British Educational Research Journal
Vol. 35, No. 4, August 2009, pp. 497–517

Research quality assessment in education: impossible science, possible art?

David Bridges*
University of East Anglia and St Edmund's College, Cambridge, UK

For better or for worse, the assessment of research quality is one of the primary drivers of the behaviour of the academic community with all sorts of potential for distorting that behaviour. So, if you are going to assess research quality, how do you do it? This article explores some of the problems and possibilities, with particular reference to the UK Research Assessment Exercise and discussion around the proposed new Research Excellence Framework and the ongoing work of the Framework 7 European Education Research Quality Indicators project (EERQI). It begins by asking whether there are any meaningful generic criteria of quality which can be applied to research, and tension between such criteria and the diverse and sometimes contradictory requirements of educational research. It then looks at attempts to identify measurable indicators of quality, including consideration of the location of the publication, citation and download counts, and approaches based semantic analysis of machine readable text—but finds all these quasi 'scientific' attempts at quality assessment wanting (hence the 'impossible science'). This is all the more the case because of their attachment to extrinsic correlates of quality rather than intrinsic characteristics of quality and hence the probability that the measures will induce behaviours not conducive to quality enhancement. Instead the article turns to a different approach. This is better expressed perhaps as quality 'appreciation', 'discernment' or even 'connoisseurship' and is rooted in the arts and humanities rather than in (quasi) science. It considers whether this might offer a better approximation to the kind of judgement involved in quality assessment of a piece of research writing than the sort of metrics approaches favoured in current discussion.

An appreciation of quality

> She was handsome, but the degree of it was not sustained by items and aids: a circumstance moreover playing its part at almost any time in the impression she produced. The impression was one that remained, but as regards the source of it, no sum in addition would have made up the total. She had stature without height, grace without motion, presence without

*Email: d.bridges@uea.ac.uk

ISSN 0141-1926 (print)/ISSN 1469-3518 (online)/09/040497-21

DOI: 10.1080/01411920903111565

> mass. Slender and simple, frequently soundless, she was somehow always in the line of the eye—she counted singularly for its pleasure. (Henry James 1902/1947, p. 6)

The central role of research quality assessment in higher education

For those who have gone through the recent experience in the UK of the Research Assessment Exercise (RAE)[1]I probably do not need to labour the observation that research quality assessment is one of the main drivers of the behaviour of higher education institutions and their staff. It controls access to publication in books and journals and the presentation of one's work at academic conferences; it plays a central role in the appointment and promotion of academic staff (notwithstanding some worthy attempts to give due credit to, for example, teaching quality); it determines who gets grants from research councils and other bodies; it determines through the RAE process the core and infrastructure funds which universities receive for research; and it plays an important part in shaping the recruitment of postgraduate students. If, in the words of the old song, 'money makes the world go round', it is research quality assessment which, in the higher education context at least, makes the money go round.

It is argued with some justification that 'academic research is probably one of the most rigorously and consistently evaluated sectors in modern society' (Blockmans, 2007, p. 89), and this is, of course, partly what provides its warrant as basis for public understanding. However, in the process, this assessment actually shapes the very construction of knowledge in the academy. What counts as good research subsumes a set of principles about what will count as research at all. The UK RAE has in fact allowed—with admirable liberality of judgement—that a website or piece of software, a painting, a sculpture, a curated exhibition, a musical composition or a novel might all be acceptable as 'research outputs' alongside books and journal articles. In so doing it has aligned such activity with the more traditional work of the university: but they might not have done so. To take examples closer to our own interest, the What Works Clearing House (www.w-w-c.org) has notoriously measured research quality against the standard of the double blind controlled experiment with the result that vast swathes of educational research (which would normally pass as of high quality but different methodology) have been excluded. Even the EPPI Centre (www.eppi.ioe.ac.uk) 'systematic reviews' have 'systematically' excluded from consideration in most of their reviews, for example, historical and philosophical work, case studies and biography, critical theory, discourse analysis, deconstruction and small-scale action research. Over a period of time such inclusions and exclusions begin to affect what knowledge, what research, there will be a place for in the university.

If research quality assessment is going to play such an important role in controlling the destinies of individual researchers, of university departments and indeed whole universities and of the knowledge which has a place in these universities, then it rather points to the importance of assessment practices which are fair, valid, reliable and consistent with the values and principles which inform the practice of higher education—but this is easier said than done.

Some complexities in assessing the quality of educational research

I shall focus here on educational research, but I think it fair to say that much of what I have to say about the diversity of educational research (which is perhaps at the heart of the problem) would be applicable to other areas of social science research (especially applied social science) and, in these postmodern times, to the humanities. I suspect that the natural sciences too are considerably less homogeneous than is often presented and that they would be more obviously diverse were it not for the huge resources which have been dedicated to work occupying the central ground of experimental scientific method. (What space, for example, is there for a phenomenographic study of patients' experience of illness compared with the resources available for a pharmaceutical cure?)

In our own field the last 40 years have seen a remarkable proliferation in the intellectual resources brought to educational enquiry. This is partly the result of the fragmentation of disciplines which were in the 1970s temporarily held together. Thus different ingredients of 'Sociology' have crystallised around identities such as ethnography, critical theory, large population studies and multilayered modelling, feminist and post-colonial research, case study, etc., etc. 'Psychology' was always divided by tensions between, for example, behaviourist, cognitive and psychoanalytic approaches. Today two of the most influential currents come from phenomenography (at one end of a kind of humanistic to scientistic continuum) and neuroscience at another. Mix into this already diverse set of ingredients some literary theory and discourse analysis; some biography, autobiography and 'little stories' (including fictional ones); some history (including 'contemporary history'); some iconography and 'connoisseurship'; not to mention (my favourite from the 2001 RAE exercise) 'Narcissus myth and deconstruction'—and you are really faced with an overwhelming diversity in the theoretical framing of educational research and in the epistemological and ontological assumptions which underlie it, let alone the methods employed and the forms in which it is subsequently represented. 'Ours is a field characterised by paradigm proliferation and, consequently, the sort of field in which there is little consensus about what research and scholarship are and what research reporting and scholarship should look like', wrote the then editor of the house journal of the American Educational Research Association, *Educational Researcher*, in a piece entitled 'Educational research in an age of paradigm proliferation: what's a journal editor to do?' (Donmoyer, 1996, p. 19). If a journal editor has this problem (and one response to the proliferation of paradigms has been the proliferation of journals) how much more of a problem and a responsibility lies with those responsible for research quality assessment across a whole national system? How is one to respond?

