Imbalancing the academy: the impact of research quality assessment

Ian McNay

Emeritus Professor, Higher Education and Management, University of Greenwich, London, UK

The author

Ian McNay has worked in Catalunya and Brussels, in an international organisation, as well as six HEIs in the UK. He has delivered programmes in another 20 countries, currently in a major project on leadership development in Ukraine. This article represents 25 years of research and departmental leadership on the issue; other research interests include organisation analysis, leadership and strategy in universities, and policy analysis, particularly on widening access and student experiences. He has over 150 'outputs' to his name, including 3 edited books from SRHE/Open University Press. He edited *Research into Higher Education Abstracts* for over 15 years, and is a Fellow of both the Society for Research into Higher Education and the Higher Education Academy, and a Trustee of the Universities' Association for Lifelong Learning.

Imbalancing the academy: the impact of research quality assessment

Ian McNay, Professor Emeritus, Higher Education and Management, University of Greenwich, London, UK

Abstract

The UK exercises in research quality assessment since 1986 have had ill-defined objectives. For the first exercises the results were simply 'to inform funding', seeking value for money in an evaluative, regulatory state. That aim remained, almost word for word, until the Research Assessment Framework (REF) in 2014, when there was an additional articulated intent 'to change behaviour'. There had already been much changed behaviour over the years, perhaps unintended, apparently unexpected, as staff adapted their responses to changing 'rules of the game' and the meaning they attributed to these. This article outlines the changing mechanics of successive exercises – the process means to ill-defined policy ends, and analyses the impact of design features which have affected the staff, and distorted institutional strategies of both policy development and control of delivery. The cumulative effect is to imbalance the system to favour a small elite, leading to isomorphism and funding concentrated to an extent that risks loss of diversity and stifling of challenges to established ideas, failing to recognise a variety of excellences.

Key words: performance management, diversity, ends and means, design defects, unexpected consequences.

Imbalancing the academy: the impact of research quality assessment

Introduction

It is thirty years since the first national approach to assessing research quality in a university system was introduced. The pioneer was the United Kingdom, in 1986, within a context where the Thatcher government had an agenda of 'value for money' in public services, of containing the autonomy claimed by professionals, and increasing the accountability for the use of public funds. That applied not only to university academics and their institutions, but to school teachers, medics and others in the public sector, as elements of both the evaluative state (Neave, 1998) and the regulatory state (King, 2007) developed. At that time, the University Grants Committee, the buffer body between the then universities and the state, assumed that academics spent one third of their time doing research, and funded them accordingly. The first exercises served to audit that belief with the aim of 'informing funding decisions'. The belief was misplaced. Successive exercises took place in 1989, 1992, 1996, 2001, 2008 and 2014. Geuna and Piolatto (2016) claim that the Italian system for research quality assessment, VQR, operated through ANVUR (Agenzia Nazionale di Valutazione dell'Universita e della Ricerca), was inspired by the UK scheme where the 'experience, errors and improvements since 1986 have been used to inform policies [and] initiatives in other European and world countries' (p263), but warn about cultural impacts on policy transfer and the time needed to gain experience and develop appropriate skills.

The aim of this article is to review some of those errors and experiences. It examines shifts in the UK exercises and the way individuals and institutions responded. In doing so, it records the lack of clear stated *ends* beyond the economic – value for money - and the constant changes to criteria, data requirements and processes of evaluation – the *means* to the ends. This is similar to, though not totally congruent with the distinction made by Glaser in his contribution to this issue, 'between strategic autonomy (the autonomy of selecting goals) and operational autonomy (the autonomy to select approaches to goal attainment)'. There is evidence that the research assessment approach displaced more appropriate goals that universities might have set for the quality of their activity and even distorted the strategic choice of what research to pursue.

The article is based on a mix of policy and documentary analysis, field work with those involved, using surveys, questionnaires, focus groups and other group discussion methods and participant observation. It has been a personal and professional journey: it is 25 years since I started preparing the first submission I was involved in, as head of a Research and Development (R+D) unit, for 1992. I was then commissioned by the Higher Education Funding Council for England (HEFCE) to conduct a major project examining the impact of that 1992 exercise on individual and institutional behaviour, submitted before, but not published until *after* submissions had been made to the 1996 exercise (McNay, 1997), just as the 2001 evaluation was published too late to affect the 2008 exercise. I led submissions for 1996, and, in a different institution, for 2001, when I was also a member of a sub-panel making judgements on submissions in Education. For 2008 I was an advisor to departments

in several universities preparing submissions, as well as having my work included again, as it was in 2014. Throughout the period, I continued with a number of smaller research projects on research quality assessment, analysing changing policy, institutional strategies and staff experiences, and organised major national conferences on the topic. I was a member of the panel assessing research bids for the Economic and Social Research Council, and for a major initiative – the Teaching and Learning Research Programme. This article draws on those experiences and others' work in various countries to record how successive adaptations have changed the landscape of HE in England. I should declare another interest: the two most recent staff members in the English Funding Council (HEFCE) with lead responsibility for research policy are both former students of mine, one giving considerable credit to that experience for his subsequent career progression.

The article pursues a roughly chronological approach to changes between exercises and then treats factors having an impact on the balance of activity at institutional and individual levels: research, teaching and management. Some comparisons are drawn with other countries. The conclusion is that despite the stated principles underpinning the exercise, there has been little coherence, nor continuity between exercises, nor consistency between academic areas, nor credibility to those being assessed (RAE, 2001;McNay, 2015). There is also a lack of connection between the exercise and other activities of universities and a failure to articulate a 'fit' with a national strategy for science. The main positive is in *post facto* transparency whereby the bulk of the submissions are accessible online as are the judgements and feedback commentaries from panels (REF, 2014c). This provides a rich source for researchers!

The basic framework: the (changing) rules of the 'game'

My starting point is the 1992 exercise, which is when all HE Institutions could enter following system restructuring, including university designation of former polytechnics. All were funded by four unitary councils; assessment continued to be UK wide, but funding was delegated to the four constituent parts of the UK, since education was a devolved responsibility. Before that date the two exercises had been low key, 'light touch', a monitoring audit rather than the comparative and competitive approach that came to dominate in a marketised system with a monopoly customer – the state. In 1986, for example, the Research Selectivity Exercise (as it was called initially) asked for responses to a four-part questionnaire on income and expenditure, planning priorities and output, with the focus mainly at the level of the university as a whole (Bence and Oppenheim, 2005). Only 5 representative outputs were submitted by each subject area. From 1996, the norm became four per person.

In 1992, funding based on grades awarded in the RAE/REF, known as QR funding – Quality Related – was separated from funding for teaching. Previously, UGC universities had been given an unhypothecated lump sum to cover both. It was intended to support infrastructure, a 'research environment', and complemented the other arm of the dual support system of state funding of research through research councils, which funded projects ranging in size from single researcher enquiries up to major centres with significant staff levels. There has been a shift by research councils in recent decades, from responsive

funding – reviewing bids from Principal Investigators in open competition – to inviting bids within themes agreed with government as it became more directive, linking research to national strategic priorities. Until 2008, QR funding was based mainly on overall quality ratings of outputs such as articles, artefacts, patents, papers to conferences. Such output came from individuals, aggregated into 'Units of Assessment' (UoAs) mainly coinciding with discipline-based departments. Data on context at unit level – doctoral completions, other research income, policies for staff support, strategic plans - could moderate grades at threshold points. Such units got a single overall grade. Funding went to the corporate university without conditions attached – an acknowledgement of institutional 'autonomy' – though *expectations* were increasingly stated. So, *individuals*' output was assessed for a *departmental/unit* grade for *institutional* funding. Research teams were unrecognised elements in the structure of research activity and assessment. For only a handful of elite universities did it form a significant element of income: before the trebling and trebling again of undergraduate fees, and withdrawal of most funding council support for teaching in England, only four universities got more money from QR than from teaching funds. As one vice-chancellor/rector put it, the exercises were about 'fame and fortune'; but not much of the latter for many units, despite international level work. Until 1992, funding related to student numbers. With extension of coverage, that changed to 'research active' staff numbers, so as not to fund those not engaged in research, particularly in the newly designated modern universities which did not have a research tradition and had had no funding, but which provided most of the growth in student numbers.

The first exercise in my period required details of ***all*** publications by staff being submitted, with two indicated as 'best' with two others for further examination if needed. The assumption was, then, that quantity was a factor – to show that funds were being used productively – so, an evaluative/productive approach. My evaluation of impact (McNay, 1997) showed that journal editors were, as a consequence, deluged with articles, often irrelevant to the aims of their journals, and of variable quality. Research findings were published on a drip feed basis, with what was called a 'salami slicing' of publications to maximise the quantitative indicator, as staff 'played the game' (Lucas, 2006). Articles often had considerable content overlap with others. Hundreds of new journals appeared, to accommodate the surge of submissions as 'publish or perish' became a reality for staff when publication gives readily visible evidence of work and is used in performance management. The evidence of increased research was less convincing than for increased publication, and a surge in quality of output was also not evident (Talib, 2000, Bence and Oppenheim, 2004).

