International Journal of Research & Method in Education
Vol. 30, No. 3, November 2007, pp. 243–269

From Procrustes to Proteus: trends and practices in the assessment of education research

Alis Oancea*
University of Oxford, UK

This article is a reflection on an area of particular interest in the current research environment, but which has not yet been explored satisfactorily in the education literature: the evaluation of educational research. The particular focus is on the UK context, but the article is informed by comparative evidence from six countries (gathered through analysis of policy and administrative documents, literature review, informal discussion and written requests for information from key persons). It identifies eight recent trends in the evaluation of education research (from performance-based funding and institutionalisation of assessment, to the de-sensitivisation of research assessment) and it explores the benefits and perils of three types of assessment procedures (peer review, bibliometrics and econometrics) as they operate at a micro, meso and macro level. The article argues that current evaluations of educational research (particularly those aimed at supporting funding decisions) tend to operate from an instrumental standpoint that largely ignores the epistemic specificity of the various fields, modes or genres of research, the assumptions about knowledge with which they work, and the cultural and social dimensions of research evaluation as a practice.

The current context of research evaluation

Several initiatives for reforming research assessment in the UK were made public in 2006. The Chancellor's budget statement, in March 2006, announced 'plans for radically simplified allocation of the research funding that goes direct to universities' (HM Treasury, 2006a). The plans were outlined in the 'science and innovation investment framework 2004–2014: next steps' (HM Treasury, 2006b):

> Recognizing some of the burdens imposed on universities by the existing Research Assessment Exercise (RAE), the Government has a firm presumption that after the 2008 RAE the system for assessing research quality and allocating 'quality-related' (QR) research funding to universities from the Department for Education and Skills will be mainly metrics-based. (p. 3)

*Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, UK. Email: alis.oancea@educational-studies.oxford.ac.uk

ISSN 1743-727X (print)/ISSN 1743-7288 (online)/07/030243–27

DOI: 10.1080/17437270701614766

A follow-up consultation was launched by the DfES later that year (June 13, with a deadline of October 13) and concentrated on five possible funding models for STEM (science, technology, engineering, mathematics and medicine) subjects, all based on indicators of external research income (DfES, 2006a). Questions were also asked about which metrics may be appropriate to the assessment of arts, humanities and social science subjects. For such subjects, a 'basket of metrics' was suggested in Appendix 2 to the consultation document (DfES, 2006a), including the following potential measures: input metrics—research grants and contract income; volume metrics: PhD completions and staff; quality and output metrics: bibliometrics, user impact, peer esteem, Research Council evaluation; institutional assessment. An expert group set up by Higher Education Funding Council for England (HEFCE) and Arts and Humanities Research Council (AHRC) to investigate the use of metrics in arts and humanities noted that there were no clear-cut differences between STEM and A&H subjects in terms of research assessment, and concluded that peer review should be preserved in the assessment of research in those fields, but that metrics were also needed for research management (with the caveat that bibliometric indicators were still underdeveloped and that conventional metrics do not capture non-traditional forms of research and research outputs). The group proposed that the following indicators were considered: research outputs; research infrastructure expenditure and other funding of the research environment; wider social, cultural and economic significance of the research process; PhD completions per research active member of staff; peer-reviewed external research income and esteem indicators (HEFCE/AHRC, 2006).

In preparation of its response to the DfES consultation, the Economic and Social Research Council launched its own consultation, with 4 September 2006 as the deadline. ESRC's questionnaire prompted potential contributors to distinguish between forms of research assessment according to their main purpose (quality assessment vs. funding allocation), and asked for specific 'appropriate' metrics for each form (ESRC, 2006).

The responses to the consultation (287 received by the DfES, more if the individual responses incorporated in the joint submissions, such as that of Research Councils UK, are also counted) varied in scope, focus and depth. While the British Academy and the Royal Society rejected the proposed metrics-based system in favour of peer review (British Academy, 2006[1]; Royal Society, 2006), the Research Councils welcomed the initiative in general terms, but cautioned against a rushed introduction of simplistic metrics before the future role of QR funding is clarified (RCUK, 2006a). In its response to the ESRC consultation, the British Educational Research Association (BERA) had also expressed concern about a simplistic approach to metrics, and supported 'a basket of metrics plus a light touch peer review' (St. Clair, 2006, p. 22). Other responses accepted the need for more significant changes, but argued for alternative models of funding and assessment; for example, the Wellcome Trust saw income metrics as appropriate for funding allocation within the discipline of biomedical research, but not for other levels of assessment (international quality assessment and institutional-level assessment) or

for non-STEM subjects (Wellcome Trust, 2006). Universities UK also rejected a system based on income metrics, in favour of a combination of discipline-tailored indicators, with added 'quality depth', and expert assessment; it also called for debate around the principles underpinning the funding and assessment system and for the preservation and consolidation of the dual support system (Universities UK, 2006).

The reaction of the government went partly against the grain of the consultation responses. It announced the setting up of a double assessment system: expert-monitored metrics (research income, postgraduate students and bibliometrics) for science, engineering, medicine and technology (STE) subjects from 2009; and a rating system using a combination of metrics (research income and postgraduate students, but no bibliometrics) and light-touch expert review, for all other subjects (including mathematics) from 2013 to 2014 (DfES, 2006b).

This article considers the assumptions underpinning these proposals in the wider context of international trends in research assessment. It then outlines the three types of assessment procedures (peer review, bibliometrics, economic and financial metrics) that were at the core of the debates surrounding the research assessment consultation and reviews their potential benefits and limitations. Finally, it suggests that much of the recent debate around these procedures (and the government's proposals) remained within an instrumental framework that tended to reduce research to its production aspects and evaluation to quantification of performance.

Unpicking current discourses of research assessment: eight problematic issues

The emphasis on research assessment in the public arena is not peculiar to the UK. Throughout the world, researchers are urged by a variety of agencies, and with particular insistence by governments, to be more accountable for their work, to contribute to, and/or endorse, a public set of criteria by which to be held to account and to participate in the fine tuning of technologies for research assessment.

Part of this sense of urgency is given by a shift in the frame of reference for research assessment, from traditional concerns with scientific worth and contribution to the advancement of the discipline, towards considerations of quality, productivity, performance, economic efficiency, impact and use, feasibility and capacity. As such, research evaluation (either *ex ante* or *ex post*) is expected to satisfy with equal efficiency the needs of a variety of (more or less traditional) contexts and agencies: from publishing, refereeing and editing, awarding degrees and titles, and employment, promotion and career development; to funding decisions, including the distribution of public funds for research (regionally and nationally); and to systematic reviewing and the use of research in practical settings and in policy formation.

As illustrated in Table 1, these developments have been accompanied by a process of specialisation and stratification of assessment into macro (international, national, multidisciplinary and disciplinary), meso (organisations, research units, programmes) and micro (teams, individuals, projects, products and outcomes)

Table 1. Stratification and specialisation of research evaluation

Level	Scope	Rationale	Evaluator	Strategies and procedures
Macro	International, national, multidisciplinary and disciplinary	Policy and strategic decisions; resource allocation; field identity and status	International organisations, professional associations, funding bodies	Econometrics, bibliometrics, expert descriptions, scenarios, peer review, consensus conferences, consultation, public debate
Meso	Organisations, research units, programmes	Allocation of funds within organisations; management decisions; human resources decisions; organisational identity, competitiveness and prestige	National strategic bodies; funding bodies; quality assurance and audit bodies; professional evaluators; management; external evaluators; public; media	Rating; peer review; bibliometrics; econometrics; international standards; accreditation; impact and use studies; benchmarking; total quality management; advisory boards; case studies
Micro	Teams, individuals, projects, outputs, and outcomes	Access to funds; publication; career and professional status; awards and recognition; decisions on: implementation, follow-up, dissemination, reviewing, etc.	Peers; human resources departments; management bodies; professional associations; grant awarding bodies; editors and referees; users and partners; public, media	Peer review; human resources management; case studies; public debates

levels. Read horizontally, Table 1 gives a description of how specialised these levels are becoming. They vary in scope (unit of evaluation), rationale (implications and utilisation of evaluation outcomes and the types of decisions that they are expected to inform), agents (agencies and individuals involved in commissioning, monitoring and carrying out the evaluations—these may not necessarily be one and the same for any particular evaluation) and strategies. Read vertically, Table 1 illustrates the stratification of assessment by showing how the different attributes of research evaluation may vary as we move from the macro to the micro level of evaluation.