Are there meaningful generic criteria of quality which can be applied to research?

The UK Research Assessment Exercise was constructed around three quality criteria which were believed to be applicable to the full range of research—not just in Education, but across the board. These were the criteria of rigour, originality and significance.[2]

Each subject panel was required to spell out how it would interpret these criteria in their own areas. The Education panel did this at some length:

> 26. *Originality* is a characteristic of research which is not merely a replication of other work or simply applies well-used methods to straightforward problems, but which engages with new or complex problems or debates and/or tackles existing problems in new ways. So, for example, a review of existing research can demonstrate originality if it analyses and/or synthesises the field in new ways, providing new and salient conceptualisations. Originality can also lie in the development of innovative designs, methods and methodologies, analytical models or theories and concepts.
>
> 27. *Significance* can be judged in different ways according to whether the research is basic, strategic or applied. Research has, or has the potential to have, considerable significance if it breaks new theoretical or methodological ground, provides new social science knowledge or tackles important practical, current problems, and provides trustworthy results in some field of education. These results might be empirical or analytical and theoretical, providing new (and sometimes challenging) conceptualisations, and evidence for audiences ranging from academics to policymakers and practitioners. Ways of evaluating the significance of research include judging its effects or potential effect on the development of the field, examining contributions to existing debates, and assessing its impact or potential impact on policy and practice. The nature and degree of immediate impact on policymakers or practitioners will provide some useful indication of significance in terms of 'value for use'. However, there may be reasons for high impact that are not dependent on research quality; and, equally, in many cases the observable impact of high quality research is achieved only over the longer term. Theoretical and more analytical research can also be of high significance if it takes forward the state of current 'international knowledge in its field, and has influenced, or has the potential to influence, the work of other theoreticians. In education it is possible that such significant theoretical advances also influence practitioners and/or policymakers, although it will probably need a deliberate strategy to ensure that this happens. Where appropriate, evidence of any of these forms of significance should be provided in the 'Other relevant details' field of RA2.
>
> 28. *Rigour* can be judged in many ways, and can helpfully be associated with methodological and theoretical robustness and the use of a systematic approach. It includes traditional qualities such as reliability and validity, and also qualities such as integrity, consistency of argument and consideration of ethical issues. It certainly entails demonstrating a sound background of scholarship, in the sense of familiarity and engagement with relevant literature, both substantive and methodological. Different dimensions of rigour will be important in different types of research but rigour can best be assessed on a case by case basis using whichever dimensions are most appropriate. In the case of outputs that are primarily directed towards utility, it is still the rigour of the underpinning research work that will be assessed and will need to be evident. (Higher Education Funding Councils, 2006, pp. 32–33)

Several things are clear from this attempt to formulate the criteria in terms appropriate to the education research community. First—and this seems entirely appropriate in the light of the picture I have presented of the diverse forms of educational research—the rubrics are at pains to stress the *variety* of ways in which these criteria might be satisfied: significance 'can be judged in different ways'; rigour 'can be judged in many ways', etc. Secondly, the list is expressly non-exclusive: *examples* and *illustrations* of the variety are offered but not a list of absolute requirements; what *can* be the case, *might* satisfy the requirements or *could* satisfy them are indicated but not what

must be a feature of work assessed of high quality. In this way and others the rubrics struggle to satisfy the requirement that they are open to the diversity of educational research and the different expectations which might be attached to these.

There is of course, from some perspectives anyway, a cost to this liberality and responsiveness. Working with criteria expressed in this way is not a mechanical process which can be reliably replicated in a straightforward way. It requires judgement on the behalf of the assessors and judgement which is responsive to the diversity of material based on a sensitive, open-minded and reasonably careful reading of the material—and here is the rub.

Reading—indeed a 'detailed reading'—of research was at the heart of the UK RAE, as indeed it is at the heart of most research quality assessment. Peer review, based on such reading, is, after all the very foundation of the credibility of academic publication. Of the final research profile for 'units of assessment' (for most purposes university departments), 70% of the profile was based on the reading and assessment of up to four 'outputs' (in this case mainly journal articles, book chapters and books) from each member of staff submitted for assessment.[3] The Education panel of 20 assessors (four of them from the research user community) read over 7000 'outputs' between them (from 82 institutions).

Of course they did not work entirely independently of each other. Indeed, one of the most important parts of the process—begun two years before there was anything actually to assess—was talking to each other about the criteria and their interpretation, comparing and discussing our assessment of a variety of pieces of work (drawn from outside the UK) and, as we got into the assessment process proper, comparing and moderating our assessments. One of our number said at an early stage that we would never arrive at a form of words (e.g. in our rubric) which removed all ambiguity in terms of what we might mean by 'rigour', 'originality', etc., but that we would have to build a 'community of meaning', a shared language through talking together in the way which I have described. I believe that this is exactly what happened and I don't think that there is any substitute for this process.

Let us suppose (what I would claim, but would not expect everyone to accept) that through this process of discussion and reading we were able to achieve a pragmatically acceptable level of consistency in our judgements (reliability) and of responsiveness to the variety of forms which educational research could take (validity). Do we then have a satisfactory system for quality assessment?

Well, apparently not—because the UK Higher Education Funding Councils (HEFCs) among others are desperately seeking an alternative. The main problem which such attempts are seeking to address[4] is that it all takes an enormous amount of time and, consequently, that it carries an enormous cost.[5] First of all it requires a major effort on behalf of universities to screen, select, collect and return the submitted publications as well as substantial accompanying documentation. Then it requires a huge administrative operation at the HEFCs to handle all this and to arrange for its distribution to assessors and eventual return to the universities. It also requires, of course, a huge investment of time on behalf of research assessors—most of whom had little time to do anything else in the six-month period of the most intensive activity.

Alongside this is a lurking sense that there is something not quite scientific about an assessment which relies on an individual reading of a piece of work (even with some moderation).

Even while the 2008 RAE was still under way, therefore, the government initiated a review of the system for funding research in higher education. The main features of proposals which were put out for consultation in November 2007 are captured in the following paragraph from *Research Excellence Framework* (Higher Education Funding Council for England [HEFCE], November 2007 and www.hefce.ac.uk/research/ref/):

> The new assessment and funding framework will be based as far as possible on quantitative measures. There will be an overarching framework within which differences between the disciplines will be accommodated. For the science based disciplines, funding and assessment will be driven by bibliometric indicators of research quality and data about external research income and research students. For the arts, humanities and social sciences, there will be a light touch peer review process, informed by metrics. (HEFCE 2007, p. 4)

HEFCE had already commissioned some background study (Centre for Science and Technology Studies, Leiden University, 2007 and Adams *et al.*, 2007) which gave it some confidence in the practicality of this approach, though the evidence was almost entirely garnered from science and technology. (Both reports are available at www.hefce.ac.uk/research/ref/.) By May 2009 and on the basis of further studies and work by 'expert advisory groups' HEFCE retained a commitment to bibliometrics, but this was clearly seen as 'informing' and not replacing expert review: 'There was a strong consensus that bibliometrics are not sufficiently mature to be used formulaically or to replace expert review, but there is considerable scope for citation indicators to inform expert review in the REF' (HEFCE, 2009a, p. 2).