The outcome provoked consternation among traditional universities. The level of core funding had not changed significantly despite the doubling of the number of submitting institutions. The previously unfunded polytechnics and colleges – known as 'modern' universities - produced some good results, and were given funding beyond the small supplementary amount earmarked for them, reducing the amount available to their more prestigious competitors. (In a later exercise, History at Oxford Brookes University – the former polytechnic – outscored its more ancient neighbour, creating a local frisson of delight or despair depending on location).

For 1996, each staff member identified their four 'best' outputs for submission – a shift to an evaluative/qualitative approach. In Italy, for the Valutazione della Qualita della Ricerca (VQR) exercise, it had been three for university staff and six for those in Public Research Organisations (Bonaccorsi and Malagrini, undated); for the 2011-2014 exercise that reduced to two and four respectively. UK grade criteria made much of the proportion of work at three levels of quality – international, national, and sub-national - but a single grade was awarded, and funding withdrawn from lower graded units. After 1992 results came out, there was evidence of *internal* decisions to allocate income more equally, but perhaps less equitably, to support improvement in those lower graded units, and tax those who generated it. After 1996, my own unit had its QR allocation halved by senior managers: none of the lower graded departments to whom the money was then allocated got a funded grade in 2001, and we only just clung on to our funded grade. Investing in weakness is rarely a good strategy. The practice faded away.

Funding was related to grade and numbers of staff submitted – quality x volume, with three price levels for different clusters of disciplines. Initially there were 5 grades, with the funding gradient a straight line graph. Later, there were 7 and successive exercises cut funding from the bottom rungs and changed the funding gradient (it differed across the four administrations) (Brown and Carasso, 2013). Such decisions were announced *after* bids had been submitted, which made planning strategies for the bid difficult to calculate a balance of advantage between selectivity or inclusivity of all 'research active' staff who had the basic requirement – four 'outputs', of which panel members read between 10 per cent and 50 per cent (McNay, 2007). Between 2009 and 2012, the ratio changed annually, so that the ratio for 2*, 3* and 4* went from 1:3:7 in 2009-10 to 0:1:4 in 2015-16 (Geuna and Piolatto, 2016). 2016-17 will see the 3*/4* gap widen further. Unlike, say, New Zealand, there was no requirement for all academic staff to be submitted; nor did individuals get a personal grade.

When the grade was a single average, the criteria for each grade created traps in their wording (McNay, 2003, 2007), again only clarified *after* results were announced. So a grade 5 required 'international excellence in *up to half* of research and national excellence in *virtually all* of the remainder': 'virtually all' was quantified as 95 per cent; 'up to half' remained unspecified. A grade average could be dragged down by including too many lower quality outputs to the extent that the increase in the quantity multiplier did not compensate for the drop in the quality indicator. One major unit dropped to 5 from 5* in 2001 because there was seen to be no flexibility in 'international excellence in more than half of research and national excellence in [*all of*] the remainder'; a handful of outputs, out of several hundred, fell below national excellence. That calculation of the equation became increasingly important as the funding gradient steepened, under pressure from government (Geuna and Piolatto, 2016); getting it wrong could mean paying an expensive price for motivational staff management.

The QR grade, though dominated by outputs, was sometimes crucially affected by context factors described above. I had experience as a panel member of this: one UoA in a modern university was held down a grade because the activity was new, but good despite that, and an older university was given a higher grade than output merited because it was 'well found'

– i.e. had academic capital and a past record. I can make a case for those decisions being reversed on the basis of equity, with allowance for context in the same way as students from less advantaged background should be given credit for performance for overcoming that, and regret that I did not make such a case more forcefully at the time.

Little changed between 1996 and 2001. User representatives were introduced to some panels, with a variety of experiences, many critical about the attitudes of other panel members, and the exclusionary nature of academic discourse (Lewis, 2002; McNay, 1998b). The funding gradient steepened with fewer grades attracting any money, and a small number of institutions getting the bulk. In England, in 2005, four - Oxford, Cambridge, Imperial College and University College, London - got 29 per cent of the total allocated for research, but only 6.5 per cent of that for teaching. In Wales, Cardiff has consistently had nearly 60 per cent of QR funds. This is true of other systems using RQA approaches to funding: in New Zealand, 2 universities – Auckland and Otago - got 51 per cent of research funding in 2013 (TEC, 2013), despite one objective being 'to prevent the undue concentration of funding' (Ministry of Education, 2013). That objective disappeared soon after (Ministry of Education, 2014). In Portugal, half the universities are excluded from funding after the first round of a two-stage process (Deem, 2016).

For 2008 there was another development in the UK – an objective 'to fund excellence wherever it is found', which the modern universities welcomed as a commitment to a more equitable recognition of their developing profiles and a re-balancing of distribution of funds. The single grade, which had concealed the range of quality behind an average, was replaced by a profile with four levels, showing the percentage of work deemed to fall within each band. There were some derisive comments about criteria applied in defining 'world-leading', which Johnston (2008), another former vice-chancellor, wanted to show revolutionary science, a paradigm shift, mainly beyond achievement in a 5-year perspective. Within the – perhaps less exigent - framework applied, 150 out of 159 UK institutions had at least one unit with at least 5% of its work deemed to be 'world-leading', based on rigour, significance and originality. Adams and Gurney (2010) note wryly that 'it appears that the policy of selective funding, while leading to a fair (sic) degree of concentration of research funds, had not led to quite the concentration of research excellence that might have been expected'. Despite further adjustment of the grade:funding ratio to favour the top grades, 2016-7 allocations per full-time equivalent (FTE) academic show a reduction of over 5 per cent for Oxford, Cambridge and Imperial College, as submitted staff numbers were reduced to protect the headline quality profile, and modern universities included more staff and achieved considerably improved grades, so claiming a higher proportion of the funds available in the zero sum game. Some more than trebled their allocation from the previous year (Ruckenstein *et al*, 2016).

In the 2008 exercise, three elements of output, context and esteem factors were separately graded. Esteem factors were new, and covered such things as prizes, fellowships, office in learned societies, editorial board membership, encouraging academic citizenship activities. The balance of contribution to the overall grade of these three factors was left to panels to decide, leading to considerable variation (RAE, 2006; McNay, 2007). For 2014, esteem

disappeared, and the contributory elements became output, environment and impact, weighted on a standardised ratio across all panels of 65:15:20 and separately reported and funded. 'Impact' complemented 'significance', the latter being seen as recognition *within* academia (probably through citation counts); impact covered effects *beyond* academia: 'change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life' (REF, 2014), and was assessed on the basis of 'reach' and 'significance'. Environment covered strategy, resources and infrastructure with criteria of 'vitality and sustainability'.

The unit discussed above, where I had had regrets, had a higher grade in Education in 2014 for *output* than its more prestigious neighbour in the alphabetic order listings of results, but lost out in the *overall* score on the other two factors, so the overall grade reversed their rank order, raising issues about how quality is defined and what was being judged. As Deem (2016) comments: 'research has become what REF measures'. The environment grade depended heavily on previous funding and 'research capital'. Impact – a concept I support and campaigned for in the context of professional research – was assessed through case studies of evidence from the period 2008-2013 of research done since 1993, *in the same, submitting, institution.* So, anybody who had moved institution in a 20 year period (as had I) lost at least part of their claim to impact. The institution/s they had left could still claim the credit for the work of staff who had left, because the ownership of research was seen as resting with the institutions. This also meant that the 'stars' recruited in a short term perspective to boost profile could not count for impact, but their outputs could be used because they belonged to them and so were portable. This raises questions about ownership of intellectual property.

There have been other issues related to the *means* by which judgements are reached: while panel feedback about what quality research involves in Education stresses large, quantitative projects based on developing and using large statistical data bases, which require a team approach, for 2014 the number of people from the same UoA who could submit the same output/article as evidence of quality of their contribution was limited to three across the social sciences. If co-authors came from elsewhere in the university, and submitted to a different UoA, or were from another university, there were no limits. So a leader's strategy to encourage colleagues to collaborate internally, based on approaches of previous exercises, was undermined. The criteria for impact specifically excluded impact on one's teaching – a difficult hurdle for those who research teaching in higher education! Esteem factors, which may have assessed that among other things through academic peers, disappeared as a specific element, possibly in to the 'environment' narrative. Impact was judged on case studies, with a minimum of two for all submissions and one required for approximately every 10 staff submitted over 15. This led to some staff not being submitted, because the threshold would be passed to require an extra case study which could not be demonstrated or, perhaps, not at an acceptable level for the four point scale (Kerridge, 2015). 23 per cent of subpanel members were user representatives or impact assessors. They were more generous and less fine-grained in their grades than academics, leading to concern about a possible lack of rigorous audit trail of inflated claims (Deem, 2016, Oancea, 2016). Greenhalgh (2015) reviewed 162 case studies linked to community-based health

sciences (Panel A2), aiming to define a template of 'good' impact. She found that claims of impact on clinical guidelines or clinical practice were not supported by data on outcomes - in morbidity or mortality rates - from such changes. In a quarter of submissions, no active efforts to achieve impact were described. Work by Derrick and Samuel (2016) also casts doubt on the validity and reliability of impact assessor judgements across health sciences. The user representatives and impact assessors in Education raised queries about 'the value for money of such a time consuming exercise' (REF 2014).