This section will unpick some of these developments by highlighting eight trends in research evaluation and research at the international level that, despite their spread and their consistency with other aspects of research policies, may in fact be highly problematic. The analysis is based on comparative evidence from six countries, gathered through analysis of policy and administrative documents, literature review, informal discussion and written requests for information from key persons. It will not seek to provide a historical account of their evolution and the factors contributing to it, but, rather, it will be aimed at a snapshot of the current landscape of research evaluation in Britain and internationally. Obviously there are other ways of describing this landscape, but the eight issues outlined below will suffice to show that the aspirations

towards increased efficiency, accountability, management principles and quality assurance, mostly taken for granted in the recent debates about research assessment, may be a double-edged sword when it comes to nurturing sound, viable research cultures.

Performance-based allocation of research funds and externally defined indicators

Most developed countries have already implemented or are currently investigating the feasibility of such initiative, systems and procedures for performance-based allocation of funds for research. Throughout the 1980s–1990s, an important component of reforms of public sector financing in the UK and elsewhere was the introduction of devolved management accompanied by increased selectivity and concentration, and a move from negotiated budgets to 'performance budgeting'.[2] Based on the assumption that 'institutional management (principally rectors, presidents and deans) are rational actors, and that they maximize whatever is rewarded' (Johnstone, 1998, p. 20), performance budgeting introduced criteria other than the number of students and of staff in the allocation of public budgets (e.g. external research income, postgraduate completions and peer-reviewed quality assessments), presumably in an attempt to disincentivise institutions from expanding the volume of their activity to the detriment of its quality. This has sometimes been institutionalised as periodical research assessment exercises based on peer review and/or on metrics. While there may have been an input from professional and research communities to the development of the indicators and criteria used in these exercises, they tended to be the subject of centralised decisions and to reflect current policy moods and priorities and external quality measures (see Slowey, 1995). The UK Research Assessment Exercise epitomises this trend and, for some, its advent was related to the country's self-assumed role as a herald of New Public Management principles and practices in Europe (Leisyte *et al.*, 2006, p. 17).[3]

Pressure towards increased transparency and accountability

Recent decades have brought pressures towards increased accountability of researchers and research units to funders of research, users and taxpayers, accompanied by detailed reporting and well-documented audit trails. For Ranson (2003) this amounted to a 'revolution in accountability', to the extent that accountability structures ceased to be 'part of the system' and tended to become 'the system itself' (p. 460). At the core of this so-called 'revolution' was a gradual shift from traditional forms of professional accountability to forms of neo-liberalism based on partnership privatisation and strengthened corporate power. This entailed changes in the concept of accountability, from its meaning as 'communicative reason' to hierarchical answerability or 'holding to account', as well as in its structures, mechanisms and regulation. The 1997 Dearing Report in the UK endorsed the move towards increased accountability of higher education institutions to public bodies and stakeholders through public evaluations based on 'externally decided benchmarks and indicators' (Leisyte

et al., 2006, p. 27). Despite the good intentions peppering its development, the ensuing high-stakes accountability regime, with tight targets and monitoring and close focus on performance, affected research communities and research institutions in profound and often 'perverse' ways. Perhaps the most evident danger is that it becomes internalised in practice, to the extent that transparence is obfuscated by rhetorical fabrication; intrinsic excellence subsides in favour of an obsession with external effectiveness (e.g. meeting targets, creating wealth, gaining competitive advantages), and professionals are redefined as 'service deliverers'.

Institutionalisation of research evaluation

Over the past 20 years many countries have increased the amount of regulation around the assessment of research quality, not only for the purpose of management and quality assurance, but also to inform the distribution of public resources for research. To this aim, various dedicated structures have been created or consolidated and embedded in public research funding and governance systems. Table 2 illustrates this with examples from Europe, Australia and New Zealand. Some countries show a definite commitment to national assessments (with the corresponding structures already in place, e.g. the UK, New Zealand), while others only indicate a gradual shift towards more formal evaluations of research quality, volume, impact and relevance (Australia, France, Denmark, the Netherlands).

Universities UK, based on evidence from Geuna and Martin (2003), claims that 'the UK has developed one of the most advanced research evaluation systems in Europe' (UUK, 2006, p. 3). The extent to which an assessment regime such as the UK RAE has benefited policy-making and research is, however, open to debate, but I shall not engage with either argument here. In the case of education research in the UK, tighter assessment regimes seem to have stoked already smouldering criticisms of research communities and of their outcomes. The Research Assessment Exercise arguably contributed to stirring debates about the aims and value of education research and to maintaining a sense of unease and apprehension in research units throughout the country (particularly in those departments who fell immediately under the post-RAE 2001 funding threshold, such as those rated 3b, and were not beneficiaries of HEFCEs subject-related 'Capability Funding',[4] either—see Dadds & Kynch, 2003). The troubled public recognition of education research throughout the 1990s, following high-profile interventions from OFSTED (Tooley & Darby, 1998), the DfEE (Hillage *et al.*, 1998), Hargreaves (1996), Clarke (1998) or Blunkett (2000), made it difficult to campaign for the future of research in the field (see the activity of BERA and of its presidents and officers, as well as other interventions, such as Hammersley, 1997, analysed in Oancea, 2005).

The assessment regime has been, however, not only a source, but also a target of criticisms. Puxty *et al.* (1994) and Humphrey *et al.* (1996) described it as a 'surveillance system', with dire consequences for academic freedom, autonomy in research and vitality of research cultures throughout the country.[5] Such criticisms, and the tension and feelings of apprehension with which each assessment exercise had been

Table 2. Institutional formalisation of public research evaluation[a]

Country	Core public research funds allocated on the basis of	Institutional formalisation of public research evaluation	Examples of evaluations of public research for purposes of project and programme funding
Australia	National assessment of research quality and impact, still being trialled (announced in 2004 as a replacement for the previous research block funding scheme, based on educational and research volume); six-year cycle	Research Quality Framework (Australian Government, Department of Education, Science and Training): peer review and some metrics; 5-point scale	Competitive grants programmes—National Competitive Grants Program, Australian Research Council; National Health and Medical Research Council; Performance management frameworks (e.g. that of the Commonwealth Scientific and Industrial Research Organisation)
Denmark	Educational volume combined with performance-based assessments; move towards more formalised quality and relevance assessment, combining self-assessment with periodical external review	Danish Council for Research Policy's 'Tool for assessing research quality and relevance'; the Danish Councils for Independent and for Strategic Research; The Danish Research Coordination Committee	Competitive programmes—Danish Scientific Research Councils; the Danish National Research Foundation
France	Volume (four-year contracts) and performance—recent move towards allocation based on external evaluation	Institutionalisation of evaluation and articulation of criteria and procedures are in progress—the Agency for the Evaluation of Research and of Higher Education (created 2006) replaced the former CNER (National Committee for Research Evaluation), the CNE (National Committee for the Evaluation of Public Scientific, Cultural and Professional Establishments) and the MSTP (ministerial scientific, technical and pedagogical mission)	Project funding by the National Agency for Research (established in 2007)

Table 2. *(Continued)*

Country	Core public research funds allocated on the basis of	Institutional formalisation of public research evaluation	Examples of evaluations of public research for purposes of project and programme funding
The Netherlands	Traditionally, allocation based on educational volume. Recent move towards a hybrid system that includes quality, productivity, relevance and vitality assessment: 3-year cycle of internal evaluations and 6-year cycle of external evaluation	A Standard Evaluation Protocol between KNAW—Royal Netherlands Academy of Arts and Sciences; NOW—Netherlands Organization for Scientific Research; and VSNU—Association of Universities in the Netherlands, proposes periodical self-assessment plus external validation	Seven Research Councils under the umbrella of the NWO
New Zealand	Allocation based on educational volume was superseded by performance-related allocation informed by a national assessment exercise (2003, 2006, 2012)	Quality Evaluation (through peer review of 'evidence portfolios' submitted by institutions) for the distribution of the Performance-Based Research Fund (Tertiary Education Commission). In addition to QE (weighted at 60%), the allocation of the PBRF considers research degree completions (25%) and external research income (15%)	Competitive awards—Foundation for Research, Science and Technology; the Health Research Council of New Zealand, etc.
United Kingdom	Though traditionally based on educational volume, now mostly quality-related (ex-post evaluation through informed peer review). Since 1986, periodical national assessment exercise (1986, 1989, 1992, 1996, 2001, 2008)	Research Assessment Exercise, based mostly on peer review (Higher Education Funding Councils). Move towards a metrics-based system (from 2009)	Competitive grants from the Research Councils (dual support system), charities, British Academy, etc.; calls for tender from government departments and agencies

[a]This mode of presentation was inspired by Geuna and Martin (2003), p. 41.

expected, were very public and have created a rather paradoxical situation within the boundaries of official discourses of research assessment: the recent intensification of formal research assessment in the public domain, it seemed, had somehow failed its own standards of accountability, effectiveness and public endorsement. In the UK, this situation stimulated calls for reform of the assessment system (see House of Commons Science and Technology Committee, 2004; Royal Society, 2003; the Roberts Review, 2003).