The EU Framework 7 European Education Research Quality Indicators (EERQI) Project (www.eerqi.eu) is in search, similarly, of 'bibliometric' answers to the question of research quality assessment and is, as we shall see, exploring a wider range of approaches. So the question is posed, can we not find some more 'objective', perhaps more 'scientific', measure of research quality which can be administered much more cost effectively?

Measurable indicators of quality?

This ambition has pointed towards a number of possible indicators (some would say 'metrics') of the quality of research texts. These include: (i) the location of their publication; (ii) the number of times they have been cited; (iii) the number of times they have been downloaded. None of these seem, however, very satisfactory proxies for reading-based assessment. Let me discuss each in turn.

(i) The location of publication

Can we make a judgement of the quality of a piece of research simply by noting where it is published? After all, in certain publications a paper or chapter will not appear

unless it satisfies certain criteria applied through a process of peer review, so publication (in this context) is already an indication that it has satisfied a certain quality threshold judged on the basis of a reading of the piece in question.

Clearly a number of universities in mainland Europe think so, because this is precisely the metric used in many contexts as a basis for decisions about appointments and promotions. The University of Ghent, for example, in common with other Flemish universities, has a system of points awarded for each piece of work published in refereed journals, with publications in English language international journals earning as much as 12 times the credit as national publications in Flemish. (For a discussion of some of the distorting effects of this practice see Smeyers and Levering [2000] and Bridges [2006].) In France the Ministry of Education has introduced a system under which it awards a grade on a four-point scale as a judgement about the quality of academic journals—and the new Australian system of research assessment is employing a similar metric. This in turn serves to indicate the quality of papers published in them and, hence, decisions about appointments, promotions etc.

The European Reference Index for the Humanities (ERIH) has until recently used the letters A, B and C to divide journals into 'highest ranking international publication' (A), 'standard international publications' (B) and those with important 'local and regional significance' (C) (Corbyn, 2009, p. 7). However, the European Science Foundation (EFI), which is behind the index, has agreed to drop the letters in favour of a written description of the differences, which, it insists (against widespread scepticism), are intended to describe the different character of the journals and not to place them on a hierarchy of value. Sixty-one editors of international journals in the field, however, committed to publishing editorials in their first issues of 2009 dissociating themselves from the plan and requesting that their journals be withdrawn from the scheme. 'We want no part of this dangerous and misguided exercise' reads their joint editorial. '[This is] an expression of our collective dissent and our refusal to allow our field to be managed and appraised in this fashion'.

The Education panel in the UK RAE took a conscious decision to ignore information about the location or form of publication and judge every piece of work on its own merits. There were a number of reasons for this decision and these begin to challenge the appropriateness of relying on place of publication as an indicator of quality—at least in research in Education. To begin with, unlike perhaps some areas of scientific research, there is no clear or agreed hierarchy of English language publications which could serve as a proxy for a more direct judgement of quality—and this reflects, among other things, the great diversity in both methodology (as illustrated above) and substantive focus of research in Education. Secondly, even if publication in some highly esteemed research journals might provide some basis for assessing a piece of work to be of good quality, it certainly does not follow that publication in other locations (which might be motivated by a wide range of considerations) indicated poor quality. Third, the UK RAE required some fairly fine-grained quality assessment above the threshold that one might reasonably expect a good journal to maintain. The HEFCs require assessment to be conducted against a five-point scale

as follows, and invited each panel to provide its own gloss on how it would interpret the given criteria (see HEFC, 2006, for some examples):

> 4* Quality that is world leading in terms of originality, significance and rigour
>
> 3* Quality that is internationally excellent in terms of originality, significance and rigour but which nonetheless falls short of the highest standards of excellence
>
> 2* Quality that is recognised internationally in terms of originality, significance and rigour
>
> 1* Quality that is recognised nationally in terms of originality, significance and rigour
>
> Unclassified Quality that falls below the standard of nationally recognised work. Or work which does not meet the published definition of research for the purposes of this assessment. (HEFC, 2005b, p. 31)

The point is that, while we might with reasonable confidence judge that a paper published in a number of refereed educational journals would have come up to 1* or even 2* standard, it would be difficult to take the place of publication as providing sufficient evidence that the piece was of 3* or 4* standard. No educational journals publish material consistently at that level. If, as in the case of some continental European systems, it is sufficient that it has reached a threshold of academic acceptability, then it might be easier to be guided by the place of publication, even if this is still not entirely reliable. Even so, there remain issues about the way in which books are evaluated in such a system, and these are particularly significant in the humanities and social sciences where books play a more important part than in the natural sciences and are an expected form of publication of at least senior researchers.

(ii) Citation

There is a considerable body of opinion in favour of citation as a metric of research quality, not least, I suspect, because it is quantifiable and allows direct and commensurable comparisons to be made on a single scale. In particular this seems to have gained acceptance in the natural sciences (though not universally so even there). It also has certain claims to validity in so far as it reflects the attention which an academic's peers (at least) have given to a piece of work. In this sense citation (of journal articles) indicates two levels of peer review. In the UK, as I have indicated, the Higher Education Funding Council for England has taken a favourable view of citation as, at least, data which should inform if not replace expert review.

There are nevertheless a number of objections to the use of citation as a proxy for quality assessment. The first relates to the narrow range of journals on which current citation indices are based and the apparently arbitrary selection. In philosophy of education, for example, only one out of, perhaps, four leading journals in the field appears in the Thompson/Reuters Web of Science Index. I have already noted the diversity of forms which are taken by contemporary educational research—and these have spawned an equally diverse array of journals, but only a fraction of these appear in the citation indices. In the humanities and social sciences (in contrast to the natural sciences) importance is attached to books as a form of academic expression,

but these have no place in existing citation indices. From a European and wider international perspective the problem gets worse, because existing indices acknowledge only work published in English (see Smeyers and Levering [2000] on the impact of this on Dutch language researchers). The Framework 7 EERQI project is seeking to address some of these issues by piloting a European Education Research Citation Index which will include books and book chapters and French, German and Swedish language publications as well as those in English, but there are a number of technical difficulties to be overcome before this can operate in any significant way.