There was also a risk that network power (Bourdieu, 1988, Kahler, 2009) would have operated in their selection, and that partners in large organisations, working with higher rated units – the network from which members were dominantly drawn - would be over-represented. Previously, professional research had been under-regarded, a point made in my 1997 report and accepted by HEFCE in its response (HEFCE, 1997). This was still true nearly 20 years later, according to a report endorsed at parliamentary and ministerial level (Furlong and Oancea, 2005; ESRC, 2004). 'Near-market' research had been excluded from submission; now this could count for 20 per cent of the quality grade. Conversely, encouragement to be active in the wider research community – learned societies, editorial boards, school research presentations – was discouraged because it brought no rewards recognised in the utilitarian world of politicians and senior managers, who expected a different balance of effort.

Across the exercise, judgements are made by panels of 'peers' from within higher education, mainly from the UK, even though funded work has now to be deemed 'world-leading'(4*), or at least of 'international quality' (3*); being 'internationally recognised' (2*), whatever it means, is not enough for funding, but is difficult to judge from a domestic viewpoint. For 2014, international members were added to the four main panels, though not enough – 36 units of assessment, down from 69 in 2008, shared 23 people - to give a full international perspective on those terms in the sub-panels where grades were determined (Deem, 2016*).* Those terms are not used in grading environment, nor impact, but four point scales are used for both, suggesting similarity, and the terms are used to cover 'overall performance' of a unit when all three are aggregated into a blended grade. This exaggerated the implied quality of work in headline reports, since the ratings for environment and impact were much higher than for outputs – see Table 1.

Table 1 here

The impact scores for Science and Medical subjects were much higher than for social sciences, with an average of 60 per cent gaining 4* ratings for both elements. The scores for the elements other than outputs were much higher than the two used in 2008, though they are not directly comparable. So, across all disciplines the proportion of *outputs* getting a 4* rating went from 14 per cent to 22 percent, but *overall performance* scores at 4* went from 17 per cent to 30 per cent, a much bigger rise, because of shifts in the other two elements. For example, in Education, in 2008, the top score for 4* grade in *context* was 75 per cent and only 5 institutions had a score of 50 per cent or more. In 2014, 18 institutions scored 50 per cent or more, of which 8 scored 100 per cent for their Utopian *environment*. Similarly, for *esteem* the top rating for 4* was 40 per cent, awarded to 2 units; in 2014, 27 units

scored 40 per cent or more and 3 scored 100 per cent for *impact*. This shows differences across time and panels, yet two of the key principles for the exercises claimed by the funding councils are consistency across panels and continuity between successive exercises (RAE2001). The validity and reliability of the process are both called in to question by these data, yet little is said because university staff do not want to challenge the upward trajectory of the quality scores. In Deem's (2016) view, the environment narratives have had less and less scrutiny, and the transparency that is one of the best things about the exercises, reveals what might be labelled 'creative accounting' in the case study claims for impact, where many institutions brought in specialist authors to sex them up (Oancea, 2016).

In Italy, there are fixed quotas for the four grades, so normative approaches are used rather than the supposedly objective criterion referenced approach in the UK. The top 20 per cent are 'excellent', the next 20 per cent are 'good', the next 10 per cent 'adequate' and the bottom half, 'limited'. These grades inform only about 2 per cent of funding, of which, in 2013, 63 per cent went to the top quintile (Geuna and Piolatto, 2016), which may reduce the temptation to 'game' the exercise by packing the sample with lower quality work to improve the chances of a higher grade for *relatively* better work. The financial impact of the Portuguese exercise in 2013 is more significant where, in a two-stage process, half of units do not get through to the second stage and lose all funding: a consequence of restricted funds in an economy under challenge. Of those that did get to stage 2, none were given the top rating of 'exceptional' (Deem, 2016)

Means and Ends

Such changes to processes and data requirements reflected the technicist approach adopted by the UK funding councils responsible for the exercises. They were irritating and disruptive to good research planning, but if there was adjusting of the *means* - of process - a more serious problem with the exercises has been over *ends* - their lack of clear objectives (McNay, 2015a). This distinction is similar to that made by Glaser (2016) in his article in this issue between strategic autonomy and operational autonomy. The first objective set was 'to inform funding decisions', which simply shifts the question to the objectives of funding strategy. The *operational* strategy has been ever increasing concentration of funding in fewer institutions. This may indicate an unstated intention to protect the 'elite' institutions as a subset, when student participation is now at mass levels in Trow's (1974) typology. League tables, which weight research heavily, create pressure to keep that elite group in the top ranks internationally, the level to which comparisons and competition have now moved for such universities. At a conference I organised after the 1996 exercise, the head of the funding council admitted that 'you never know how it will all turn out', which aligns with several writers on unexpected consequences (e.g. Merton, 1936, Elton, 2000, Krucken, 2014), and then claimed that one objective was 'to improve research quality' (McNay, 1998). That has *never* been stated in any official documents. Statements from HEFCE officials did retrospectively claim a causal link with some developments – accountability for public investment in research and its benefits, the establishment of yardsticks and

benchmarks, and leading a cultural shift, though, they noted, the last more among managers than researchers! (HEFCE, 2014).

In 2014, the introduction of 'impact' was overtly intended to change behaviour at institutional, departmental and individual level (HEFCE, 2009), though in what way was not specified. Funding remained a dominant element, and there appeared to be an axiomatic assumption, articulated in other policy arenas, that competition would improve quality – a tenet of New Public Management, as Glaser (2016) notes in his contribution to this volume. Since the total fund was fixed, there was a zero sum game, and funding for lower Q levels was removed or reduced in successive exercises.

The lack of clarity over ends was compounded by changes in factors to be used as evidence of quality and by lack of agreement on means for judging quality. There was a constant debate, which Italy has also experienced (Bertocchi et al*,* 2014) as to whether greater use of metrics would be more efficient, and just as effective, as peer review, if not more so (OECD, 2004: Sastry and Bekhradnia, 2006; Donovan, 2007; Scott, 2007). In the UK, pressure to use metrics comes from politicians who, within New Public Management approaches want easy qualitative indicators. This goes back for more than a decade, but whenever an independent review of the issue is commissioned the outcome is to keep peer review and metrics in a hybrid or symbiotic relationship, with the lead role being taken by peer review. This is true of the latest independent review, which was published just before the deadline for submission of this article (Stern, 2016). McNay (2009: 42-44) has a brief summary of the claimed strengths and weaknesses of the two approaches in isolation. Wilsdon (2015) has an extensive literature review of several hundred articles, admittedly mainly from English speaking contexts, concluding that 'individual metrics give significantly different outcomes from the REF peer review process, showing that metrics cannot provide a like-for-like replacement for REF peer review'. Academics are divided, often along disciplinary lines: the Swedish Research Council has recently proposed that funding should be based on peer review through the FOCUS model every six years, to replace the current metrics base. Government response is awaited, but the academic union, SULF – Swedish Association of University Teachers - is opposed on grounds of efficiency and effectiveness, supported by an Italian academic working in a Swedish university, who had reviewed papers in the last ANVUR exercise to arrive at ratings consistent with the number of citations on Google Scholar (Maukola, 2016). Baccini and De Nicolao (2016) challenge the conclusion by ANVUR on the Italian exercise that outcomes from the two approaches had a good degree of agreement; they claim that the use of the dual system may have introduced biases in the results. They agree with Wilsdon, which is significant because he has recently been asked to chair an expert group on alternative metrics to feed in to the Open Science Policy Platform of the European Union.