Specialisation (professional, methodological) of research assessment and the emergence of assessment expertise

The aforementioned trend towards formalisation and institutionalisation of research evaluation cuts even deeper than the creation of dedicated structures, agencies and mechanisms and their eventual integration into national evaluation systems. It also entailed moves towards formally agreed explicit assessment criteria, standards and quality thresholds, and towards detailed reporting of research 'performance' at national and international level, accompanied by a complex statistical apparatus. Such approaches presuppose the commensurability and comparability of research 'performance' over time, between countries and between fields or subfields and modes of research, as well as widespread acceptance of evaluation as a legitimate component of the overall research environment and of the everyday life of academic communities. At the same time research assessment has become increasingly specialised, employing a plethora of evaluation experts supported by an ever-growing corpus of theoretical, methodological and administrative literature and documentation (including dedicated journals, handbooks and detailed guidebooks and regulatory documents).

Mismatch between economic & strategic and professional & academic indicators

However, the indicators used and the practices preferred in various evaluations of research are eclectic, reflecting the multitude of interests and demands that compete in setting the boundaries of accountability in research activities, rather than the nature of these activities themselves and of the particular forms of knowledge to which they contribute. The disjunction between economic and strategic criteria and indicators, and professional and academic ones, is particularly poignant. There is an implicit acceptance that indicators derive from incompatible discourses, and that therefore they cannot act as mediators of intrinsically problematic communication processes (e.g. between funders and researchers, between policy-makers and researchers, between administrators and academics).

Leaning towards peer review in micro evaluations and metrics in meso and macro

A trend, the consequences of which are not yet clear, is that of the polarisation of research assessment strategies according to the level of decision that they serve and

the scope of the evaluation process involved. As such, in evaluations at the micro level (e.g. for publication, awarding degrees, project funding, etc.) and in professional contexts, peer review is still the preferred and most trusted strategy, while at the level of external evaluations of research units or at the regional, national and international levels, quantitative procedures tend to take precedence (see Table 1). As we move from internal evaluations, professional contexts and small-scale assessments towards external evaluations, administrative and policy contexts, and large-scale reporting, productivity, competitiveness and medium-term impact become core values of research assessment. Even countries with long tradition in the use of peer review at meso and macro levels, such as the UK, are currently attempting to increase the reliance on metrics in research assessment, perhaps in the hope of balancing stronger intervention of the state in research arenas with what is presented as a gain in 'objectivity', consensus and legitimacy.

Diversification and variation of modes, criteria, procedures and agencies of research evaluation

As Table 2 indicates, most of the countries considered in the study reported here have a multitude of institutions organising and carrying out evaluations of research, as well as a growing diversity of procedures and criteria. As a consequence, a particular research project, research unit or network may find itself the object of a wide range of different evaluations, often simultaneous (Frederiksen *et al.*, 2003, p. 150). The recent consultation concerning the post-RAE 2008 assessments in the UK reflected this situation quite clearly; similarly, in France, the former National Council for Research Evaluation (now part of the Agency for the Evaluation of Research and of Higher Education) commented on the tendency towards multiple evaluations: 'having multiple evaluations should enrich our understanding, as it involves diverse judgements taking into account a variety of dimensions of research. However, there is no cross-evaluation in place to ensure the coherence of the results' (CNER, 2003, p. 49, my translation).

The strategies, criteria and potential impact of research evaluation are prone to a degree of variation as we move from one particular context to another. For example, the size of the national research community and the volume of research can play a significant role in determining the success of one or other research assessment model in a country. In countries where the research community is very small, its tight internal connectedness may undermine peer review processes (see, e.g. Frankel and Cave's (1997) comments on central European states) and a system of metrics may have wider support. The variation becomes, however, more problematic when the status and agendas of the agencies commissioning and carrying out the evaluation have a strong impact on its actual shape and direction. This may carry great potential for conflict, for example, if there is (unresolved) tension between the agenda of these agencies and that of those under evaluation or if the evaluations of different agencies favour and thus incentivise different approaches to research.

De-sensitivisation of research evaluation

Another source of concern is that the diversification of evaluation, described above, tends to remain external to the intrinsic diversity of disciplines, fields and modes of research. National and regional research assessment strategies have a rather low level of disciplinary sensitivity. This can partly be explained through shifts in the disciplinary space itself, such as the tendency towards cross-disciplinary hybridisation (see Dogan & Pahre, 1990; Thompson, 2004) and complex patterns of specialisation (i.e. narrower substantive coverage but wider theoretical and methodological palettes). Nonetheless, this may well have more to do with policy-originated pressures towards reducing the distance between the social and natural sciences in evaluation (see CNER, 2003, p. 26, for the French context), often in a poorly disguised attempt to extend standards and practices of assessment that are accepted in the natural sciences to the social sciences (see, e.g. the STEM-inspired metrics proposed for the UK system).

Even more striking is the low sensitivity of strategies and criteria for public assessment to differences between the modes of research under evaluation, to their nature, aims and claims—see, for example, the transfer of criteria from traditional definitions of 'basic' research towards applied and practice-based research (see also Oancea & Furlong, 2007). This is a cause for concern for many who believe that research evaluation systems should develop contextually, taking into account the social and organisational processes among which knowledge is generated (Hansson, 2002, p. 15).

The myth of the perfect indicator

Much of the public debate, surrounding the Research Assessment Exercise and its implications for education research in the UK, as well as the consultation concerning the announced shift to a metrics-based system (DfES, 2006a, b), has focused on choosing the right combination of techniques and indicators and weighting their benefits and limitations. Expert assessment through peer review and quantitative models based on a wide range of metrics (e.g. bibliometrics, technometrics, econometrics, sociometrics) were the two main contenders (perhaps combined, in an eclectic model). Other possible approaches, such as case studies and prevision models may have been mentioned, but never considered a viable alternative (one of the few examples of introducing case studies to public debate as a serious contender, on a par with bibliometrics, peer review and economic indicators, is a UK Evaluation Forum (2006) report on assessing medical research). Perhaps not surprisingly, even fewer mainstream solutions[6] ever made it into the public debate in any notable way; there has been little, if any, mentioning of, for example, historiographical or critical evaluation. Instead of exploring such alternatives, the problems identified in connection to each individual approach (e.g. bibliometrics or econometrics) prompted even more concentrated efforts to find and refine 'perfectly' reliable indicators (see the five possible models of research funding based on external research income indicators offered for discussion by the 2006 DfES consultation paper). With the focus firmly on the

technical detail of public assessments of research, the basic assumptions of the current assessment regime were in fact preserved.

This section will outline briefly the three main contenders—peer review, bibliometrics and econometrics—and the weighting of their advantages and disadvantages that was at the core of recent debates about research assessment in the UK.

Peer review

Peer review is still the dominant procedure in research assessment, particularly at the micro and meso levels. It has long established roots in academic practice and academic traditions and it benefits from the wide support and trust of the research community. Its origins can be traced back at least to the mid-seventeenth century, when, after the establishment of the Royal Society in Great Britain, the first editor of *Philosophical Transactions*, Henry Oldenburg, insisted that each article was reviewed by experts in the field prior to publication. The use of peer review was extended in the USA from publication decisions to the assessment of research proposals as early as the first half of the twentieth century, when the National Advisory Cancer Council (United States of America Congress, 1937, section 4) was authorised to evaluate applications for research funding (Cerroni, 2003; Hackett & Cubin, 2003, p. 3). Nowadays, peer review is perhaps the most widespread procedure for the assessment of academic papers, proposals and reports, and has developed into a variety of forms from traditional expert judgement to strategic evaluations (such as those developed by the European Commission for its Framework Programmes) and to 'informed peer review' (i.e. expert assessment that draws on a range of quantitative indicators).[7] To this day, the Royal Society remains one of the most fervent advocates of peer review; an enquiry group of the Royal Society concluded in 1995 that peer review is for science what democracy is for the good functioning of a country (Lachmann *et al.*, 1995, p. 2; Royal Society, 2006).