A second objection also points to the unreliability of the databases which provide evidence of citation. The report on the May 2009 discussions within the HEFCE Expert Advisory Group meetings observes that 'Differences in the two citation bases (Web of Science and SCOPUS) led to some marked differences in the results (HEFCE, 2009a, p. 5). It remains to be seen whether other databases more widely used in different disciplines (such as ArchiV and Google Scholar) would produce even more diverse outcomes, but given the diversity of material which is included in these different databases, it seems likely that they would.

A third set of objections to the use of citation metrics to assess quality is to do with the invalidity of the metric. It is argued that work may be cited not in any way in recognition of the quality of its contribution to the field but because of its crass stupidity; it may get attention because it is especially controversial (e.g. people continue to cite the Millgram experiments, not because of the quality of the work but as examples of unethical research). In short, comparatively high citation counts may have nothing to do with the quality of the work being counted. Semantic analysts working on the EERQI project are exploring the possibility of identifying the semantic features of 'negative citation'. I guess that the proximity of the words 'absolute rubbish' to the citation might be one such indicator, but I am not sure, given the nuanced way in which academics tear each others' work apart, that they are going to be able to produce a very reliable tool for this purpose.

Fourth, citation rates vary enormously from one subject to another (a discrepancy which can, however, to some extent be taken into account).Of work published in 1998 (i.e. mature work which has had a chance to receive critical attention) the average citation rate for molecular biology (the highest scoring) is 46.06, for social sciences 7.48 and for mathematics (the lowest scoring) 5.73. The rates also steadily decline as you approach the most recent year. The comparative figures for 2008 publications were: molecular biology 0.75, social sciences 0.16 and mathematics 0.11 (source: Thomson Reuters' Essential Science Indicators Database as presented in the *Times Higher Education Supplement*, 2009, p. 23). It becomes questionable when you start to deal with the very small numbers in the social sciences, for example, how significant these figures really are. It would be very easy in what purports to be an indicator of quality for the figures to be disturbed by other quite irrelevant features of the article in question.

There are, of course, technical means available to deal with these variations through statistical 'normalisation' of the results (e.g. citations for a specific paper

may be normalised so that they are compare with rates in the same field, in the same sub-disciplinary area or even the same journal). But these distinctions raise complex questions about what is the right level of normalisation; what level of granularity or aggregation should we employ in assessing the impact of a piece of writing by reference to the number of citations. An HEFCE paper on 'The use of bibliometrics in the REF' (HEFCE, 2009c, pp. 25–26) provides an example of a paper published in the *British Journal of Haematology* in 2002 which received 24 citations to the end of 2007. When normalised to the average citations for that journal volume its impact was 1.76; normalised for the Web of Science subfield of Haematology the impact was 1.52; normalised for the Thompson current contents category for the field of Haematology it was 1.38; normalised for the Thomson ESI Field category of clinical medicine it was 1.72. The report observes:

> There is no simple answer to the question 'what is the right level for normalisation?' Common sense suggests that it should not be too fine a level, which becomes unduly self-referential. Nor should it be too coarse a level, which loses any sense of disciplinary and cultural context. But in between there are important nuances about the relative significance of fields and sub-fields.

Moreover, 'The answer is not solely a technical one, if it is technical at all. It is also, perhaps, largely political. Decisions about the level of normalisation will be value judgements' (HEFCE, 2009c, p. 26)—and ones, I might add, of considerable significance for the organisation and construction of knowledge in the academy.

A fifth concern is to do with the ways in which academics might try to buck the citation system by, for example, establishing 'citation clubs', i.e. networks of researchers who agree to boost each others' tally by citing each other. Already, it is argued, 'citations are the currency through which scientists pay others back'. Moreover '[they] have become a currency which is convertible into dollars and cents' (Figà-Talamanca, 2007, pp. 83, 84). It may be comparatively easy to develop software which will pick up patterns of mutual citation. However, given that such cross-referencing between networks of academics is a common feature of academic practice—especially in the somewhat fragmented and ghettoised structures which are characteristic of contemporary academic life—it would, I think, be difficult to distinguish between a natural and an artificially constructed network of cross referencing.

(iii) Downloads

Counting downloads of articles from the Web is sometimes offered as another metric of research quality and even vaunted as a 'democratic' form of assessment, since it is an indicator of what has interested a wide community of potential readers. However, this metric shares many of the limitations of citation, without, perhaps, some of the strengths. To start with, some publications are downloadable free of charge, some at a cost, and some not at all. These differences clearly impact on the number of downloads in a way which has nothing to do with the quality of the article (or book) in question.

Downloads are driven mainly by the connection between key words in the title and the interest of the person accessing the paper. The most frequently downloaded article among recent issues of the *Journal of Philosophy of Education* (attracting, apparently, particular interest in Pakistan) was one on the wearing of headscarves in schools. Now this is a perfectly good piece in a respected journal, but the fact that it has attracted such interest has probably more to do with the topicality and controversiality of the topic than its particular excellence.

Downloads have in any case a fairly random character. The person may not even bother to go any further when he or she sees the paper itself, especially if references to 'teenage sex' result in a rather dusty analysis of a survey rather than the anticipated set of erotic images. Even if the person accessing a downloaded paper does read it, he or she may well conclude that it is a pedestrian and uninteresting piece of work of low quality. My point, again, is that, whatever else they tell us, download counts are invalid and unreliable as measures of quality.

I have probably not exhausted the list of possible metrics of research quality—and will welcome consideration of any other alternatives—but an examination of the claims of some of the most frequently offered (and widely regarded) metrics does not give one much confidence in them. They simply lack validity as proxies for quality assessment.

From measurement to machine readable semantic analysis?

All the measures I have considered so far have related to things that happen outside the text which is to be assessed, but we normally expect to read a text, carefully, if we are serious about assessing its quality. Perhaps, then, there may be more promise in an approach to quality assessment which focuses closely on the text and its features.

One thread of the EU Framework 7 EERQI project is an exploration of the possibility of identifying the features of machine readable texts deemed to indicate good quality educational research. One measure of success would be that they could identify features which, while not conclusive evidence of quality, would nevertheless draw attention to the work as likely to be high quality. For the moment the ambition seems to be limited to being able to provide a reader with some data which might assist a more traditional form of assessment (i.e. one based on reading).