Figure 1 about here

If one develops a matrix plotting clarity of and agreement on ends/objectives of the exercise, including quality criteria, against degree of robustness and acceptance of means, the four quadrants produce different behaviours:

- 1. ends and means both clear and agreed: rational decision making. This was the assumption of the funding council, seeing the approach as helping financial accounting and operating discretely, without recognising the impact on other actors, nor setting the exercise in context, relating to wider national science policy, nor to teaching, consultancy and other university activities.
- 2. ends defined, agreed and consistent; means less so: trial and error experimentation. This is seen in the changes between exercises recorded above such as the change in the funding model, the inclusion of user representatives in judgements, reduced recognition of jointly authored papers, the exclusion of teaching as an arena of impact.
- 3. disputes over quality and ends of the exercise, but confidence in particular means and measures. This resulted in a lot of lobbying by what Watson (2009) labelled the HE 'gangs', and Kahler (2009) the exercise of 'network power', particularly as traditional universities tried to retain their dominance, resisting impact as a criterion of quality and pressing for quality to be deemed an organisational characteristic, and so proposing the exclusion of units below a minimum critical mass, advocating the inclusion of all eligible staff in a submission, and supporting selectivity in funding with its exclusionary consequences. This was helped by the membership of panels, dominated by those from research intensive universities – the dominant network - with very few having more than one member from a modern university, and occasional total absences. Those who set the canon and established the parameters of disciplines and methodologies acted as gatekeepers, judging work that might challenge them, and sometimes keeping the gate closed. This is evident in the most recent review (Stern, 2016), which supports those changes; the review steering group had no members from modern universities, but six from the Russell Group of research dominated universities, plus one from Princeton
- 4. extended lack of agreement: 'garbage can' behaviour [Cohen, March and Olsen, 1972) with changing actors/governments in an uncertain decision making process, inconsistent processes and conflicts of requirements, conflicting evidence and opinion on the balance of qualitative and quantitative data. Until recently, this had been evident in Australia, with several plans for change never implemented as the political pendulum swung. The UK system comes close, particularly when changing politicians try to micro-manage.

<u>Designed to distort?</u>

Staff in the funding councils adopted a mix of 1 and 2. The English council is a funding body, not a strategy steering board, they insisted. That led to problems inherent in design factors being denied, ignored or neglected. The simple technicist approach survived through several exercises, despite evidence suggesting the need for fuller review. The HEFCE review following the 2014 exercise, despite criticisms, including my own, also intended little change (McKenzie and Gallodi, 2016): a 5* grade for work better than 'world leading' – perhaps the USA hyperbole of 'stellar' might serve; an increase in the impact of impact to 25 per cent of the overall score. That review was stopped when government announced its own (Stern, 2016). Their policy preference is seen as aiming at reducing the cost of the exercise through

more emphasis on metrics, a policy proposed by all shades of government since before the 2008 exercise, but resisted by academics on the balance of evidence (Wilsdon, 2015). That, too, ignores other, more important, design defects.

Disconnected

One major issue is that the exercise has been conducted as a discrete activity, not located in the broader landscape and operational activity of universities. At a workshop at the 2008 conference of research managers, I conducted a quick survey on the links between research and other activities. In a group of 60, no more than three indicated their institution made good, clear links between research and any of…teaching, enterprise, users, knowledge transfer, economic development, regional regeneration, international competitiveness. Things may have moved on since then, but that was 15 years into the period under review and after five rounds of assessment.

Impact on teaching

My report (McNay, 1997) pointed to issues of concern, acknowledged by the funding council (HEFCE, 1997), but which remain unresolved: treatment of interdisciplinary work, recognition of professional research and, particularly, the damaging effect on teaching. 62 per cent of heads of department in my survey said the effects on teaching had been negative. It became low priority because the RAE/REF produced extra funding for quality without extra productivity, but extra income for teaching could only come through extra students, and quality teaching rarely led to promotion. Only 28 years later, the government's consumerist perspective led it to recognise that 'because some universities see their reputation, their standing in league tables and their marginal funding as being principally determined by scholarly output, this can result in teaching becoming a poor cousin to research in parts of our system' (DBIS 2015, para 1.10). At institution and departmental level, the elite universities emphasised their commitment to 'research-led teaching'. This is not tested by quality assurance of teaching, and risks distorting teaching to give more prominence to research – often beyond undergraduate level – than to the needs of students and expectations of employers, those two being the main 'clients' of universities. Evidence of this was offered by *Times Higher Education* (Havergal, 2016), where the data team modelled the potential results of TEF using the government's indicated approach, and produced a league table. Cambridge fell outside the top ten, Oxford outside the top 20 (below Oxford Brookes), and University College London, St. Andrews, LSE and King's College London well outside the top 50, with less familiar names in the top ten: Loughborough, Aston, De Montfort, Swansea, Kent, Coventry, Keele, Surrey, Bath and Lancaster.

Government intends to introduce a Teaching Excellence Framework to equalise esteem and effort, 'to bring better balance…including stimulating greater linkages between teaching and research' (1.13) because 'research and teaching should be recognised as mutually reinforcing activities' (1.3). Given the reduction of government investment in R+D, there may be a view that better graduates will stimulate economic productivity (a key recurrent theme) as an alternative strategy. But…the only reward for excellence is permission to

charge students higher fees, just as state financial support to students is being reduced. This article is not about the TEF, but it has problems not dissimilar to REF. The intention is to use metrics, but the three proposed, based on current data availability are: student retention and completion, graduate employment, and student satisfaction. The government's paper acknowledges, briefly, that 'these metrics are largely proxies rather than direct measures of quality…and there are issues around how robust they are' (3.13). That is very true – all are contaminated by factors beyond the university's control. Nevertheless the government proposes to adopt the trial and error approach of quadrant 2 above. It needs to learn from previous mistakes. Yet a technical consultation on for the operation of the Teaching Excellence Framework has been issued, not for the first year, where the intention is to press ahead with that in 2016-17 before claiming it works and implementing it fully in 2017-18. The consultation is on Year Two, with replies needed before the *start* of year one. Beyond that, 'the development of future iterations of the TEF will draw upon lessons learned from Year Two' (DBIS, 2016:40, paragraph 156). Such government attitudes to evaluation are seen elsewhere: Glaser (this volume) records the decision in Germany to continue the Excellence Initiative before an international panel of experts had conducted its evaluation, and despite views that it caused imbalance between large and small units, between research and teaching and between the authority of professors and managers.

RAE/REF had a system wide impact on teaching, because poor results on research assessment, with consequent reductions in research funding, were followed by closure of teaching departments, so that, in some areas, courses in key strategic subjects were not easily accessed and the supply of skilled personnel in some regions was reduced, as well as relevant research activity (AUT, 2003). That effect is also imbalanced, geographically, in favour of London and the south-east (Adams and Smith, 2004), contrary to successive governments' espoused commitment to greater regional development, currently to a 'northern powerhouse', where support to productivity – a major driver of higher education policy – is more needed The concentration of research funding means that many departments have become effectively 'teaching only', with consequences for staff recruitment and deployment. That aligns with the reduced requirements in the new Higher Education and Research for designation as a university, aimed at encouraging private entrants to the market, since they are not required to have any research activity, a major, imbalancing, revision of the concept of a university in the UK since the second half of the last century (Robbins, 1963).

Impact on research

Conversely, the framework of assessment panels was based on teaching - the subject discipline structures of undergraduate degrees, increasingly not relevant even there. Researchers working in multi-disciplinary project teams had to disconnect from the team and the context of their work, to be assessed in a structure not fit for that purpose – a basic criterion of quality. Mega-panels were developed to review work in clusters of allegedly cognate disciplines, but those involved are sceptical about their efficacy (Deem, 2016). For 2008, Education was linked with Psychology and Sports Studies, presumably on some memory of a private school education based on *mens sana in corpore sano.*

The exercise is judgemental, not developmental, looking back on *performance*, not forward on *potential*. Yet those at 4* level probably started with some 2* outputs; it is unclear how the exercise supports that progress. I have noted the oral commitment to improving quality, but its absence from official documents. There has been a matching lack of overt commitment to developing research capacity. Despite increases in the numbers of academic staff in the system, the number of staff submitted has declined since 1996, in Education from over 2500 to 1440 full-time equivalent staff (-40 per cent). In Sociology, the number of units submitting fell from 39 in 2008 to 29 in 2014, and staff numbers from 927 to 704, a drop of 24 per cent. Part of that is down to a shift in staff contracts. In 2014, for the first time, while there was a 4.5 per cent increase in numbers of academic staff from the previous year, as recruitment for REF submission reached a peak, academic staff with job descriptions embracing both teaching and research fell and became a minority (48.6 per cent). There has been an increase in functional specialism (Locke *et al,* 2016): over 3,200 extra staff were appointed on research only contracts, which would include 'stars', and nearly 6,000 on teaching only contracts, to cope with increased undergraduate numbers after the 'cap' controlling enrolment levels was removed. That last figure will have included some switched from a dual role to reduce the base of REF- eligible staff, and so increase the apparent intensity of research activity if numbers of those submitted as a proportion of those potentially eligible becomes a metric in the future. Productivity also declined: staff submitted an average of 3.75 outputs in 2008; in 2014, this fell to 3.41, partly because of greater use of exceptional circumstances provision. Nor are prospects good, since half of full-time PhD students are not UK citizens and many will be forced to return to their home countries because of xenophobic visa controls by the Home Office. That will worsen, now that the UK will leave the European Union. High debts from high undergraduate fees make UK graduates reluctant to incur more through immediate doctoral study. The original policy of enhancing value for money is increasingly called in to question (McNay, 2015b).

