchies of journals;
- the diversion of academic researchers into trivial pursuits aimed at scoring points on indices which have nothing to do with quality;
- the privileging of some forms of academic production (for example, journal articles) over others (for example, books, which do not appear in citation indices or have machine-recognisable features of quality); and
- the privileging of some genres of research and the exclusion of others by the kind of assessment criteria that are advanced – a problem which begins to take on even a somewhat sinister dimension if, as in a lot of ‘systematic reviews’, the research which is excluded is that which offers the most thoroughgoing critique of policy and its underlying assumptions.

This is, indeed, not the only significant political dimension to different approaches to research quality assessment. One always has to answer the question: Whose power over the production of research is enhanced and whose is diminished by different approaches to research quality assessment? Crudely – or by way of illustration – the more emphasis is placed on peer review, the more control is left in the hands of the academic community; the more emphasis is placed on citation, the more power lies in the hands of the major international publishing interests that control the selection of the journals for inclusion in the indices; and the further one can move towards machine-readable text and/or machine-operable assessment tools, the more one offers research quality assessment into the hands of the bureaucrats. If I am right in arguing that quality assessment is a powerful driver of academic behaviour, then those who control the mode of assessment also control the pattern of academic behaviour.
These are pretty weighty considerations. They warn of caution and circumspection in any kind of research quality assessment. It should always come with a warning: research quality assessment can seriously damage the health of the research community.

Notes

[1] This article was first presented as a contribution to a keynote symposium on ‘The Impact of RAE 2008 on Educational Research Units’, convened by Dr Alis Oancea at the annual conference of the British Educational Research Association, University of Warwick, 2 September 2010.

[2] The aim of the three-year EERQI project ‘was to reinforce and enhance the worldwide visibility and competitiveness of European research by developing new indicators and methodologies to determine the quality of educational research publications’. For further details, see http://www.eerqi.eu. This article was written and submitted several months before the conclusion of the EERQI project in March 2011.

[3] This refers to the work of the Centre for the Study of Research Training and Impact at the University of Newcastle, Australia, which developed and analysed a database of 1042 English-language journals in the field of education (Fairbairn et al, 2008).


References


David Bridges

http://dx.doi.org/10.1111/1467-8527.00134


DAVID BRIDGES is a Professorial Fellow in the Faculty of Education at the University of Cambridge and Emeritus Professor at the University of East Anglia (where he was previously Pro Vice Chancellor), Norwich, United Kingdom. He is a Council member of the European Educational Research Association, a former Council member of the British Educational Research Association and an Honorary Vice-President of the Philosophy of Education Society of Great Britain. He served on the 2001 and 2008 Research Assessment Panels for Education. His publications include *Fiction Written under Oath? Essays in Philosophy and Educational Research* (Kluwer, 2003) and, with Paul Smeys & Richard Smith, *Evidence-based Education Policy: What evidence? What basis? Whose policy?* (Wiley-Blackwell, 2009). Correspondence: David Bridges, Faculty of Education, University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, United Kingdom (db347@cam.ac.uk).