One of the discussions in the wider project concerns the threshold or thresholds of quality which might be observed in this way. Clearly, to start with, the machine has to be able to recognise whether a particular text is indeed about *education* (something it can probably do by reference to a lexicography of educational terms). Then it needs to be able to determine whether it is indeed *research* at all. This becomes slightly more complicated unless one accepts some rather superficial distinguishing features such as the presence of an abstract and a set of references. Only then can it begin to determine whether it is *good quality* educational research (in one of its diverse forms). Here it becomes difficult to know at what threshold or thresholds to apply the criteria. It is difficult to imagine that a machine could be programmed to make the sort of fine distinctions in quality assessment that were made by the UK RAE panels.

So how might you go about programming a machine to detect indications of quality in an educational research text? One approach, which the EERQI project is still exploring, followed this sequence of steps:

- clarify the criteria of quality which you want to apply—e.g. rigour, originality, significance;
- develop a more specific set of descriptors of what you would look for in a text as evidence of quality, for example, text segments that describe the research problems and the authors' contribution to them;
- read an educational research text and highlight the words or strings of words which indicated to you that it had these features;
- programme these into the computer so that it can then search for the same or similar semantic features in other texts.

Ágnes Sándor and Angela Vorndran who are responsible for part of this work within EERQI describe their ambitions cautiously as follows: 'Our approach consists in highlighting key sentences in the articles that can be regarded as the logical backbone of the article. Our tool does not evaluate, but aims at focusing the evaluator's attention on the parts of the texts that are relevant as a basis for his/her judgment' (Sándor & Vordran 2009, p. 1). There are, however, some profound conceptual and practical problems inherent in the wider ambition to, as it were, mechanise research quality assessment. Let me just indicate three of them.

First, it is extraordinarily difficult to accommodate in this process the very diverse forms which educational research actually takes. If we take the criterion of 'rigour', the rigour in, for example, a piece of experimental psychology, is probably demonstrated by what is said explicitly in an explanation of the way in which the research was conducted. One might be looking at things like the size and representativeness of the sample, the piloting of the method to iron out any problems, care in the elimination of the influence of the researcher, the use of an appropriate statistical tool in the analysis of the data, etc. However, *none* of these would feature in, for example, the assessment of the rigour of a philosophical argument, historical writing or a piece of (auto)biographical research—all of which have a legitimate place in the educational research community. There are parts of the educational research which seek to render the researcher absent (or at least invisible), other parts that require a prominent acknowledgement of his or her presence, a biographical positioning. There are parts that require the researcher to be detached from social and political commitments, parts that require the researcher to pursue social justice through the very practice of research itself (Griffiths, 1998). Could one ever accommodate this sort of diversity and internal contradiction in any thesaurus of semantic features—and if you could, how would the machine know which it was appropriate to apply in a particular situation?

The second problem is, if anything, even more deeply problematic, because it raises issues about what happens and what needs to happen in a reading of text. I went through the sort of exercise described above with a number of texts provided by the EERQI team. I did not have too much difficulty identifying strings of words which

indicated rigour (except when it came to a more philosophical piece which I had provided where I had to point out that the rigour lay not in any one place but in the way different parts were put together as a whole). Things became much more difficult, however, when it came to identifying features of the text which indicated either originality or significance.

The problem is that recognising these features in a piece of published research is not just about seeing what is in a text, but seeing it in relation to a much wider understanding of what is in other texts (not present) and what is going on in the wider world of education and politics. Bernard Crick's early work on citizenship (e.g. in Crick & Heater, 1977) took on a whole new 'significance' in the UK when, some two decades later, his former student from Sheffield University, David Blunkett, became Secretary of State for Education and Science, decided to put citizenship education in the curriculum and called in his old mentor to advise on its content. Nothing changed in the text during this time, but it took on new significance because of what was happening in the world. As a professional educator I can know about this, but can the machine reader? And can it assess the 'significance' of a piece of research in the absence of this wider knowledge?

Even more narrowly, in the application of the criterion of originality, we are faced with a similar—and it seems to me insurmountable—difficulty, as indeed the EERQI team has acknowledged. If I judge a particular piece of work to be original, this is partly on the basis of what I read in that work, but the judgement also relies absolutely on what I also know about other work in the field. I need to feel confident that what I have spotted in the one text is not something which has already been observed or argued in another somewhere. Of course, we already have software which can detect crude plagiarism, but in applying a criterion like 'originality' we are concerned with more than an assurance that it is not plagiarised. We are not just concerned with the replication of a form of words, but with ideas that can be expressed in an infinite variety of ways.

And this brings me to my third problem, which is to do with the reading of text. As I understand it, machine-based semantic analysis can identify vocabulary, can recognise strings of words and can identify semantic features of these strings of words—and for some purposes such capacity may be very useful. However, when human beings read text they do much more than is implied by this sort of recognition. They bring all sorts of understanding and experience to text which allows them to interpret layers of meaning and significance which are not mechanically available in the text. They can observe irony and pastiche, they can pick up on the regional or social class reference which is embodied in a particular choice of vocabulary, the emotional force behind a choice of language, the political alignment implied by the author, the rhetorical as well as the logical force of an argument; they can read the meaning which is embodied in the paper as a whole as well as in particular sentences. Now perhaps I underestimate the power of semantic analysis, but I have yet to be shown that it is capable of 'reading' text in this meaningful way. I would argue, further, that unless it is, it is going to be incapable of making judgements about or even usefully informing judgements about the quality of a piece of research writing.

Impossible science?

The reason for one part of my subtitle will by now be apparent. I am unpersuaded that any of the metrics proposed provide a satisfactory proxy for the assessment of quality in educational research. I am unconvinced that machine-based semantic analysis of text will provide any very useful basis for such assessment—though I acknowledge that these are early days in the investigation of the possibilities. This is why I regard research quality assessment as 'an impossible science'.

There is a further set of arguments which reinforces this scepticism of the search for what are typically referred to as 'quality indicators'. The trouble is that what start off as perhaps empirically grounded 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. 'Goodhart's Law', which formulates this as an economic principle, 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.[6] When the research community and university managers know that, for example, citation in particular sources is what gets rewarded, then all sorts of distorted behaviour, only some of which is anticipated in the consultation document, will be produced to achieve high scores—a phenomenon well documented in the assessment literature, especially that on language assessment, as 'washback' (see, for example, Alderson & Wall, 1993; Cheng *et al.*, 2004). The *Times Higher Education* ran a piece in the context of the consultation on HEFC's proposals for a revised form of research assessment which quickly assembled an imaginative range of 'dirty tricks' that academics already planned to use to profit from a system based on citation bibliometrics (Corbyn, 2008). De Montfort University, in anticipation of future moves under the new Research Evaluation Framework, has already developed a 'Citation Strategy' and provided lunchtime workshops for its staff to help them improve their citation scores (as something quite independent, of course, from 'How to improve the quality of your research'). Journal publishers are already issuing guidance to their editors and contributors on how to raise their citation score. None of this, however, has anything to do with quality.

