Science and Public Policy, 45(5), 2018, 673–682
doi: 10.1093/scipol/scx091
Advance Access Publication Date: 22 January 2018
Article

Exploring the degree of delegated authority for the peer review of societal impact

G. E. Derrick* and G. S. Samuel

Centre for Higher Education Research and Evaluation, Department of Educational Research, Lancaster University, LA1 4YD, Lancaster, UK

*Corresponding author. Email: g.derrick@lancaster.ac.uk

Abstract

This article explores how panel expert reviewers' evaluative practice was influenced by external, political considerations during the assessment of a societal impact criterion. The results showed that prior to the evaluation process, participants demonstrated a strong preconceived, political belief that the results of the evaluation process must 'showcase' the value of British research to the public and policymakers as part of a rationale designed to ensure continued public-based research funding. Post-evaluation interviews revealed how, during the societal impact assessment, evaluators drew on these strong beliefs which informed a group-based strategy of 'generous marking' of submissions. We discuss the implications of external motivations that influence the direction of research audit exercises where the definition of the criteria is untested, unclear, and unfamiliar to evaluators, as well as discuss the suitability of peer review as an evaluation tool. Both have implications for the future legitimacy of impact assessment as a formalized criterion.

Key words: peer review; impact assessment; societal impact; expertise; research evaluation.

1. Introduction

Peer review is largely based on the political assumption that assessments are made by a highly specialized group of experts acting within a role of 'delegated authority' regarding the distribution of public funding. Its relevance as an evaluation tool is questioned when the level of expertise of peers in relation to specific criteria under their jurisdiction is questioned. This leads to peers requiring external influences to generate the assessment.

Shifts in research policy, and an increased pressure on government spending has meant that research is being increasingly viewed in terms of productivity, 'economic efficiency', accountability, and delivering 'value for money' (Shore 2008). Over the last few decades, as Governments endeavour to ensure academic research is both accountable and beneficial to society, a political drive to incorporate the criterion 'societal impact' into such assessments has emerged (Watermeyer 2016). As a result, a new social contract has now arisen between science and state, which revolves around notions of accountability, relevance, and value (Demeritt 2010; Guena and Martin 2003; Strathern 2000); and research evaluation has become a prominent exercise to exercise and enforce new, Mode 2 inspired concepts of research value (Hill 2016). For the UK, the inclusion of societal impact criteria in grant application (ex-ante) and in the Research Excellence Frameworks (ex-post) highlights that in addition to being academically excellent, research must also prove its social worth in order to guarantee ongoing, and/or increased future funding.

As the gold standard in research evaluation, peer review is largely used as an evaluative tool for societal impact criteria as a strategy to promote academic community participation, and assure stakeholders that evaluation outcomes are fair (Chubin and Hackett 1990; Chubin 1994; Collins 2010a; Penfield et al. 2013; Wilsdon 2016b). As part of the research community, peer review plays an important political and academic role by balancing the public's need for accountability with the academic community's need for self-governance. Underpinning this is the assumption that the people making the assessments although not directly accountable for these decisions regarding public funding, would be acting within a realm of technical expertise that enables them to act with a degree of delegated authority (Jasanoff 1990, 2003a). This 'delegated authority' is neither elitist or relativist and works by allowing experts to act on the behalf of elected, publically accountable figures (Jasanoff 1990, 2012). Publics and publically accountable figures, therefore, grant experts a carefully circumscribed power to speak on their behalf on matters requiring specialized judgment. For the sake of peer review, the boundaries of this delegated authority relate to an assessment of academic excellence; its potential to make a significant contribution to the discipline as judged by a fellow disciplined researcher (peer), as well as judgements about whether the project is feasible in terms of scope and budget. More importantly, it is accepted that these decisions are made by those within the academy. This essential characteristic of peer review is echoed by many scientific organizations, that judgements are made '… by those with demonstrated competence to make such a judgement' (British Academy 2007).