The UK government ideological axiom, that competition improves quality, so the aim does not need to be stated was disputed in a survey of some 300 academic staff (McNay, 2008): 52 per cent of respondents agreed that 'quality assurance processes have encouraged low-risk conformity at the expense of innovation, independence and "difference"'. That confirmed findings in Germany and Australia (Laudel, 2006, Glaser and Laudel, 2005). The UK Commission on the Social Sciences, set up by the learned societies in the field, was severely critical:

> The academy treadmill, driven by excessive accountability burdens, the Research Assessment Exercise and other factors, has reduced the originality and quality of much academic research and constrained interaction with various communities. (CSS, 2003)

There is also a belief that competition produces diversity because of an entrepreneurial drive to develop new products (Geiger, 1986, in Horta *et al,* 2008). The conclusion by Horta and colleagues was that a single excellence model for assessment promotes isomorphism – a drift to conformity to a norm. One vice-chancellor/rector, and a former head of the English funding council, claims the English have a genius for turning diversity into hierarchy; the

reduction in numbers of research groups with any significant QR funding will reduce that diversity. Perceptions of panels' views will add to that. Staff spend time trying to detect panel members' preferences/biases and adapting their work to match. In Education, the panel feedback (REF, 2016),as well as favouring large scale quantitative projects, suggests a lack of sympathy for professional classroom based research, condemning it for lack of theoretical rigour, when action research and other approaches develop grounded theory through rigorous field work on a Mode 2 approach linked to real world problems (Gibbons *et al,* 1994). Yet, there are a diversity of excellences, depending on purpose and fit with circumstances. For 2014, Psychology was located in the Panel covering social sciences, when much work fitted better with the Science Panel, but that meant lower funding for clinical research in mental health than applied for similar work on physical health (Matthews, 2016b). There had been previous claims that lab-based research in Psychology scored more highly than other methodologies appropriate to different branches of the field (Marks, 1995), as there was for econometrics scoring more highly than less statistical approaches (Harley and Lee, 1997), linked to a constricting 'approved' list of journals, a practice that continues in some panels. Yet the Australian Research Council has now withdrawn such lists, since most of the work cited in the listed journals comes from work in journals outside the 'approved' outlets. The risk of imbalancing the nature of the subject was clear.

Such perceptions also continued the guessing game that endured for several iterations about which panel to choose, given the 'garbage can' of turnover of panel members and variable criteria and standards (McNay, 2007). Deem (2016) records the disputes within panel meetings, drawing on Lamont's work (Lamont, 2009). Sato and Endo (2014) conclude that the 'cat and mouse' process of trying to model panel preferences has led to 'formative reactivity' to gain better grades. They see this as sub-optimisation, where there is 'a mismatch between what is good for each HEI and what is good for society as a whole' (p91). Such 'gaming' may be endemic to such processes: Middleton (2009) uses 'PBRF-able' (fit to submit to the Performance Based Research Fund assessment exercise) to define 'the process in which activities and research outputs of academics in New Zealand tend to be moulded into the patterns that are expected to attain favourable assessment results' (in Sato and Endo, 2014, p95). Rebora and Turri (2013), comparing the UK and Italy, claim that institutional leaders use New Public Management approaches to orient staff behaviour to perceived external requirements: neo-organisational sociology and operational control theory suggest that organisations adapt and imitate to achieve acceptance and legitimacy.

Matthews (2016a) records the views of Jonathan Adams, the leading researcher in this field, on the risk that concentration of funding in a small number of universities brings:

> 'The peak [of research excellence] only works because there is a platform of very good research supported right across the country in different regions, areas and institutions. You can't just restrict your focus on the elite institutions, [without investing in others]; you don't have a feedthrough of younger researchers [because progression is usually within the university of first degree study – transfer to research intensive universities is less usual]…A loss of structural diversity is a loss of

capacity to respond flexibly when priorities change or when opportunities appear. Diversity builds in sustainable performance'.

Yet, as noted, the response of government, through HEFCE (less so in the other three administrations) to evidence showing that concentrated funding did not work well, was…to try to increase such concentration. That is reminiscent of the English abroad, when they are not understood: shout LOUDER.

Adams also criticised looking only backwards without strategic forethought: 'awarding more funds to institutions that did well last year is a safe bet only so long as next year looks similar', his report says (Digital Science, 2015). The report also recommends a diversity of funding streams, so attempting to pre-empt likely government policy of moving the REF to the research councils. Such a move is now included in the Bill before parliament.

One quantifiable indicator of isomorphism is in profiles of outputs, and here there has been a shift in outputs submitted towards journal articles: another imbalancing. The relative volume of articles among submitted outputs grew from 62 per cent in 1996 to 75 per cent in 2008 and further in 2014. Outputs declining in prominence are: conference papers in technology – a quick way to get results out in to the field given long journal lead times, though excluded from some citation indices used to provide metrics; and scholarly monographs in the social sciences and humanities, also excluded from citation indices (Adams and Gurney, 2014). In Education, between 1996 and 2014, the percentage of outputs which were journal articles went from 40 per cent to 78 per cent, with a decline in reports for external bodies, and in website content. In Business and Management, journals as a share of output went from 59 per cent to over 95 per cent in the same period. The report from Digital Science (2015) cites the journal rankings of the Association of Business Schools as exhibiting a lack of diversity and favouring mono-disciplinary research, another congruence with Germany (Glaser, 2016). The conclusion was that 'this may result in researchers tending to comply with disciplinary authority and be pressured into writing papers to fit a narrow core of …journals' (p5).My colleagues at Greenwich are currently subject to such pressures from senior management. Other units were more resistant to pressures: in history, in 2014, books rose in prominence from 2008, though the longer term trend was for journal articles to gain in share from 30 per cent in 1996 to over 40 per cent in 2014. Similarly, in Area Studies, articles rose slowly but steadily from 34 per cent (1996) to 57 per cent (2014). (I am grateful to Ikuya Sato for these figures, drawn from www.ref.ac.uk/2014output).). It is not clear how any process based mainly on bibliometrics would capture output not held in data banks, producing one imbalancing on the nature of 'output', since most output other than journal articles, and not all of those, is not included and so not used in citation counts. Another imbalancing is that work in languages other than English is usually excluded, so pushing those working in other languages down any league tables that use this data as part of their criteria.

A later report from Digital Science (2016), which appeared as I prepared to submit this article, drew on 921,254 outputs from the exercises from 1992 to 2014, plus references included in 6737 non-confidential impact case studies. It reports two main findings across the four main panels of Science, Engineering, Social Science and Arts/Humanities:

- It confirmed the trends reported above, including the drop in using conference papers as outputs. In Engineering this went from 26.9% to 7.9% in the period; for the other three panels, the 2014 figure was under 1%. Books and chapters in both Science and Engineering went for over 13% to under 1%; in social science from 46% to 15.9%. The report suggests the shift was out of conference proceedings for engineering, from mono graphs for social sciences and from media for arts. Patents and artefacts will also have features less, as well as commissioned research reports.
- The output cited to show impact outcomes differed significantly from that used to show academic excellence. A minority of 42% of references used in impact case studies had also been submitted as outputs to any of the exercises, though the figure appears to be higher in professional disciplines, which is logical. That emphasises the dual audience for work and underlines the need for a 'double discourse' as respondents to my work (McNay, 1997) emphasised. Yet, again, professional journals often have lower status in citation indices. Or are excluded altogether; that needs to change to redress the imbalance.

For the ANVUR exercise covering 2004-2010, 73.5 per cent of 'products' were articles and 19.9 per cent were books or chapters of books (Bonaccorsi and Malgarini, undated). One characteristic that emerged in my early work (McNay, 1997) is the impact on research publishing, and then on research focus, of the hegemony of the English language in citation counts, so that researchers on the continental mainland publish less in their native languages, which has an obvious effect on dissemination and knowledge transfer and research use, and the neglect of local issues in trying for 'international' quality, relevance and impact.

The periodic batch processing model creates further reactive behaviours. Publications peak before the census date and decline after it, perhaps as staff get back to doing research to generate the next batch. But the fixed period may discourage blue skies research, arising from the curiosity that is a strong motivational force for researchers. There is no guarantee of 'results' before a deadline, so there is pressure to do work which has a shorter horizon and may be more limited. The fixed period, compounded by the lack of continuity and certainty between exercises, also fosters a serialist/episodic approach to strategic planning and directive approaches to staff management based on a target culture. From the start, one major impact of the exercises has been on management (Williams, 1991, McNay, 1997), as it has been in Germany (Glaser, this volume). For most departments and their universities, QR income is marginal at best, but the exercise has a symbolic value for both staff, who want to benchmark against the best, and managers looking to polish the institutional brand and league table position. The next two sections consider the managed and their managers.