Advocates of peer review see it as a means to ensuring the stability of a field[8] and respecting 'the structure and culture of scientific communities' (CNER, 2003, p. 46, my translation). Their advocacy is based on a number of assumptions about the nature of peer review and of the relationships underpinning it, such as the benefits of a judgement based on competence and of parallel assessments of a single piece of research and the independence of professional judgement from political agendas and state control. Peer review is generally seen as offering the advantages of effectiveness and efficiency, shared accountability, trustworthiness, fairness, detail, flexibility and strategic value (Chubin, 1994; Kostoff, 1997, p. 9). Unlike other approaches to assessment, which work best at aggregate levels of data (e.g. citation analysis), systematic peer review 'remains the best available method for assessing the quality of individual pieces of work' (UK Evaluation Forum, 2006, p. 21). It also helps protect the autonomy of the research enterprise by creating a 'bumper' zone between academe and the policy arena (Hackett & Cubin, 2003, p. 13), and has great potential to stimulate excellence in research and filter out weak proposals and outputs. Based on such considerations, the British Academy stated, in its response to the DfES

consultation on the reform of research assessment, that 'simplistic metrics' are a threat to research in the humanities and social sciences, and that in these fields 'no alternative to peer review panels for assessing the quality and significance of outputs is credible' (British Academy, 2006, Para 2).

While peer review may be a very rewarding process, this depends, however, on several pre-conditions. First, it presupposes a degree of agreement on what constitutes good research, on the basis of which to overcome the tensions inherent both to research and to its evaluation, such as those between transparency and confidentiality; effectiveness and efficiency; sensitivity (to fields and modes of research) and selectivity; innovation and tradition; confirmed merit and promise; methodological rigour and social solidity and so forth (see Hackett & Chubin, 2003; Oancea & Furlong, 2007).

Second, in order to be effective at meso and macro level, peer review needs considerable resources to secure and manage the necessary investment of time at high levels of seniority. The administrative burden can be considerable. A recent RCUK consultation on the efficiency and effectiveness of peer review in informing funding decisions by the research councils put forward several proposals for improving cost-effectiveness (RCUK, 2006b, c), including larger multi-project awards; institutional-level quotes for grant applications; controlled resubmission (recycled proposals) and greater use of outline proposals. Also, the councils hoped, more emphasis may be put in the future on assessing the potential economic impact of the proposals to be funded (a 'red herring for the ESRC', according to BERA—St. Clair, 2007, p. 12). For many, the proposals rang alarm bells; they may amount to a 'dangerous economy', warned Sastry (2006), and end up sacrificing the distinctiveness of the Research Councils' role in the national research environment for the sake of limited cost reductions. In addition, they may unduly tip the balance away from quality and towards efficiency, to the detriment of research capacity and research cultures across the country (a point made in relation to education research by BERA—St. Clair, 2007, p. 12).

Third, for peer review to work optimally, its organisation and conduct need to be almost faultless. However, it is in fact a rather opaque process based on mutual trust rather than formal check points and errors or misconduct may pass unnoticed. For example, in choosing reviewers competence is key, but a 'network' effect may also play a role (CNER, 2003, p. 47). Further, peer review entails a risk of bias. Often the 'peers' are a part of the same competitive market or niche for research, and so impartiality is not always easy to achieve (and equally, it is not always easy to prevent voluntary or involuntary data or theory 'leaks'). The evaluation may be victim to a form of halo effect (e.g. the so-called 'Matthew' effect was identified, according to which senior researchers from prestigious institutions may be favoured in the review process, even when it is purportedly 'blind'—see Johnes, 1994, p. 213; Kostoff, 1997, p. 13; Cerroni, 2003, p. 6), or it may be a strongly conservative process, to the detriment of new and interdisciplinary approaches and fields of research. Also, expert evaluations may be used strategically to make education research more 'useful' to policy—sometimes to the detriment of its educational value (see, e.g. the push, in the USA, towards tackling 'what works'—type questions through experimental designs, particularly RCTs, the putative 'gold standard' in research).

Finally, the predictive power of peer review, in terms of citation success and potential impact, is low. For example, many studies, which compared the conclusions of peer review of articles proposed for publication with the number of citations received by the published article over a period of time, invariably found only weak correlations.

Nonetheless, many of the above concerns about peer review fail to fully acknowledge the fact that its main function is not to ensure stakeholder consensus or to provide accurate predictions of future performance, but to preserve professional autonomy and stimulate excellence from within the research communities.

Bibliometrics and technometrics

Bibliometrics (focused on publication outputs) and technometrics (focused on patents) are relatively new inventions.[9] They rely on quantitative indicators of research productivity, often drawn from large-scale dedicated databases (e.g. Thompson ISI,[10] Ulrich,[11] ERIC—Education Resources Information Center in the USA; BEI—British Educational Index in the UK; FRANCIS, a 1972-created French multidisciplinary database for the social sciences and humanities; other local and specialised databases,[12] etc.). The indicators can be descriptive (indicators of research productivity, volume and recognition, such as the number of publications or of citations) or relational (indicators of interaction, co-operation and cohesion of a field or research community, e.g. co-authorship, co-citation analysis, co-classification, co-nominalisation) (for a detailed presentation see Gauthier, 1998). Nowadays, the field of bibliometrics is in constant international expansion. The four main publications in the field (*Scientometrics*, *Research Policy*, *Science and Public Policy* and *Research Evaluation*) are subscribed by most large academic libraries. Research evaluations across levels, fields and countries are increasingly incorporating such indicators, particularly in STEM subjects (as Braun *et al.* anticipated in 1995), but also in the social sciences, though to a smaller extent (see Glänzel, 1996), while attempts are being made to develop new indicators that overcome some of the weaknesses traditionally associated with bibliometrics.

The rise of bibliometrics was historically fuelled by the aspiration towards objectivity and cost-effectiveness in research assessment. Bibliometric indicators were deemed a cost-effective instrument for policy, for strategic and tactical decisions at all levels, and for the positioning of an institution within a field or in a research environment. They also provided a neat and clear tool for mapping the research landscape of a discipline or a country and for creating (comparable) profiles of research networks (for instance, RCUK claims that progress has been made by the research councils and the Office of Science and Innovation towards using bibliometric analyses at high levels of aggregation—'super unit of assessment'—RCUK, 2006a, p. 3). Such maps and profiles not only describe concentrations and mobility of expertise, but they can also help identify patterns of collaboration, rivalry and competition in a field. Such expectations make bibliometrics very attractive to governmental and transnational (e.g. European, OECD) evaluations of research. For example, the response of the government to the 2006 DfES consultation on the reform of higher education

research assessment and funding put forward plans to use bibliometric indicators as a core component of the metrics-based model for science, engineering, medicine and technology, but for the time being to contend with more traditional indicators and peer review in all other disciplines, for which reliable bibliometric measures 'were yet' to be developed (DfES, 2006b). This was despite concerns from, for example, the Royal Academy of Engineering that the extensive use of bibliometric output data would be detrimental to a high proportion of 'mode 2' applied research in the field (Royal Academy of Engineering, 2006, p. 5).

As was the case with peer review, the effective use of bibliometrics and technometrics also relies on a number of assumptions, such as: that the number of publications and patents is a valid estimate of R&D intensity in a discipline or a research unit; that the number of citations is a valid indicator of the impact or importance of a publication or piece of research; that mutual citations and co-citations are an estimate of intellectual connections between authors, research units and fields of research, and so on (see Narin, 1994; Kostoff, 1998).

However, caution is needed when deciding whether bibliometrics offers the best set of indicators in particular exercises of research assessment or whether it should be replaced by, or at least combined with, peer review or other procedures (e.g. external income metrics). If either of the above assumptions is challenged or if the indicators are misused, a number of problems emerge (Adam, 2002). For example, though bibliometric indicators are essentially quantitative and do not purport to offer more than information on productivity and popularity/citation impact, sometimes an abrupt jump is made from quantitative description to quality judgments, as it was the case with some of the 1990s–early 2000s criticisms of education research in England. Similarly, flawed is the use of journal impact factors to estimate the research performance of individual authors (Seglen, 1997; CNER, 2003) and even departments; citation counts are meaningful at higher levels of aggregation (Sastry & Bekhradnia, 2006), and, even so, they are not a proxy of research quality.