These considerations point to the importance of ensuring that, whatever form of quality assessment you have or whatever 'quality indicators' you employ, they reinforce rather than distract from behaviour which is actually conducive to the production of high-quality research. To this end it is important to distinguish between characteristics of a piece of research writing which are intrinsically good (type A indicators) and characteristics which may at some point in time be shown to correlate with these, but which are extrinsic (type B). Thus, for example, if I am assessing the quality of a cut diamond and this quality consists partly in its clarity and purity, then these indicators would be intrinsic or type A. I may, however, observe that high-quality diamonds are normally very expensive and that cost correlates with quality. Cost may in this way be a type B indicator of quality, but it is not an intrinsic feature of that quality. If I know people are judging the quality of my diamonds on

the basis of their cost, I can easily price them higher, but this does nothing to enhance their quality.

Similarly, in educational research, evidence of rigour or originality in a piece of writing would be a type A indicator of qualities intrinsic to the quality of the work. It may be shown that other characteristics ('indicators') correlate with judgement made on this basis—the location of the publication or the institutional base of the researcher are, as we have seen, two which are sometimes suggested. These are type B indicators. The problem is (as with the diamonds) that if I know that people are judging the quality of my work by reference to such extrinsic features, I will be driven to change these features of my work as far as I can (as in the example of the De Montfort citation workshops) but in so doing I will do nothing to improve the quality of my work. Ergo, any quality assessment system needs to ensure that this assessment is based on type A indicators, on characteristics which are intrinsic to the quality of the work, rather than indicators which are not intrinsically about quality, the use of which will distort academic practice but do nothing to improve quality.

But how can we assess such qualities, and, if a quasi scientific approach to research quality assessment appears impossible, what are we left with? At this point my paper becomes rather more speculative, but let me begin to explore some ways forward.

Possible art?

Perhaps we are looking entirely in the wrong direction in expecting even a quasi scientific answer to a question about 'quality' assessment. In ordinary circumstances we do not employ a scientific language to describe such a process. We talk about '*appreciating*' the quality of something; such appreciation is based on '*discernment*', which suggests an '*intelligent and informed appreciation*', one capable of observing relatively nuanced differences between one object of appreciation and another. We may talk about '*relishing*' or '*savouring*' the quality of something—there is almost an erotic dimension to such appreciation (and this is of course fully exploited in advertisements for chocolates, perfume or a finely designed and manufactured motor car, the quality of which is revealed when it is gently caressed by a beautiful model).

My point here is not that we should re-envisage the contents of the *British Educational Research Journal* as an erotic turn on (notwithstanding the glossy red lipstick cover) but that we might take the cue from the language I have illustrated that perhaps quality assessment owes more to aesthetics than to science—that is perhaps more akin to *connoisseurship*[7] than to measurement. The extract from Henry James's novel, *The wings of a dove*, with which I introduced this article rather beautifully illustrates this contrast, in this case with an eloquent appreciation of a handsome woman:

> She was handsome, but the degree of it was not sustained by items and aids: a circumstance moreover playing its part at almost any time in the impression she produced. The impression was one that remained, but as regards the source of it, no sum in addition would have made up the total. She had stature without height, grace without motion, presence without mass. Slender and simple, frequently soundless, she was somehow always in the line of the eye—she counted singularly for its pleasure. (Henry James 1902/1947, p. 6)

Note in particular in this extract: (i) that quality is 'not sustained by items and aids', i.e. we are looking at 'pure' quality unadorned; (ii) and most significantly—that the *qualities* the writer attributes to the woman (stature, grace and presence) are discerned in spite of the absence of their quantitative measures (height, motion and mass); (iii) that the total impression is explicitly something quite different than could be achieved by adding up the sum of its parts; and (iv) that the observation of these qualities elicits a response of pleasure in the beholder. James has quality assessment in a nutshell.

I suggest that it is fruitful to extend this picture of sensibility and discernment to a wider concept of connoisseurship when looking at the judgement of quality. The connoisseur, after all, has to be equipped with wide knowledge of and experience in the field in which she is making her assessment. She needs to be able to know what sort of thing it is she is looking at and to judge whether this is indeed the sort of thing which she can appraise or whether she needs to pass it on to someone with a different kind of expertise. She then needs to judge it in terms appropriate to what it is (is it good *of its kind*?), not against inappropriate standards or criteria. She needs to be responsive to novel features of the particular case—perhaps something she has not come across before. She will see things that the lay person will probably have missed—and she will understand the significance of things which again will have passed over the untutored eye. Then, though she may comment on specific features of the object, she needs to have a sense of how they all combine to give an overall effect. (The child in the picture may be out of proportion to the adults, but the overall effect, given that this is the Christ child, is right.) And if the work is indeed of quality she will respond to it with appreciation, with delight even.

The suggestion that forms of educational evaluation might be conceived in terms of connoisseurship is, of course, not a new one. Most significantly, it is a view advanced by Elliott Eisner through a number of publications (Eisner, 1979/1985), in particular in his essay on 'The Forms and Functions of Educational Connoisseurship and Educational Criticism' (Eisner, 1979/1985). At its worst the notion of connoisseurship can suggest something entirely subjective or at least non-transparent, an opinionated judgement which it seems impossible for the lay person to gainsay, but Eisner manages to anticipate some of the potential critique of the notion of *connoisseurship* by linking it to the more public process of *criticism*. This is how he explains the two terms and their relationship:

> Effective criticism, within the arts or in education, is not an act independent of the powers of perception. The ability to see, to perceive what is subtle, complex, and important, is the first necessary condition. The act of knowledgeable perception is, in the arts, referred to as connoisseurship. To be a connoisseur is to know how to look, to see, and to appreciate. Connoisseurship, generally defined, is the art of appreciation. It is essential to criticism because without the ability to perceive what is subtle and important, criticism is likely to be superficial or even empty. The major distinction between connoisseurship and criticism is that connoisseurship is the art of appreciation, criticism is the art of disclosure. Connoisseurship is a private act: it consists of recognising and appreciating the qualities of a particular, but it does not require either a public judgement or a public description of those qualities. (Eisner 1975/1985, p. 219)

The connoisseur/critic of educational research requires, then, the qualities of discernment and appreciation that I have been trying to convey, but these need also to be linked to a capacity to express, explain and defend the grounds for any appreciation, assessment or evaluation in a public forum—to reveal to others what the connoisseur has appreciated for himself or herself, to point to what is good, bad or indifferent and to explain convincingly the grounds for such assessment. It is this connoisseurship 'made public' (to echo one of Stenhouse's requirements for something to count as research [Stenhouse, 1980]) and the susceptibility to peer review that goes with that that protects it from some accusations of being esoteric, purely subjective or unaccountable.