Impact on staff

I have already noted some impacts on academic staff. Many whose output is judged internally to be of ***only*** (sic) 2* quality – 'internationally recognised' – will not be submitted and so deemed not 'research active'. Many research intensive universities were draconian in adopting such measures, aiming to improve their headline quality profile as REF income

became a smaller element in overall budgets. There may be other reasons – or rather causes, because they are only 'reasonable' within a framework without a duty of care to employees: work that does not fit the strategic narrative of the submission, even though it is high quality (Lucas, 2006); or where submission of, say, the twenty-first staff member would require a further impact case study, which cannot be developed. To boost submissions, there is an active trade in recruiting 'stars', many from outside the UK – making the lack of encouragement to development of younger staff, noted above, even more serious. This is similar to the football transfer market, where big money is paid to imported stars without sufficient matching investment to 'grow' local talent. In research, many such stars are not fully committed to developing others. They continue to pursue single-mindedly their own career, and are often absent from campus.

Thomas (2007), another vice-chancellor, is scathing about such 'star based' tactics:

> Even a superficial analysis shows such investment incapable of producing an economically sustainable return…A sort of boom and bust cycle occurs with major investment before the [exercise] and significant retrenchment afterwards because the outcome has not delivered the expected increase in income(44)

Meanwhile, other staff survive. Academics, including researchers, are remarkably resistant to attempts to change their behaviour (Trowler, 1998). Those in professional fields often have a double pressure, first to continue activity in their field, where their performance had been a factor in recruitment, and which is often essential to continuing professional recognition, in some cases needed to be able to teach on courses leading to a licence to practise. Yet they are also expected to develop a research profile with, sometimes, a PhD, rare among such professionals, being required before appointment. Boyd and Smith (2016) surveyed staff across departments serving health professions and found that only 17 per cent aligned with the established policy to focus on research activity. 39 per cent were subverting policy in operation by various stratagems, with 4 per cent simply rejecting it. The rest – 40 per cent – are labelled 'dissonant' being out of sympathy with the policy but slowly/reluctantly adapting to its requirements. That often had high personal cost. One of my projects (McNay, 2007) highlighted four groups:

- the confident/assertive, who were good enough to set their own agenda, be independent and entrepreneurial, usually working in highly graded units;
- the carefree/autonomous, also setting their own agenda, but in low graded units with no funding, so that managers were grateful for *anything*, often done in private time;
- the controlled/oppressed, driven by others, where managers were desperate to improve scores, by whatever means, or were fearful of dropping a high grade and believed in 'strong' management.
- the positive, who welcomed pressure from assessment and a motivator, wanted to see how they compared against a benchmark set (difficult with grades not given ast individual level) and had a supportive 'fit' with their location.

Leathwood and Read (2013) use a three-part categorisation of contestation, compliance and complicity, with staff almost universally contesting policy particularly over small-scale, critical, innovative projects, many with a feminist perspective, which were under threat, with early career researchers a group under extra pressure. And yet…

> …despite high levels of contestation, almost all…were complying with the demands of research audit and performativity, often at significant personal cost. Compliance [was] seen as being the only way to continue to do the research they loved, and to remain in employment [p1172).

Impact on management and the impact of management

Middle managers then become complicit in promoting this compliance, to protect staff from penalties. Even those within departments scoring 100 per cent at 4* on environment ratings are not exempt, as one of my research respondents reported:

> The approach was to move to be 'selective' this time round, having been 'inclusive' in the last RAE. This was achieved through a centralisation of decision-making where decisions were not effectively open to challenge. No senior member of the School questioned that the game had to be played according to the rules (though some whinged about the rules) so of course we became agents of the state. (McNay, 2014, p31)

The fact that grades continue to improve is seen as justification for such management approaches, but Thomas (2007), is again sceptical as well as critical:

> …evidence of a causal relationship between the RAE audit process and improvement in research quality, and leadership and management cannot be proved…on the other hand, there is evidence that the current RAE precipitates behaviours which can be damaging to universities and staff (46).

My report to HEFCE (McNay, 1997) identified views that research management had improved, though there was a lesser claim by researchers that research in general, other than their own, had improved in quality. Even then, nearly 60 per cent of researchers agreed that 'the research agenda – programmes and priorities – is now defined by people other than the researchers themselves', which runs counter to the Haldane principle, long established in UK higher education, that the best people to decide what research to do are the researchers, another issue echoing Glaser in this volume. More recently, it is evident that Thomas's final assertion still has relevance (McNay, 2015).

One damaging behaviour is to let research become 'what REF says it is' (Deem, 2016) and letting the REF dominate strategic thinking, with 'gaming' the next exercise seeming to take precedence over longer-term strategic planning of research and support for it in the wider context of university/departmental values and stakeholders. That has been clear in some recent development work with research leaders in several modern universities. If, for 'fame and fortune' as objectives, are substituted 'reputation and resources', for many departments the key actors in their reputation are not readers of league tables but stakeholder-partners in other activities – teaching, student placements, graduate

employment. In my own unit, income from RAE/REF did not cover the staff cost of preparing the submission. More income derives from work in partnership with clusters/consortia of schools and colleges from where research questions are posed by those who need evidence to support decision-making to improve policy and practice. This does not rate highly with the Education panel in REF. What it does do is build on established contacts, adding value to them, and involves staff and sometimes students who are in the research arena for other purposes, making a research 'top-up' cheaper than starting from the beginning. It also makes impact easier and quicker, because those involved in implementation have been involved from the beginning in design, development and delivery of the research. The long-term relationships are a firmer foundation for continued engagement than trying to read the chicken entrails and divine what the next exercise might involve. Funding from partners is greater, more assured and open to negotiation. Working is collegial in an entrepreneurial culture as more contacts/clients/partners are added. That contrasts with the corporate bureaucracy approaches of many senior managers and research 'support' offices (McNay, 1995).

That aligns with a 'supporter' strategy identified in an article that appeared as I finalised this one (Silander and Haake, 2016: pp7-9), which combines top-down and bottom-up initiatives – an 'interplay between organic growth and strategic vision', where co-operation with the regional business community is important. By contrast, their 'gold-digger' label is based on 'follow the money' – a top-down strategy to find the currently fashionable areas – 'gold-shimmering' - and promote them, disregarding current activity profile or linkages. Their third label – 'inclusive' – comes close to *laissez-faire* with resources spread thinly. They found universities in Sweden that matched all three strategies for developing a research profile.

If managers read the research, the evidence is strong in identifying the optimal management culture. In a project funded by the Leadership Foundation for HE, itself funded by the sector, so that institutions are the intended users and where impact will be measured, Franco-Santos *et al* (2014) noted that universities 'are currently becoming more short-term and results/output driven due to the increased pressures to perform' (p7). That implies what they label an 'agency' approach, with greater monitoring and control, which academics find 'unhelpful and dysfunctional' but which was preferred by professional staff because it gave 'clarity and focus'. On the other hand, 'stewardship' approaches foster long-term outcomes through people's knowledge and values, autonomy and shared leadership within a high trust environment' (p7). Such approaches are 'associated with…higher staff wellbeing…higher research excellence results, students' satisfaction…employability and financial results' (p7). So, economy, efficiency and effectiveness. Those findings are supported by many others: Kallio and Kallio (2014) on performance management; Edgar and Geare (2013) on high performing departments and collegial culture; Kok and McDonald (2015) on leadership and management behaviours to underpin excellence. The list could be long. Yet the drift to New Public Management continues. And yet professional researchers (and teachers for student learning) have continued to deliver.

<u>Other models</u>

Reference has already been made to where practice differs in other countries. The recent review (Stern, 2016) commissioned a study of models close to REF, but that has not yet been published. There are those in the Anglo-Saxon tradition and on the European mainland which could be examined for lessons.

Approaches in other countries differ from the UK in several respects. In the Netherlands research and teaching are assessed together and panels visit the institutions. Ratings follow a similar four point scale, but are not used for funding, nor, officially, for rankings. The strategic plan for 2016-2020 from the French higher education and research evaluation agency (HCERES) says it will evaluate work in a broader setting, and adopt a strategic approach to funding, not the formulaic processes of the UK funding bodies. In Hong Kong, an holistic approach uses all four of the Carnegie scholarships – discovery, integration, application, dissemination (Boyer, 1990). In the USA, where there is no national equivalent of a REF, applications for federal research funding, to the National Science Foundation and the National Institutes for Health, have to show how findings will be used in teaching, and will broaden research participation and enhance capacity. Australia and New Zealand look forward and see their exercises as developing quality and the capacity of professionals who deliver it as well as advancing national strategic agendas and world status (McNay, 2015). So, criteria in Excellence in Research in Australia include 'identifying emergent areas and opportunities for further development' (ARC, 2014). In New Zealand, one objective is 'to increase the average quality of basic and applied research (Ministry of Education, 2014:7). Its achievement can be measured: 'the number of staff whose evidence portfolios have been assigned a funded Quality Category increased by 41.6% between 2003 and 2012'. (TEC, 2013:6). Middleton (2009), cited in Sato and Endo (2014) points out that 'the impact of research assessment runs deeper than mere measurement of "what is there"; such processes are productive or formative'. In the light of the English approach, it is worth noting that the 2013 objectives had included to 'prevent the undue concentration of funding'. Both antipodean countries make specific reference to support for students and newer researchers. So, there are lessons for the UK to learn from others, despite its view that is a leader in the field, by virtue of being the pioneer, and setting the model for others to follow.