However, these boundaries of expertise as dictated by learned societies within the academy (British Academy 2007) relate primarily to judgements of socialized indications of academic excellence, and does not extend to judgements beyond this experience such as for notions of societal impact. Within these non-academic notions, pinpointing excellence is therefore more sensitive to highly subjective influences (Derrick 2018; Lamont 2009; Lamont and Huutoniemi 2011b). As with processes that demand nuanced, non-traditional notions of excellence when technical expertise is ineffectual, other forms of expertise are then needed to drive the assessment (Collins 2004; Collins and Evans 2007; Collins 2010b). Here, interactional expertise, as a form of tacit knowledge is operationalized and is constructed from all evaluators that takes into account all subjective and conflicting viewpoints around the excellence that exist within a peer-review panel. The existence of interactional expertise is used to legitimize peer review as a tool for assessing interdisciplinary research (Collins 2010a) where no obvious, direct peer exists. In this way, the value of a peer is that they bring their intellectual, political, and social insights to be operationalized together, therefore, acting on interactional expertise as a substitute for technical expertise. However, for Impact, this can be dangerous as this mutual construction transcends a specified degree of delegated authority risking a process driven by bias instead of any form (technical, interactional or otherwise) of expertise.

For notions of societal impact, previous research highlights how little technical expertise evaluators have with this criterion (Derrick and Samuel 2016c) and their resulting nervousness and apprehensiveness towards the evaluation. Here, evaluators perceived the criterion as something outside their expertise and were openly concerned about how to resolve the variety of values, views, and beliefs held by individuals about societal impact as an object to judge and value (Derrick and Samuel 2016c; Derrick 2018; Derrick and Samuel 2017). This declaration from evaluators of an absence of technical expertise, and their fear that interactional expertise would be insufficient, questions the boundary of delegated authority permitted for societal impact criteria. Unlike other ambiguous evaluation criteria (Derrick 2018) such as Interdisciplinary research (Collins 2010a; Huutoniemi 2010; Klein 1990; Laudel 2006; Laudel and Origgi 2006; Porter and Rossini 1985), assessing societal impact does not require judgements to be made against socialized norms of research excellence through interactional expertise. For interdisciplinary research, assessment of excellence (publishing in good journals, measure of esteem, etc.) are made against socialized academic thresholds but for societal impact, this habitus is not available for evaluators and/or is insufficiently independent from external, non-technical influences on the judgement. This is not to suggest that it is possible that peer-review practice is free from external influence either for Impact, or more traditional criteria, but the legitimacy of peer review as a tool lies in its perceived immunity to such forces. It remains to be seen how the borders of an acceptable degree of delegated authority and level of expertise are influenced by such non-academic influences in an assessment of the societal impact of research.

This article examines an example where the political and academic roles of peer-review clash resulting in panel deliberations that are infiltrated by external motivations to advocate in favour for promoting research's public value, thereby influencing the evaluation outcomes from the peer-review process. We define these external influences as those not based on an assessment of the explicit characteristics of submissions but that stem from evaluators' consideration and sensitivity to political and social factors of the assessment process. By doing this we explore how these influences unconsciously force evaluators to act beyond the boundaries permitted by their delegated authority (Jasanoff 1990, 2003a, 2003b, 2012).