Using citation indexes for purposes outside those originally intended is problematic in any case (see Johnes, 1994, for an early analysis), and this may be one of the main weaknesses of hastily used bibliometric measures. For example, a growing body of literature points out the shortcomings of such unwarranted use of, in particular, the ISI database. Such studies comment on a range of issues: its Anglo-centrism (Johnes, 1994; Rey-Rocha *et al.*, 2001); its incompleteness, classification errors and disproportion in the coverage of different disciplines (Seglen, 1997; CNER, 2003); its selection criteria, which tend to work to the disadvantage of emerging fields and beginning researchers (van Raan, 1997); its exclusions (books, conferences, electronic publications, 'grey' literature, etc.—CNER, 2003); the lack of disciplinary sensitivity of its formulae for calculating indicators (e.g. using a two-year period as the basis for calculating the impact factor, and thus failing to take into account the variable speed of citations accumulation and of obsolescence, as well as the specific patterns of publication—for example,. favouring monographs, fast succession of articles or reports, in different fields—Moed *et al.*, 1995; Cerroni, 2003), and its indexing inconsistencies (e.g. in tracing changes of title and publisher for a given

periodical; or in distinguishing between authors with identical surnames—CNER, 2003). Such problems are compounded by the natural variation in the citation behaviour of individual authors and by the inadequacies that may arise in the citation process: abusive or strategic citations (up to 10% of the total number of citations, according to Marx *et al.*, 1999); selective citations, with recent works and literature reviews more likely to be cited than classics whose work has already become part of the 'common knowledge' of a field (up to 40% of actual sources remaining uncited—MacRoberts & MacRoberts, 1989; Kostoff, 1998; Marx *et al.*, 1999; Patsopoulos *et al*, 2005); erroneous citations (10–50% of total citations—MacRoberts & MacRoberts, 1989); self-citations (up to 10%—Marx *et al.*, 1999); negative or critical citations; non-substantive and token citations, etc. Finally, publication patterns may vary for reasons that are not performance- or impact-related; for example, from one disciplinary or national context to another (MacRoberts & MacRoberts, 1989), from one period of time to another (Seglen, 1997), from one type and topic of research to another (the number of citations was found by Marx *et al.*, (1999) to vary in direct proportion with the degree of currency of a research topic and in inverse proportion with the degree of specialisation), as well as idiosyncratically (as an issue of the personal style of researchers—Lindsey, 1978).

Economic and financial metrics

This is a highly specialised field and this article is not the place for going into a great depth of detail on such metrics. Nonetheless, due to the high profile they currently enjoy in the context of the recent debates about the future of research assessment in the UK, it is worth commenting on their benefits and limitations in relation to education research.

Generally, economic analyses informing research evaluation can rely on input or output measures, and support *ex post* judgements of performance, as well as forecasts and feasibility studies (Capron, 1992). The type and weight of the indicators and the formulae used in each case may differ considerably. Two of the most common approaches to such analyses are: cost-benefit and rate of return studies (employing techniques that compare financial estimates of net costs and benefits in relation to an initial time T_0); and production analyses (which attempt to provide an estimate of the value added by research on the basis of regression techniques using time series of outputs, capital, labour costs and direct research expenses) (see e.g. Mansfield, 1980; Kostoff, 1997). A distinctive category of economic metrics are those concerned with estimating the economic and financial impact of research (rates of return). Various macro and micro economic models have been developed in this respect (e.g. Buxton *et al.*'s (2004) model for medical research, adopted by UK Evaluation Forum, 2006, covers cost savings, trained workforces, commercial developments and wider technological and societal benefits); none has, however, gained general support.

Economic and financial analyses can be an effective means of monitoring investment in research and informing funding decisions, particularly in relation to technological research. However, they have limited accuracy in relation to 'blue-sky', long-term,

historical and theoretical research (including research in the humanities and the social sciences), that is, research where it is extremely difficult to anticipate and accurately estimate costs, timelines and results. Further, they presuppose a rather linear relation between research quality, productivity and added value, and work with an incomplete model of research and of research policy that does not account fully for their cultural, social, professional and practical dimensions. Through their emphasis on what is (economically) measurable and tangible, economic indicators can be rigid and conservative, and their use—while striving for accuracy—may leave out societal benefits that are in fact crucial to the aims and nature of the research under evaluation (including here educational benefits—see the Dutch research assessment model's efforts to incorporate qualitative assessments of societal impact) (UK Evaluation Forum, 2006, p. 32). Finally, any economic analysis for research assessment purposes will by confronted with classical problems of statistical reporting, such as the distinction between correlation and causation, and of accounting, such as ways of costing complex and developing processes, the measurement and quantification of 'intangibles' or the attribution of impact (see Mansfield, 1980; Kostoff, 1997; Piric & Reeve, 1997; Sastry & Bekhradnia, 2006; UK Evaluation Forum, 2006).

Like bibliometric indicators, economic and financial metrics are also subject to strategic use by those evaluated. They are likely to impact on the behaviour of those evaluated, by virtue of their status as privileged measures of performance informing funding and resource allocation decisions in research. Individual and collective actors can therefore be expected to develop strategies to improve their ranking along each indicator rather than their actual activity. This is the so-called Goodhart's (1975), according to which pressures for control and target-setting can lead to the collapse of previously valid measures and thwart statistical regularities (Johnes, 1994; McIntyre, 1999).

Economic and financial metrics, and in particular measures of external research income (believed appropriate to STEM subjects), formed the core of the proposed move towards a metrics-based funding system in the UK. However, it is essential not to overlook the shortcomings of financial metrics, and the fact that each approach to research assessment may have its own applications that cannot be generalised to the entire system.

A Higher Education Policy Institute report (Sastry & Bekhradnia, 2006) explored the financial implications of the DfES proposals in more detail and concluded that, while the proposed metrics-based system would have the advantages of increased transparency, focus on research that is of public interest, and greater concentration of funding, it would come with considerable problems without necessarily solving the old, RAE-related, ones.[13] For example, the authors argue, linking QR funding to grants income would be a costly process; the costs will be augmented by the need to finance extra data collection, expert check-ups and administration, as well as a separate arts, humanities and social sciences exercise. Further, the claimed stability of the system may not materialise, as metrics can be volatile and the new funding models (by putting pressure on Research Councils funding) can generate significant financial fluctuations for the institutions concerned. Finally, the proposed system is likely to

impact both on the nature of the research carried out (e.g. inhibiting research that is irrelevant or hostile to the interests of the major funders; changing the nature of the bids put forward by institutions, perhaps towards more expensive projects, with the risk of 'volume inflation'[14]) and on the non-research activities of an institution (e.g. potential negative effects on teaching, on employment patterns, on university policies) (Sastry & Bekhradnia, 2006, pp. 5–14; see also Royal Society, 2006, p. 5).

Questionable technicalities? The Procrustean drift in research assessment

The proposed reforms of research assessment in the UK (DfES, 2006 a, b; HM Treasury, 2006a, b) are based on the hope that the 'right combination' of (quantitative) indicators is achievable and able to solve the problem of assessing research across disciplines in a cost-effective manner. The forthcoming RAE (2008) was expected to be an improvement on its predecessors by achieving better specification of criteria and better weighting of standards and objects of assessment and by making greater use of metrics where thought appropriate (UK Funding Bodies, 2005). Also, the RCUK proposals concerning peer review (2006b, c) went to great lengths to control and limit the conduct of what is traditionally a form of professional self-regulation.

An often neglected by-product of these processes, and of the trends outlined in the first half of this article, is that the problems and dilemmas of research evaluation are gradually relegated to the domain of the technical. What is more, the public meaning of the 'technical' itself has been shrinking over the past decades, to an ever-closer approximation of instrumentalism. As argued elsewhere in relation to applied and practice-based research (Oancea & Furlong, 2007), the instrumentalisation of assessment also came with a narrowing of the 'official' concept of *quality*[15] to 'measurable performance', of *research* to 'production and delivery' and of *research assessment* to 'quantification'. The plethora of standards, shopping-list criteria and 'cut-off point' guidelines currently in circulation illustrates it sufficiently. This is what I shall call the 'Procrustean' drift in public research assessment in Britain.

The story of Procrustes, gruesome as it may be, is worth a brief retelling. Procrustes ('The Stretcher') is mentioned in the myth of Theseus as a robber living in Attica, who enticed passers-by with deceptive hospitality and after having them lay down to rest on his iron bed, would make them 'fit' the length of the bed by either 'stretching' them or chopping off their limbs.[16] It is the contrast between the rhetoric of accommodation and the outcome of arbitrary, forceful conformity that makes the analogy to Procrustes' bed appropriate to the above-mentioned narrowing of 'quality' and 'research' to fit the requirements of a technically-defined mode of research assessment.

In other words, while public discourse insists on 'ensuring that appropriate measures of excellence are developed which are sufficiently wide as to capture all types of research, including practice-based research, applied research, basic/strategic research, interdisciplinary research' (UK Funding Bodies, 2005, Para 3c), the indicators proposed for the metrics-based system are hardly sensitive to this diversity. Further, even in relation to the old RAE, many critics have commented on what they

perceived as 'undue emphasis' 'on academic publications rather than applied work',[17] and a 'bias against multi-disciplinary research in favour of theoretical and against applied research'.[18] It was not the actual acts of evaluation undertaken by members of the RAE panels that were being questioned here; rather, the concern was with the overall framework of the official discourse that constrained and interpreted their work. Far from breathing fresh air into public research policy, the recent proposals for the reform of research assessment may end up reinforcing those constraints even more tightly.