Envoi: looking forward

This article has attempted to offer evidence that the policy behind the UK REF, and particularly in England given the importance of its funding model, is ill-defined. The ends are loosely stated; the means are constantly changing, creating uncertainty that operates against innovation. The evidence of value for money is missing; the effect of concentrated funding has strong risks attached; competition leads to sub-optimal isomorphism and loss of a diversity of excellences. The overall effect, over the years, has been to imbalance HE activity at several levels. A consultation on the future of the exercise was started in 2015 by HEFCE, but suspended when the government announced its intention to commission it own review, when it was clear what would emerge (McKenzie and Gallardo, 2016). It seems likely that responsibility for any future exercise will transfer to a super council for research (Nurse, 2015), which a senior Treasury civil servant has now been appointed to chair. Teaching and

research will be further separated at system level, diversity of funding streams will be reduced and the elitist network will be further strengthened. Politicians think that bigger is better as a default position, and remain convinced that money can be saved by adopting metrics as a dominant means, despite the recent report to the funding councils supporting a hybrid approach with peer review retained, with different approaches providing different, complementary data (Wilsdon, 2015). Perhaps ministers think that was the wrong answer. That is the problem with doing research on policy analysis: ideology trumps rational objectivity. How big an impact factor do you think this article will have before the next exercise?

There are other issues for UK research: the turbulence of the HE system, and the consequent stress for staff, has been evident in recent weeks. Government is in turmoil following the referendum vote to leave the European Union. That could mean considerable loss of income from EU funding for research and other projects, as well as affecting international research partnerships and staff recruitment and job security. There is a Higher Education and Research Bill before parliament, which may be law by the end of 2016, to restructure the governance of higher education, and, despite those fine words quoted above on the interdependence of teaching and research, to separate policy, development, implementation and funding for the two functions even more. The restructuring of government departments by the new prime minister, Theresa May, means that the minister for higher education and science, Jo Johnson, brother of the more infamous Boris, now heads functions not just in separate units of the one ministry – formerly Business, Innovation and Skills, notably not referencing either HE or Science – but to different senior politicians in different ministries, since universities have been moved to the Department for Education, while research has remained in a remodelled ministry of Business, Energy and Industrial Strategy. There is still no mention of research in that title and the new super council for research and innovation is to be chaired by a senior mandarin from the Treasury, giving an indication of how government sees the strategic function of research.

There is some hope on the horizon. The government-commissioned independent review of research assessment reported at the end of July (Stern, 2016). It examined the purpose and benefits of REF; current problems and issues [many treated in this article]; principles and high-level recommendations for shaping future exercise recognises many of the issues raised in this article; and a vision of the organisational location of the REF. Its title – *Building on success and learning from experience* – suggests a less than radical intent, but an acknowledgement of the approach in the second quadrant of my model above, and underlying my critique, some of which it shares. Its membership was dominated by the elitist research-heavy universities, with five members, plus one from Princeton, with none from the modern universities with their different values and experiences. That is refelected in the recommendations, which cover:

- outputs, where all active staff with outputs, however many should be submitted, with an average output per head, but allowing some – the better ones - to submit more. Outputs should not be portable, [which shifts the ownership from the academic professional to the institution and will constrain staff movement], and

should continue to be assessed by peer review supported by metrics, disappointing advocates of more quantitative methods
- impact, which should include an institutional level element to allow treatment of interdisciplinary project teams that cross internal divisions. A more flexible approach should allow a wider scope of areas of impact, including teaching
- environment should have a major element focused on the institution – its research culture and corporate support for research, with unit submissions framed by that institutional assessment
- system policy should ensure more strategic use of REF results and data, linking to other data collecting exercises to avoid duplication and promote integration; to better understand the health of the research base , areas for future development and the case for strong investment in research.

How that works out over the next five years will, no doubt, be researched and researched, but…by others.

References

Adams, J. and Gurney, K. (2010) *Funding selectivity, concentration and excellence – how good is the UK's research?* Oxford, HE Policy Institute

Adams, J. and Gurney, K.A. (2014) *Evidence for excellence: has the signal overtaken the substance? An analysis of journal articles submitted to RAE2008,* London, Digital Science

Adams, J. and Smith, D. (2004) *Research and Regions: An overview of the distribution of research in UK regions, regional research capacity and links between strategic research partners,* Oxford, HE Policy Institute

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

Australian Research Council (ARC) (2014) *The Excellence in Research in Australia (ERA) Initiative,* www.arc.gov.au/era/ Accessed January 2015

Baccini, A. and De Nicolao, G.(2016) Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise, *Scientometrics,* 2016:1-21

Bence, V. and Oppenheim, C. (2004) The role of academic journal publications in the UK Research Assessment Exercise*, Learned Publishing,* 17: 53-68

Bence, V. and Oppenheim, C. (2005) The evolution of the UK's Research Assessment Exercise: publications, performance and perceptions, *Journal of Education Administration and History,* 37(2): 137-155

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A., Peracchi, F. (2014) *Assessing Italian research quality: A comparison between bibliometric evaluation and informed peer review,* www.voxeu.org/article/research-quality-assessment-tools-lessons-Italy (accessed December, 2015)

Bonaccorsi, A. and Malgarini, M. (undated) *Research Assessment and Research Evaluation in Italy: the ANVUR Experience,* ANVUR, www.vyzcum.cz/storage/att/33E23087221B49FB093AAE04C91B2B14/62-4_evaluating-research-on-a-large-scale-the-italian.pdf

Bourdieu, P. (1988) *Homo Academicus,* Cambridge, Polity Press

Boyd, and Smith, C. (2016) The contemporary academic: orientation towards research work and researcher identity of higher education lecturers in the health professions, *Studies in Higher Education,* 41(4): 678-695

Boyer, E.L. (1990) *Scholarship Re-considered: Priorities for the professoriate,* Jossey-Bass/Carnegie Council for the Advancement of Learning

Brown, R. and Carasso, H. (2013*) Everything for Sale? The marketization of higher education,* London, SRHE/Routledge

Cohen, M.D., March, J.G. and Olsen, P. (1972) A garbage can model of organisational choice, *Administrative Science Quarterly,* 17

Commission on the Social Sciences (2003) *Great Expectations: the social sciences in Britain,* http://www.the-academy.org.uk

DBIS (2015) *Fulfilling our Potential: Teaching Excellence, Social Mobility and Student Choice,* Cm 9141, London, Department of Business, Innovation and Skills

DBIS (2016) Teaching Excellence Framework. Technical consultation for Year Two, London, Department of Business, Innovation and Skills

Deem, R. (2016), Research evaluations in the UK and Portugal: methodologies, processes, controversies, responses and consequences, Presentation to UCL/IoE seminar, London, 9 February (article in preparation)

Derrick, G.E. and Samuel, G (2016) The societal impact-focused and quality focused evaluator. Balancing evaluator views about the assessment of the societal impact of health and medical research, *Minerva.* Online first

Digital Science (2015) *The value of structural diversity. Assessing diversity for a sustainable research base,* London, Digital Science and SPRU

Digital Science (2016) *Publication patterns in research underpinning impact in REF2014,* Bristol, HEFCE www.hefce.ac.uk/publications/ Accessed 30.7.2016

Donovan, C. (2007) Future pathways for science policy and research assessment: metrics vs peer review, quality vs impact, *Science and Public Policy,* 34(8): 538-542

Elton, L, (2000) The UK Research Assessment Exercises: Unintended consequences, *Higher Education Quarterly,* 54[3]: 274-283

Edgar, F. and Geare, A (2013) Factors influencing university research performance, *Studies in Higher Education,* 38(5): 774-792

ESRC (2004) 'Memorandum from the Economic and Social Research Council' in evidence to the House of Commons Parliamentary \select Committee on Science and Technology (2004) *The work of the Economic and Social Research Council.* HC 13. London: The Stationery Office

Evidence, Ltd. (2003) *Funding Research Diversity,* London, UniversitiesUK

Franco-Santos, M., Rivera, P. and Bourne, M. (2014) *Performance management in UK higher education institutions: the need for a hybrid approach.* London Leadership Foundation for Higher Education

Furlong, J. and Oancea, A. (2005) *Assessing quality in applied and practice-based research. A framework for discussion,* Oxford, Oxford University Department of Educational Studies

Geuna, A and Piolatto, M. (2016) Research assessment in the UK and Italy: Costly and difficult but probably worth it (at least for a while), *Research Policy,* 45: 260 – 271

Gibbons, M., Limoges, C., Novotny, H., Schwartzman, S. and Scott, P. (1994) *The New Production of Knowledge: the dynamic of science and research in contemporary societies.* London, Sage

Glaser, J. (2016) German universities on their way to performance-based management of research portfolios, *Sociologia Italiana,* (this volume)