My suggestion is that this account of connoisseurship/criticism taken with my earlier elaboration of the characteristics of the sort of judgement involved describes pretty well the process of quality assessment of a piece of educational research. This is, broadly speaking, what it requires. If this account is right, it is difficult to imagine any satisfactory short cut through the requirement that such quality assessment requires a direct encounter between the assessor and a piece of work, i.e. normally, a reading.

'Love's knowledge' and practical wisdom

The analysis that I have represented in this last section takes me very close to thinking which is rather beautifully articulated in a set of essays on philosophy and literature by Martha Nussbaum under the title *Love's knowledge* (Nussbaum, 1990). These relate to the wider field of ethics and are largely rooted in Aristotle's *Nicomachean ethics* (Thompson, 1955). Nussbaum might well be astonished to find the commentary applied to research quality assessment, but the work is significant and applicable because it is to do with a 'sense of life' (p. 36), of what constitutes human being and human experience and the values that lie at the heart of it and how, as a consequence, things are to be understood and evaluated. There are two particular features of what I think of as qualitative judgement that emerge from her analysis:

(i) 'The noncommensurability of valuable things'. More particularly, Nussbaum describes the tendency to reduce quality to quantity as 'ethical immaturity—at worst callousness and blindness' (p. 36). Quality in research is not reducible to a single set of values, nor representable by a single set of measures on a scale. In making qualitative judgements we have to find a way to hold a plurality of values in our minds at once and to discover such as are appropriate in the object under scrutiny.

(ii) A demand for 'a much finer responsiveness to the concrete—including features that have not been seen before and could not therefore have been housed in any antecedently built system of rules' (p. 37). 'Excellent choice cannot be captured in general rules, because it is a matter of fitting one's choice to the complex requirements of a concrete situation, taking all of its contextual features into account (p. 71). 'Aristotle', writes Nussbaum, 'stresses complexity and context, [and] both of these features call for responsiveness and yielding flexibility, a rightness of tone and a sureness of touch that no general account could adequately capture' (p. 72). Hence,

we come again to an artistic rather than a scientific model of what is involved in such judgement: 'Good deliberation is like theatrical or musical improvisation, where what counts is flexibility, responsiveness and openness to the external; to rely on an algorithm here is not only insufficient, it is a sign of immaturity and weakness' (p. 74).

Of course, neither Nussbaum nor Aristotle was writing about research quality assessment, but they were writing about perception of the good and of what it was right to do in particular circumstances, and it is my argument that this account bears a closer resemblance to what is properly involved in a judgement of the quality of a piece of research writing than is captured by any of the bibliometrics offered thus far. Further, the problem is not just with existing bibliometrics but with the whole project of reducing quality assessment to an algorithm for the interpretation of what can be measured. Quality assessment is not reducible to what can be measured scientifically. It is rather a judgement rooted in:

- an understanding of the wider historical and contemporary context of the production of such a work and its place within it;
- an appreciation of the qualities which are being sought for in the object of assessment based on prior encounters with them in other works;
- an appreciation of which of these qualities it might be appropriate to look for in the sort of work under scrutiny and what form these might take;
- an alertness to the possibility of the work exhibiting unlooked for qualities;
- perceptiveness and discernment in observing such qualities or their absence in the work;
- and, *pace* Eisner's notion of the need to ally criticism to connoisseurship, an ability to point to, articulate, explain and defend these perceptions in the public sphere.

It also requires of the person engaged in this appreciation sufficient humility to recognise when they are entering territory beyond their own actual competence and understanding—but there is no requirement for a tweed suit, flamboyant bow tie, bulbous red nose or any of the other attributes associated with 'connoisseurs' in the popular stereotype.

Whether we call such judgement connoisseurship or practical judgement, I am not too worried. It is, however, probably closer to the sort of judgement which occupies a central place in the humanities and the arts, in aesthetics and ethics than in more narrowly conceived versions of science or mathematics. However, as a helpful reviewer of this paper has properly pointed out, such 'practical judgement', including connoisseurship and the aesthetic, plays a central role in science and mathematics, properly understood, too (Hadamard, 1945; Poincaré, 1996). Part of the problem is that the twin discourses of the 'scientific' and the 'measurable' which have come to play a hegemonic role in the conversations of the educational community represent a mean and intellectually impoverished version of these two great, rich and varied traditions of enquiry.

Finally, such judgement may not be easy or uncontroversial, but it is a form of judgement with which, in a modest way, we are routinely familiar in our ordinary and professional lives: it is, in this sense, a 'possible art'.

Acknowledgements

I am grateful to Hilary Perraton, John Elliott, Cristina Devecchi and colleagues in the Centre for Applied Research in Education at St Edmund's College—as well as the anonymous *BERJ* reviewers—for their contributions to the development of this paper. Ágnes Sándor has, I hope, enabled me to correct some of my misunderstandings of the contribution of semantic analysis but will probably not agree with my assessment of its contribution.