Restoring the socio-cultural, historical and philosophical nature of research assessment

The preoccupation with fine-tuning the techniques obscures several fundamental questions: is research evaluation indeed a matter of getting the technologies 'right' (i.e. of making them effective, controllable and in perfect fit with their object)? What are the dangers entailed by the focus on the fine-tuning of techniques and criteria to objects (and the reverse), rather than on the limits of the discourse from which these techniques emerge?

Recent developments in research evaluation have made it clear that a multiplicity of interests are at stake (academia, policy, practice, public, media, market, etc.) and that there are tensions between the forms and criteria that each of these communities favours. The value and usefulness of each procedure/strategy for research assessment is therefore contextual. It has to do with the characteristics and goals of the process of evaluation and of the evaluator, as well as with the nature and claims of the research under evaluation and with the particularities of the publication/communication channel used. Moreover, as hinted above, the procedures and indicators of research evaluation are not purely technical matters, but they are also 'cultural documents, imperfectly embedded in certain historical traditions' (Thackray, 1977, p. 20). They need therefore to be read in their historicity and contextuality and in relation to the particular assumptions about knowledge and knowledge production that they embody. Recent public debates about research assessment in the UK fuelled by the mismatch policy-academe and by the financial, logistic and prestige-related rivalries between research communities and institutions, tended to play by the rules of administrative and political discourse and neglect the socio-cultural, historical and philosophical dimensions of evaluation. The final part of the article suggests that a restoration of these last dimensions of research assessment can help a dialogue between research communities and the policy and practice arenas.

If assessment is defined exclusively by its use as a means towards pre-determined ends (distribution of funding and holding to account), the consequences of this are likely to be the subordination of appraisal to measurement and of process to output—i.e. falling into the extremes of what Stake (2001) calls 'criterial thinking'. This may lead to the alienation of research evaluation from interpretations of good practice shared within professional communities.

Rather than cutting things to shape in Procrustean manner, that is, investing our time in the search for the perfect indicator, a more profitable way forward may involve shifting our overall way of conceiving of assessment and of its roles, and our understanding of quality, from a narrowly defined technical framework towards a wider (historical and philosophical) understanding. Such a shift would draw on alternative discourses about the assessment of education research. I suggested elsewhere (Oancea, 2006; Oancea & Furlong, 2007) that one way of reframing the problem may be by interpreting 'assessment' as *deliberation and judgement*, and 'quality' as *excellence or virtue*, in a classical (Aristotelian) sense of the terms. Table 3 suggests alternatives to some of the more established concepts and interpretations in current official discourses.

From such a perspective, 'excellence' in education research can no longer be reduced to the (measurable) attainment of standards of performance: it involves a synergy of epistemic, technical and practical qualities that resists instrumentalisation.

Further, the concept of technique suggested above is no longer in conflict with demonstrative knowledge (*episteme theoretike*), practical and ethical deliberation (*phronesis*) (Aristotle, 1915/1975) and professional definitions of 'good research'. It is a concept akin to Aristotle's wider understanding of *techne*, which was in fact multilayered, and thus it is not fully captured by today's commonsense concepts of technical skill or craft or by the instrumental discourse of techniques, targets and performance measures. For Aristotle, its domain covered, for example, material production or fabrication (e.g. building and shoe-making), skilful or expert intervention (e.g. military strategy or medicine), as well as performance (sports and arts). Thus extended, *techne* can incorporate an element of chance, timing (*kairos*), and ambiguity—thus, a degree of responsiveness to the conditions of one's intervention. This opens up the possibility of *techne* linking not only to *episteme*, at one, more general, level, but also to *phronesis*, at another, which is closer to the richness of the particular. In Aristotle's actual usage of the term, thus, *techne* encompasses activities, the characteristic features of which are scarcely captured by his 'official concept' (Dunne, 1993), and which bring into play, in fact, a quite different conceptual paradigm bearing strong family resemblances with *phronesis* (Dunne, 1993, p. 261).

Table 3. Discursive alternatives to instrumental accounts of research assessment

Current discourses	Alternative discourses
Hierarchical relationship between modes of research	Complex entanglement of research and practice and different modes of knowledge
Quality assurance and quality assessment	Nurturing excellence/ virtue (epistemic, technical and phronetic)
Quantification, measurement and ranking of performance	Deliberation and judgement
Assessment techniques unquestioningly produce externally specified outputs	Assessment techniques help research communities to gain increased control over the contingencies of their practice

Such an understanding of *techne* is open towards a concept of professional activity (research and research evaluation, included) as the conjunction of technique, praxis and theory, in which skill is balanced by knowledge and judgement and training by experience. If this is the case, then the recent proposals for reforming research assessment in the UK may simply be trying to squeeze complex practices into the straitjacket of externally defined, cost-efficient and easily quantifiable performance measures, in which case they will fail to either build more democratic forms of assessment or nurture a healthy research environment in the UK.

Concluding comments

As the myth goes, Procrustes got his comeuppance at the hands of Theseus, who 'fit' him to his iron bed by chopping his head off. Which only stands to show that, even in the hands of heroes, blunt, unflinching standards lead to mutilation.

This article argued that further tweaking of the established assessment procedures (such as trying to produce perfect criteria) is unlikely on its own to solve the current research assessment conundrum. The way forward suggested here is not a case of simply reversing Procrustes's approach and perhaps starting instead to 'stretch' and 'cut' the 'bed' (i.e. the assessment indicators and strategies) to adjust its length to that of the person (i.e. the research under assessment), while preserving a rigid one-to-one correspondence. Rather, the approach that is required involves changing our overall understanding of the relationship between the processes and the substance of assessment, accompanied by a discursive (Protean[19]) opening towards accepting diversity, hybridisation and versatility in research.

Acknowledgements

This article draws on work undertaken with the financial and logistic support of the Oxford University Department of Education and the Oxford Institute of Ageing. It was anticipated by presentations given to the British Educational Research Association conference, Warwick, 2006 (Oancea, 2006) and the European Educational Research Association conference, Geneva, 2006. The insightful comments and questions of the audiences to these events are gratefully acknowledged.

Notes on contributor

Dr Oancea is a research fellow working on Philosophy of Educational Research, Research Policies, and Research Evaluation in the Oxford University Department of Education and the Oxford Institute of Ageing. She is also part of the ongoing ESRC/TLRP review of the epistemological basis of educational research findings and of the independent Nuffield Review of 14–19 Education and Training. Her recent work covered: domains of quality in applied and practice-based research (Economic and Social Research Council, 2004–2005); education research capacity in the UK (British Educational Research Association, 2003–2004);

practices and criteria for research evaluation; and criticisms of education research.

Notes

1. The recent British Academy working group on peer review report, launched in September 2007, also recommended that metrics take a back seat in social sciences and humanities, and that their peer review, with proper training and appropriate consideration of costs, remains the main form of research assessment (British Academy, 2007).
2. See, for example, the 1991 White Paper '*Higher Education—A New Framework*' (Cm 1541) and the 1992 Further and Higher Education Act.
3. See also the reporting of the performance of the 'research base' in the PSA target metrics, covering inputs (including expenditure on research), outputs (including people and publications), outcomes (research recognition, citations, training and research quality), productivity—financial (outputs and outcomes related to inputs) and labour (outputs and outcomes related to other measures), and people (research capacity) (OSI—DTI, 2007).
4. 'Capability funding' was created for subjects with emerging research cultures on condition of submission of 'acceptable' research strategies: nursing, other studies and professions allied to medicine, social work and art and design.
5. Some argue that education was one of the worst hit subjects post-RAE 2001. Figures produced by the Association of University Teachers show that education, together with environmental studies and business and management studies, topped the charts in terms of percentage of units to receive no funding after RAE 2001 (60.09%) (AUT, 2003). The figure must be interpreted with caution, though, as it does not take into account the number of submissions and the characteristics of the field and of the infrastructure and staff involved. For an analysis of the results of RAE 2001, see Oancea, 2004.
6. For example, the Witness Seminars approach based on methodology developed by the Institute of Contemporary British History (UK Evaluation Forum, 2006, p. 19).
7. For an examination of different forms of peer review, see also Gibbons and Gheorghiou, 1987.
8. Hackett and Chubin (2003, p. 10) see peer review as the epitome of Kuhn's 'essential tension' between originality and tradition.
9. For example, the more comprehensive term 'scientometrics' entered public use only after the establishment of the eponymous journal in 1979.
10. A selective database that relies on criteria such as citation impact to index a limited number of publications of the almost 70,000 relevant periodicals that are currently available worldwide. It was initiated in Philadelphia in the 1960s (1963, for natural sciences; 1975, for humanities; 1979, for social sciences) and indexes over 5000 science journals, 1700 social science journals (plus references from around 3300 journals of cross-disciplinary relevance) and 1000 (plus 7000 cross-references) arts and humanities periodicals.
11. A comprehensive bibliographical database of periodicals created in 1932 in the USA and based on criteria such as periodicity and audience rather than quality and citations. The main target audience is composed of libraries and publishing houses rather than research units, and as a consequence the database includes mainly bibliographic and commercial information.
12. See also the debates around proposals for the creation of a European Social Science Citation Index (Gogolin *et al.*, 2003; Botte, 2004) and of a European Citation Index for Humanities (European Social Foundation, http://www.esf.org/).
13. Bill Rammell (2006) responded to the HEPI report by arguing that it oversimplified government's proposals.
14. In its response to the DfES consultation, the Royal Academy of Engineering also expressed concern that 'over-reliance on metrics based on institutions' income … has the potential to reward expensive research rather than good research' (2006, p. 5). By contrast, the Wellcome

Trust felt that external research income was 'an effective measure of excellence' in biomedical research, due to the particularities of the field; however, it cautioned against a resource allocation model that 'value[s] different funders in different ways' (Wellcome Trust, 2006, p. 3).