Glaser, J. and Laudel, G. (2005) The impact of evaluation on the content of Australian university research. Paper to TASA conference, December

Greenhalgh, T. (2015) Research impact in the community-based health sciences: what would good look like? Dissertation, MBA in HE Management, UCL Institute of Education

Harley, S. and Lee, F. (1997) Research selectivity, managerialism, and the academic labour market. The future of non-mainstream economics in UK universities. *Human Relations,* 50(11): 1427 – 1460

Havergal, C. (2016) A new perspective, *Times Higher Education,* 23 June, pp36-49

HEFCE (1997) *The impact of the 1992 RAE on higher education institutions in England,* M6/97, Bristol, Higher Education Funding Council for England

HEFCE (2009) Research Excellence Framework: second consultation on the assessment and funding of research, HEFCE 2009/38, Bristol, Higher Education Funding Council for England

HEFCE (2014) REF 2014 shows UK university research leads the world, News release, 17 December, Bristol, HE Funding Council for England

Horta, H., Huisman, J. and Heitor, M. (2008) Does competitive research funding encourage diversity in higher education?, *Science and Public Policy,* 35(3): 146 – 158

Johnston, R. (2008) On structuring subjective judgements: originality, significance and rigour in RAE 2008, *Higher Education Quarterly,* 62 (1-2): 120-147

Kahler, M. (ed.) (2009) *Networked Politics: Agency, power and government,* Ithaca and London, Cornell U.P.

Kerridge, S. (2015) How threshold for case studies shaped REF submissions, *Research Professional,* 15 February

Kok, S.K. and McDonald, C. (2015) Underpinning excellence in higher education – an investigation into the leadership, governance and management behaviours of high-performing academic departments, *Studies in Higher Education,* http://researchonline. ljmu.ac.uk/2191/

King, R.P. (2007) Governance and accountability in the higher education regulatory state, *Higher Education,* 53[4]: 411-430

Krucken, G. (2014) Higher education reforms and unintended consequences: a research agenda, *Studies in Higher Education,* 39(8): 1439 – 1450

Lamont, M. (2009) *How Professors Think: inside the curious world of academic judgment.* Cambridge, Mass., Harvard UP

Laudel, G. (2006) The art of getting funded: how scientists adapt to their funding conditions, *Science and Public Policy,* 33(7):489 – 504

Leathwood, C. and Read, B. (2013) Research policy and academic performativity: compliance, contestation and complicity, *Studies in Higher Education,* 38(8): 1162-1174

Lewis, J. (2002) Assessing the RAE: An expensive, (mad) lottery?, Paper to the Annual Conference, Association of University Administrators, April

Locke, W., Whitchurch, C., Smith, H. and Mazenod, A. (2016) *Shifting Landscapes. Meeting the staff development needs of the changing academic workforce,* York, Higher Education Academy

Lucas, L. (2006) *The Research Game in Academic Life,* Buckingham, SRHE/OpenUP

Marks, D. (1995) Bias in UFC assessment exercise, *The Psychologist,* July

Matthews, D. (2016a) Focus on 'elite universities "risks responsiveness of UK research", www.timeshighereducation.com/news/focus-on-elite-universities-risks-responsiveness-of-uk-research/

Matthews, D. (2016b) Mental health research 'being short-changed', *Times Higher Education,* 28 July, p6

Maukola, J. (2016) Swedish union rejects research audits, *Research Professional,* www.researchprofessional.com/

McKenzie, L. and Gallardo, C. (2016) Researchers praise 'steady as she goes' consultation, *Research Professional*, 24 February

McNay, I. (1995) From the collegial academy to corporate enterprise: the changing cultures of universities, in Schuller, T. (ed.) *The Changing University?*, Buckingham, SRHE/Open UP

McNay, I. (1997) *The Impact of the 1992 RAE on Institutional and Individual Behaviour in English Higher Education: The evidence from a research project,* Bristol, HEFCE

McNay, I. (1998a) The RAE and after: 'You never know how it will all turn out', *perspectives,* 2[1]: 19-22

McNay, I. (1998b) Challenging the traditional: professional knowledge, professional research and the RAE, *Innovations in Education and Training International,* 35(3)

McNay, I, (2003) Assessing the assessment: An analysis of the UK research assessment exercise, with special reference to research in Education, *Science and Public Policy,* 30(1): 47-54

McNay, I. (2007) Research assessment; researcher autonomy, in Tight, M., Kayrooz, C. and Akerlind, G. (eds.) *International Perspectives on Higher Education Research, Vol. 4 – Autonomy in Social Science Research: the view from United Kingdom and Australian Universities,* Oxford, Elsevier

McNay, I. (2008) The crisis in higher education: the views of academic professionals on policy, leadership, values and operational practices. *Higher Education Review,* 40(2)

McNay, I. (2009) Research quality assessment: objectives, approaches, responses and consequences, in Brew, A. and Lucas, L. (eds) *Academic Research and Researchers,* Maidenhead, SRHE/OpenUP

McNay, I. (2015a) Learning from the UK Research Excellence Framework: ends and means in research quality assessment, and the reliability of results in Education, *Higher Education Review* 47 (3): 24-47

McNay, I. (2015b) Does research quality assessment increase output and give value for money? *Public Money and Management,* 35(1): 67-68

Merton, R.K. (1936) The unanticipated consequences of purposive social action, *American Sociological Review,* 1(6): 894-904

Middleton, S. (2009) Becoming PBRF-able: Research assessment and Education, in Besley, T. (ed.) *Assessing the Quality of Educational Research in Higher Education.* Sense Publishers

Ministry of Education (2013) *In pursuit of excellence: analysing the results of New Zealand's PBRF quality evaluation.* Auckland, Ministry of Education

Ministry of Education, (2014) *Review of the Performance-Based Research Fund,* Auckland, Ministry of Education

Neave, G. (1998) The evaluative state reconsidered, *European Journal of Education,* 33[3]: 265-268

Nurse, P. (2015(*Ensuring a successful research endeavour: review of the UK research councils*, London, Department of Business, Innovation and Skills

Oancea, A. (2016) Research impacts: networks and narratives. Presentation to launch seminar, Centre for Global Higher Education, UCL Institute of Education, 3 February, www.researchcghe.org

Organisation for Economic Co-operation and Development (OECD), (2004) *A Performance Based Research Funding for Tertiary Education Institutes: The New Zealand experience*. Paris, OECD

RAE (2001) Circular letter 2/99, www.rae.ac.uk/2001/pubs/2_99/section1.htm Accessed 30.7.201

RAE (2006) *RAE 2008. Panel criteria and working methods,* Bristol, HEFCE (www.rae.ac.uk/pubs)

REF (2014) *The Research Excellence Framework, 2014: The Results,* Bristol, HE Funding Council for England

Robbins, L. (1963) *Higher Education: a Report,* London, Her Majesty's Stationery Office

Ruckenstein, A.E., Smith, M.E. and Owen, N.C. (2016) We can't go on like this, *Times Higher Education,* No. 2243, 25 February, pp35-39

Sastry, T. and Bekhradnia, B. (2006) *Using Metrics to Allocate Research Funds,* Oxford, HE Policy Institute, www.hepi.ac.uk

Sato, I. and Endo, T. (2014) From the RAE-able to the REF-able? A note on formative reactivity in national research quality assessment, *Research on Academic Degrees and University Evaluation,* 16: 85-104

Scott, A. (2007) Peer review and the relevance of science, *Futures,* 39(7): 827-845

Silander, C. and Haake, U. (2016) Gold-diggers, supporters and inclusive profilers: strategies for profiling research in Swedish higher education, *Studies in Higher Education,* http://dx.doi.org/10.1080/03075079.2015.1130031 accessed 26 February, 2016

Stern, N (2016) *Building on Success and Learning from Experience. An independent review of the Research Excellence Framework.* London, Department of Business, Energy and Industrial Strategy

Tertiary Education Commission (2013) *Evaluating Research Excellence- the 2012 assessment. Final Report,* Wellington, TEC

Thomas, E. (2007) National research assessment in higher education, in de Burgh, H., Fazackerley, A. and Black, J. (eds.) *Can the Prizes still Glitter?* Buckingham, Buckingham University Press, pp 39-47

Trow, M. (1974) Problems in the transition from elite to mass higher education, *General Report on the conference on Future Structures of Post-Secondary Education,* Paris, OECD

Trowler, P.R. (1998) *Academics Responding to Change,* Buckingham, SRHE/OpenUP

Watson, D. (2009) *The Question of Morale: Managing happiness and unhappiness in university life,* Maidenhead, Open University Press

Wilsdon, J. (2015) The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management, Bristol, HE Funding council for England

Table 1. REF 2014 Quality profile: percentages by grade

	4*	3*	2*	1*	0
Overall	30	46	20	3	1
Outputs	22.4	49.5	23.9	3	0.6
Impact	44	39.9	13	2.4	0.7
Environment	44.6	39.9	13.2	2.2	0.1

Figure 1 Ends and means in RQA