Notes

1. Full details of the current RAE (and documentation from previous assessments) are available on the Higher Education Funding Council for England (HEFCE) RAE website at www.rae.ac.uk. (One of the features of the assessment is the Council's commitment to transparency in the procedures employed.)
2. In the European Education Research Quality Indicators project (funded under EU Framework 7) we have started operating with these plus two additional criteria of 'integrity' and 'style', but I shall leave these aside for the purposes of this article. These criteria of rigour, originality and significance are themselves riddled with ambiguity, notwithstanding the attempts of the RAE panels to clarify them. I have, however, discussed these problems elsewhere (Bridges, 2003, 2009) and will not pursue these issues here.
3. A further 20% was based on evidence of the quality of 'the research environment' and 10% on 'evidence of esteem'.
4. There are, of course, many other issues, especially to do with the unintended consequences of the RAE, some of which I discuss in Bridges (2009).
5. The Roberts Review estimated the 'real terms' cost of the 2001 RAE as £5–6 million and anticipated that the cost of the 2008 one would be substantially higher (Roberts, 2003). Figures for 2008 will be published in due course. The ambition to reduce the burden of work and the cost of research assessment through the use of bibliometrics looks, however, unlikely to be fulfilled. Expert groups working on the revised Research Evaluation Framework had concluded by May 2009 that 'whichever approach to bibliometrics was used ... HEIs will want to verify the data, and therefore the burden of using bibliometrics within the REF is unlikely to be reduced compared with the RAE' (HEFCE, 2009b, para. 18). The fact that the HEFCE Expert Advisory Group anticipates that 'any additional cost of using bibliometrics would be largely absorbed by internal management within institutions' (HEFCE, 2009a, p. 7) will not immediately endear the proposal to the nation's universities.
6. See also the report by Evidence Ltd to Universities UK on 'The use of bibliometrics to measure research quality in UK higher education institutions' (Universities UK, 2007, p. 35).
7. I first suggested this model of research assessment to Derek Hicks, then HEFCE regional consultant to the East of England, following the 2001 RAE. The suggestion apparently created considerable hilarity in the usually rather subdued corridors of HEFCE, but this response only served to implant the seed of the idea more firmly in my mind.

References

Adams, J., Jackson, L. & Marshall, S. (2007) *Bibliometric analysis of interdisciplinary research: a report to the Higher Education Funding Council for England* (Leeds, Evidence).

Alderson, J. C. & Wall, D. (1993) Does washback exist? *Applied Linguistics*, 14(2), 115–129.

Blockmans, W. (2007) The underestimated humanities and social sciences, in: A. Cavalli (Ed.) *Quality assessment in higher education* (London, Portland Press).

Bridges, D. (2003) *'Fiction written under oath?' Essays in philosophy and educational research* (Dordrecht, Kluwer).

Bridges, D. (2006) The international and the excellent in educational research, in: P. Smeyers & M. Depaepe (Eds) *Educational research: why what works doesn't work* (Dordrecht, Springer).

Bridges, D. (2009) Assessing the quality of research in higher education: the UK Research Assessment Exercise, in: T. Besley (Ed.) *Assessing the quality of research in higher education: a comparative study* (Rotterdam, Sense).

Centre for Science and Technology Studies, Leiden University (2007) *Scoping study on the use of bibliometric analysis to measure the quality of research in UK higher education institutions: a report to the Higher Education Funding Council for England* (Leiden, CSTS).

Cheng, L., Watanabe, Y. & Curtis, A. (Eds) (2004) *Washback in language testing, research contexts and methods* (Mahwah, NJ, Lawrence Erlbaum Associates).

Corbyn, Z. (2008) Researchers may play dirty to beat the REF, *Times Higher Education,* 7 February, p. 6.

Corbyn, Z. (2009) Index of journals scraps controversial grades, *Times Higher Education,* 22 January, p. 7.

Crick, B. R. & Heater, D. (1977) *Essays in political education* (Lewes, Falmer Press).

Donmoyer, R. (1996) Educational research in an era of paradigm proliferation: what's a journal editor to do? *Educational Researcher,* 25(2), 19–25.

Eisner, E. (1979/1985) *The educational imagination: On design and evaluation of school programs* (New York, Macmillan).

Eisner, E. (1985) *The art of educational evaluation: A personal view* (Lewes, Falmer).

Figà-Talamanca, A. (2007) Strengths and weaknesses of citation indices and impact factors, in: A. Cavalli (Ed.) *Quality assessment in higher education* (London, Portland Press).

Goodhart, C. A. E. (1983) *Monetary theory and practice* (London, Palgrave Macmillan).

Griffiths, M. (1998) *Educational research for social justice: getting off the fence* (Buckingham, Open University Press).

Hadamard, J. (1945) *An essay on the psychology of invention in the mathematical field* (Princeton, NJ, Princeton University Press).

Higher Education Funding Councils (2006, January) RAE 2008 Research Assessment Exercise: Panel criteria and working methods, Panel K. Ref. RAE 01/2006(K) (London, HEFC).

Higher Education Funding Council for England (2007, April) HEFCE Strategic Plan 2006–2011, updated April 2007. Ref. 2007/09 (London, HEFCE).

Higher Education Funding Council for England (2007, November) Research Excellence Framework: consultation on the assessment and funding of higher education research post-2008. Ref. RAE 34/2007 (London, HEFCE). Available online at: www.hefce.ac.uk/research/ref

Higher Education Funding Council for England (April/May, 2009a) Research Evaluation Framework Expert Advisory Groups: Summary of discussion from round 2 meetings (mimeo). Bristol, HEFCE. Available online at: www.hefce.ac.uk/research/ref

Higher Education Funding Council for England (April/May, 2009b) Research Evaluation Framework Expert Advisory Groups: Expert review of outputs in the REF (mimeo) (Bristol, HEFCE). Available online at: www.hefce.ac.uk/research/ref

Higher Education Funding Council for England (April/May 2009c) Research Evaluation Framework Expert Advisory Groups: Use of bibliometrics in the REF (mimeo), (Bristol, HEFCE). Available online at: www.hefce.ac.uk/research/ref

James, H. (1902/1947) *The wings of a dove* (Harmondsworth, Penguin) (Original work published 1902).

Nussbaum, M. C. (1990) *Love's knowledge: essays on philosophy and literature* (Oxford, Oxford University Press).

Poincaré, H. (1996) *Science and method* (Bristol, Thosemmes).

Roberts, G. (2003) *Review of research assessment.* Report by Sir Gareth Roberts to the UK funding bodies issued for consultation May 2003, Ref. 2003/22 (London, Higher Education Funding Councils).

Sándor, Á. & Vordran, A. (2009) Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. Paper for workshop on text and citation analysis for scholarly digital libraries, Singapore, 7 August 2009. Available online at http://wing.comp.nus.edu.sg/nlpir4dI

Smeyers, P. & Levering, B. (2000) Educational research: language and content: lessons in publications policies from the Low Countries, *British Journal of Educational Studies,* 48(1), 70–81.

Stenhouse, L. (1980) What counts as research? (unpublished mimeo) (Norwich, UEA/CARE Archive).

Strathern, M. (1997 Improving ratings: audit in the British university system, *European Review,* 5, 305–321.

Thompson, J. A. K. (1955) *The Ethics of Aristotle: the Nicomachean Ethics translated* (Harmondsworth, Penguin).

Times Higher Education Supplement (2009) Average citation rates by field 1998–2008, *Times Higher Education,* 12 March, p. 23.

Universities UK (2007) *The use of bibliometrics to measure research quality in UK higher education institutions* (London, Universities UK).