15. Earlier in this article, I commented on the risks entailed by confusing productivity with quality and using indicators of volume as proxies for excellence. Similar risks are connected to definitions of quality in terms of natural sciences-inspired scientificity or in terms of economic potential, user impact (reduced to observable and attributable improvement in practice) or citation impact.
16. Plutarch *Lives. Life of Theseus.* Translated by John Dryden; Diodorus Siculus, *Library of History* (Books III–VIII). Translated by Oldfather; C. H. Loeb Classical Library Volumes 303 and 340. London, William Heinemann, 1935. Book IV.59, 59.5.
17. Lord Rees of Ludlow, Lords Hansard, 30 Mar 2006: Column 950.
18. Baroness Sharp of Guildford, Lords Hansard, 30 Mar 2006: Column 961.
19. Proteus, a sea-god tending the flocks of Poseidon, had 'the gift to change and change again in many forms' (Ovid, *Metamorphoses*, 2.6.) to escape the necessity of prophesying. He is a symbol of versatility, hybridisation, diversity and resistance to moulding, but also of ambiguity and ever-elusive truth.

References

Adam, D. (2002) Citation analysis: the counting house, *Nature,* 415, 726–729.

Aristotle. (1915/1975) Ethica Nicomachea, in: W. D. Ross (Trans.) *The works of Aristotle* (vol. 9) (Oxford, Oxford University Press).

AUT (2003) *The risk to research in higher education in England* (London, Association of University Teachers).

Ball, P. (2005) Index aims for fair ranking of scientists, *Nature,* 436, 900.

Blunkett, D. (2000) Influence or irrelevance: can social science improve government? Speech made by David Blunkett, Secretary of State for Education and Employment, to a meeting convened by the Economic and Social Research Council on 2 February 2000, *Research Intelligence,* 71, 12–21.

Botte, A. (2004) Achievement or performance: observation of productivity of educational research by bibliometric tools. A state-of-the-art-report, paper presented at the *European Conference on Educational Research,* University of Crete, 22–25 September.

Braun, T., Glänzel, W. & Grupp, H. (1995) The scientometric weight of 50 nations in 27 science areas, 1989–1993: Part I. All fields combined, mathematics, engineering, chemistry and physics, *Scientometrics,* 33(3), 263–293.

British Academy (2006) *Response to the DfES consultation on the reform of higher education research assessment and funding.* Available online at: http://www.britac.ac.uk/reports/rae-2006/response.html (accessed 30 April 2007).

British Academy (2007) *Peer review: the challenges for the humanities and social sciences.* Report of the working group chaired by A. Weale, September (London, British Academy).

Buxton, M., Hanney, S. & Jones, T. (2004) Estimating the economic value to societies of the impact of health research: a critical review, *Bulletin of the World Health Organisation,* 82, 733–739.

Capron, H. (1992) *Economic quantitative methods for the evaluation of the impact if R&D programmes.* EUR 14864 EN, Monitor-Spear Programme, CEC, SRD DG XII. (Brussels, EC).

Cerroni, A. (2003) Socio-cognitive perverse effects in peer review. Reflections and proposals. *Journal of Science Communication,* 2(3). Available online at: http://jcom.sissa.it/archive/02/03/F020305/ (accessed 21 March 2005).

Chubin, D. E. (1994) Grants peer review in theory and practice, *Evaluation Review,* 18(1), 20–30.

Clarke, C. (1998) Resurrecting educational research to raise standards: statement from the new minister responsible for research, *Research Intelligence,* 66, 8–9.

CNER—Comité National D'évaluation de la Recherche (2003) *Évaluation de la recherche publique dans les établissements publics français* [Evaluation of public research in French public establishments] (Paris, France).

Council of the Royal Society (1995) *Peer review—an assessment of recent developments* (London, Royal Society).

Dadds, M. & Kynch, C. (2003) The impact of the RAE 3b rating on educational research in teacher education departments, *Research Intelligence*, no. 84.

Dearing Report (1997) *Sir Ron Dearing. National committee of inquiry into higher education.* The Dearing Report, 23 July.

DfES (2006a) *Reform of higher education research assessment and funding. A consultation.* Consultation document, 13 June 2006. Available online at: http://www.dfes.gov.uk/consultations/downloadableDocs/consultationDoocument%20jcutshall2.doc (accessed 30 April 2007).

DfES (2006b) *DfES consultation on the reform of higher education research assessment and funding: summary of responses.* Available online at: http://www.dfes.gov.uk/consultations/downloadableDocs/RAE%20response%20summary%20250107.doc (accessed 30 April 2007).

Dogan, M. & Pahre, R. (1990) *Creative marginality. innovation at the intersections of social sciences* (Oxford, Westview Press).

Dunne, J. (1993) *Back to the rough ground: 'Phronesis' and 'Techne' in modern philosophy and in Aristotle* (London, University of Notre Dame Press).

ESRC (2006) *DfES consultation on reform of higher education research assessment and funding. ESRC questionnaire.* Available online at: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/dfes_consultation_questionnaire_tcm6-16002.doc (accessed 30 April 2007).

Frankel, M. S. & Cave, J. (Eds) (1997) *Evaluating science and scientists: an east-west dialogue on research evaluation in post-communist Europe* (Budapesta, CEU Press).

Frederiksen, L. F., Hansson, F. & Wenneberg, S. B. (2003) The Agora and the role of research evaluation, *Evaluation,* 9(2), 149–172.

Gauthier, E. (1998) *Bibliometric analysis of scientific and technological research: a user's guide to the methodology* (Canada, Observatoire des Sciences et des Technologies, Statistics Canada).

Geuna, A. & Martin, B. R. (2003) University research evaluation and funding: an international comparison, *Minerva,* 41(4), 277–304 (original work published 2001).

Gibbons, M. & Gheorghiou, L. (1987) *Evaluation of research: a selection of current practices* (Paris, OCDE).

Glänzel, W. (1996) A bibliometric approach to social sciences: national research performances in 6 selected social science areas, 1990–1992, *Scientometrics,* 35(3), 291–307.

Gogolin, I., Smeyers, P., Garcia Del Dujo, A. & Rusch-Feja, D. (2003) European social science citation index: a chance for promoting European research? Roundtable, *European Educational Research Journal,* 2(4), 547–593.

Goodhart, C. A. E. (1975) *Money, information and uncertainty* (London, Macmillan).

Hackett, E. J. & Chubin, D. E. (2003) Peer review for the 21st century: applications to education research, paper for a *National Research Council (SUA) workshop,* Washington, DC, February 25.

Hammersley, M. (1997) Educational research and teaching: a response to David Hargreaves' TTA lecture, *British Educational Research Journal,* 23, 141–162.

Hansson, F. (2002) Best practice in research evaluation? How to evaluate and select new scientific knowledge by introducing the social dimension in the evaluation of research quality, paper for the *European Evaluation Society conference*, Seville, Spain, October 10–12.

Hargreaves, D. H. (1996) *Teaching as a research-based profession: possibilities and prospects* (Teacher Training Agency Annual Lecture) (London, Teacher Training Agency).

HEFCE/AHRC (2006) *Use of research metrics in the arts and humanities.* Report of the Expert Group set up jointly by the Arts and Humanities Research Council and the Higher Education Funding Council for England. Available online at: http://www.hefce.ac.uk/research/assessment/reform/metrics.doc (accessed 7 May 2007).

Hillage, J., Pearson, R., Anderson, A. & Tamkin, P. (1998) *Excellence in research on schools* (London, DfEE).

HM Treasury (2006a) *Chancellor of the exchequer's budget statement,* 22 March 2006. Available online at: http://www.hm-treasury.gov.uk/budget/budget_06/bud_bud06_speech.cfm (accessed 30 April 2007).

HM Treasury (2006b) *Science and innovation investment framework 2004–2014: next steps,* March 2006. Available online at: http://www.hm-treasury.gov.uk/media/D2E/4B/bud06_science_332v1.pdf (accessed 30 April 2007).

House of Commons Science and Technology Committee (2004) *Research assessment exercise: a re-assessment* (Eleventh report of session 2003–2004) (Norwich, TSO).

Humphrey, C., Moizer, P. & Owen, D. (1996) Questioning the value of the research selectivity process in British university accounting, *Accounting, Auditing and Accountability Journal,* 8(3), 144–164.

Johnes, G. (1994) Research performance measurement: what can international comparisons teach us?, *Comparative Education,* 30(3), 205–216.

Johnstone, D. B. (1998) *The financing and management of higher education: a status report of world-wide reforms* (Washington, DC, World Bank).

Kostoff, R. N. (1997) *The handbook of research impact assessment* (7th edn) (DTIC Report No ADA296021) (Arlington, Office of Naval Research).

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation—comments on theories of citation?, *Scientometrics,* 43(1), 27–43.

Leisyte, L., de Boer, H. & Enders, J. (2006) England—the prototype of the evaluative state, in: B. M. Kehm & U. Lanzendorf (Eds) *Reforming university governance. Changing conditions for research in four European countries* (Bonn, Lemmens), 17–52.

Lindsey, D. (1978) The corrected quality ratio: a composite index of scientific contribution to knowledge, *Social Studies of Science,* 8(3), 349–354.

MacRoberts, M. H. & MacRoberts, B. R. (1989) Problems of citation analysis: a critical review, *Journal of the American Society for Information Science,* 40(5), 342–349.

Mansfield, E. (1980) Basic research and productivity increase in manufacturing, *American Economic Review,* 70, 863–873.

Marx, W., Wanitschek, M. & Schier, H. (1999) Scientometric on fullerenes and nanotubes, *Condensed Matters News,* 7(4), 3–7.

McIntyre, M. E. (1999, November 4) Audit, education and Goodhart's law or taking rigidity seriously. TV programme, *BBC,* Transcript available online at: http://www.atm.damtp.cam.ac.uk/people/mem/papers/LHCE/dilnot-analysis.html (accessed 25 March 2005).

Moed, H. F., Burger, W. J. M., Frankfort, J. G. & van Raan, A. F. J. (1985) A comparative study of bibliometric past performance analysis and peer judgement, *Scientometrics,* 8(3–4), 149–159.

Narin, F. (1994) Patent bibliometrics, *Scientometrics,* 30(1), 147–155.

Oancea, A. (2004) The distribution of educational research expertise in the UK—finding from the analysis of RAE 2001 submissions (I–II), *Research Intelligence,* 87, 88.

Oancea, A. (2005) Criticisms of educational research: key topics and levels of analysis, *British Educational Research Journal,* 31(2), 157–183.

Oancea, A. (2006) Procrustes or Proteus? Towards a philosophical dimension of research assessment, paper to the *British Educational Research Association annual conference,* Warwick, 6–9 September.

Oancea, A. & Furlong, J. (2007) Expressions of excellence and the assessment of applied and practice-based research, *Research Papers in Education,* 22(2), 119–137.

OSI—DTI (2007) *PSA target metrics for the UK research base.* Department of Trade and Industry, Office of Science and Innovation, March 2007, Leeds, Evidence Ltd.

Patsopoulos, N. A., Analatos, A. A. & Ioannidis, J. P. A. (2005) Relative citation impact of various study designs in the health sciences, *Journal of the American Medical Association,* 293, 2362–2366.

Piric, A. & Reeve, N. (1997) Evaluation of public investment in R&D–towards a contingency analysis, *OECD Conference on Policy Evaluation in Innovation and Technology Report,* Paris, 26–27 June, 49–64.

Puxty, A. G., Sikka, P. & Wilmott, H. C. (1994) Systems of surveillance and the silencing of UK academic accounting, *British Accounting Review,* 26(2), 137–171.

Rammel, B. (2006) *Speech to the HEPI conference 21 June on research funding and assessment beyond 2008.* Available online at: http://www.hepi.ac.uk/downloads/BillRammell.doc (accessed 30 April 2006).

Ranson, S. (2003) Public accountability in the age of neo-liberal governance, *Journal of Education Policy,* 18(5), 459–480.

RCUK (2006a) *DfES consultation on the reform of higher education research assessment and funding. Submission from Research Councils UK.* Available online at: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/RCUK_RAE_Consultation_Response_tcm6-16974.pdf (accessed 30 April 2007).

RCUK (2006b) *Efficiency and effectiveness of peer review project consultation.* Available online at: http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/prconsultation.pdf (accessed 30 April 2007).

RCUK (2006c) *Report of the research councils UK efficiency and effectiveness of peer review project.* Available online at: http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/rcukprreport.pdf (accessed 30 April 2007).

Rey-Rocha J., Martin-Sempere, M. J., Martinez-Frias, J. & Lopez-Vera, F. (2001) Some misuses of journal impact factor in research evaluation, *Cortex,* 37(4), 595–597.

Roberts, G. (2003) *Review of research assessment.* Report by Sir Gareth Roberts to the UK funding bodies. May 2003. Available online at: http://www.ra-review.ac.uk/reports/roberts/roberts_summary.doc (accessed 30 April 2007).

Royal Academy of Engineering (2006) *Reform of higher education research assessment and funding. Response to the department for education and skills.* Available online at: http://www.raeng.org.uk/policy/responses/pdf/Higher_Education_Research_Assessment.pdf (accessed 30 April 2006).

Royal Society (2003) *Supporting basic research in science and engineering: a call for a radical review of university research and funding in the UK* (London, Royal Society).

Royal Society (2006) *Response to the DfES consultation on the reform of higher education research assessment and funding.* Available online at: http://www.royalsoc.ac.uk/displaypagedoc.asp?id=21472 (accessed 30 April 2007).

Sastry, T. (2006) *A dangerous economy: the wider implications of the proposed reforms to the UK research councils' peer review system.* Available online at: http://www.hepi.ac.uk/downloads/MicrosoftWord-28Adangerouseconomy_proposedUKRCreforms_contents.pdf (accessed 30 April 2007).

Sastry, T. & Bekhradnia, B. (2006) *Using metrics to allocate research funds. A short evaluation of alternatives to the research assessment exercise* (Oxford, Higher Education Policy Institute).

Seglen, P. O. (1997) Why the impact factor of journals should not be used for evaluating research, *British Medical Journal,* 314, 498–502.

Slowey, M. (1995) Reflections on change: academics in leadership roles, in: M. Slowey (Ed) *Implementing change from within universities and colleges* (London, Kogan Page).

Stake, R. E. (2001) Evaluation of testing and criterial thinking in education, paper to the *American Psychological Association meeting,* San Francisco, CA, 24 August.

St. Clair, R. (2006) Replacing the RAE: the story so far, *Research Intelligence,* 97, 21–22.

St. Clair, R. (2007) BERA response to RCUK consultation on efficiency and effectiveness of peer reviewing, *Research Intelligence,* 98, 12.

Thackray, A. (1977) Reflections on the measurement of science, *Newsletter on Science, Technology, & Human Values,* 19, 20–29.

Thompson, K. J. (2004) Interdisciplinarity and complexity: an evolving relationship, *E:CO* (Special Double Issue), 6(1–2), 2–10.

Tooley, J. & Darby, D. (1998) *Educational research: a critique. A survey of published educational research* (Report) (London, Office for Standards in Education).

UK Evaluation Forum (2006) *Medical research: assessing the benefits to society*. London, Academy of Medical Sciences, Medical Research Council and Wellcome Trust.

UK Funding Bodies (2005) *RAE 2008. Guidance to panels*. Ref RAE 01/2005. Available online at: http://www.rae.ac.uk/pubs/2005/01/rae0105.doc (accessed 30 April 2007).

United States of America Congress (1937) *The National Cancer Institute Act*.

Universities UK (2006) *Universities UK's response to the DfES consultation on the reform of higher education research assessment and funding*. Available online at: http://www.universitiesuk.ac.uk/consultations/responses/downloads/ResearchAssessmentAndFundingResponse.pdf (accessed 30 April 2007).

van Raan, A. F. (1997) Presentation to the NATO *Evaluation of performance and trends in basic and applied research by advanced bibliometric methods. A science policy instrument for nations with an economy in transition*. Advanced Workshop on Science Evaluation and its Management, Prague, 25–28 November.

Wellcome Trust (2006) *Response to the consultation on the reform of higher education research assessment and funding*. Available online at www.wellcome.ac.uk/assets/WTX033895.pdf (accessed 29 August 2007).