2. Literature review

2.1 The boundaries for peer-review panels judging societal impact

The peer-review system is a central feature of both academic governance and research evaluation. It is to ensure excellence in research, as well as trust and accountability to the academic community, stakeholders, and the public (Hojat et al. 2003; Zuckerman and Merton 1971). There are three types of peer review, including individual peer review which occurs during academic publishing; group/panel peer review during funding decision-making (ex-ante); and group/panel peer review for research evaluation (ex-post) (Chubin and Hackett 1990). While these three separate processes are all referred to as peer review, the processes involved in each case are distinct. In panel peer review, which is the focus for this article (expert review for the REF), a group of evaluators jointly deliberate about the merit of proposals, with a final chair sometimes making the decision based on the common judgement of all reviewers (Olbrecht and Bornmann 2010). During peer review, a piece of research is judged by researchers working in, or close to, the field in question (Boden 1990). The basic function of peer review is to judge the value of proposed research against current knowledge, and to use yardsticks for evaluation, which stem from dominant disciplinary belief systems and paradigms (Luukkonen 2012a). The essence of the system lies on the belief that peer reviewers have a level of expertise in the field to evaluate the research and act fairly in their judgment (Research Councils 2006). If evaluators are viewed as lacking the expertise to conduct the assessment fairly and properly, then the process is called into question, and its integrity is compromised.

Despite illusions of group consensus and democracy in deliberations, if individual evaluators feel inexperienced or ill-equipped to provide an assessment, research has shown that they defer judgment to those evaluators who they perceive to be more knowledgeable, thus respecting disciplinary sovereignty (Lamont 2009; Lamont and Huutoniemi 2011a). This is the basis for how interactional expertise is operationalized. Past studies of evaluating 'interdisciplinary research' show how a panel adopts conservative outcomes when there is a lack of direct peers to assess the criteria and/or object (Luukkonen 2012b). In these cases, where interactional expertise is primarily responsible for guiding peer-review panels in the absence of direct, technical expertise, evaluators are reluctant to 'break' the boundaries of typical knowledge paradigms and go beyond epistemic and disciplinary boundaries in the assessment process (Langfeldt 2006; Laudel and Glaser 2014; Luukkonen 2012a). As such, evaluators prefer existing yardsticks for assessment based on existing paradigms of knowledge, despite conflicting assumptions about 'quality' (Huutoniemi et al. 2010). This ongoing conflict, results in the work being denounced or, when an assessment is compulsory, placed outside scrutiny (Huutoniemi 2012b). Nonetheless, assessments of interdisciplinary research are still made against common socialized norms of academic 'excellence'. Using these mutually agreed norms as yardsticks, the panel acts within a pre-determined, and granted boundary of delegated authority that recognizes that a peer's socialization within the norms of academic identifies them as best placed to make judgements about its quality. The reported

reluctance of evaluators to surrender these yardsticks, against which the quality of interdisciplinary research is judged inappropriately and therefore its true quality is still contested, is in part responsible for unfavourable and conservatives evaluation outcomes (Huutoniemi 2010; Huutoniemi 2012a; Luukkonen 2012b).

Unfortunately, these common yardsticks are not available to peer-review panels, and are useless for assessments of societal impact which requires panels to judge research's value beyond academia (HEFCE 2010, 2011), as well as to navigate a number of specific drawbacks related to the nature of impact (Donovan 2011; Penfield et al. 2013; Smith 2001). The absence of common yardsticks, and the absence of direct experience of societal impact results in an evaluation situation seemingly legitimized by expertise, but driven by neither the technical (Jasanoff 2003a) or interactional expertise (Collins 2004) used to justify the choice of peer review as the gold standard evaluative tool for other assessments of research quality and value. This questions the expertise permitted within an acceptable level of delegated authority (Jasanoff 1990) of these peer-review panels. Indeed, the newness of the societal impact as a criterion suggests a distinct novelty in terms of review processes that is vulnerable to the influence of political or other, non-academic lenses to generate outcomes.

In the case of societal impact, experts still play a crucial role in operationalizing a level of public trust that decisions are made only within the realms of a prescribed power to speak on the behalf of the public regarding a specialized judgement. The expectation is that experts operate only within the scope of their delegation and this does not extend to considerations of social and political influences. Therefore, evaluators who are led by an overly biased position are not acting within the expected level of trust that their delegated authority in a democratic society allows. Even though expertise is more than a binary consideration accounting for one's background or experience (Derrick 2018), and can also extend to differences in opinion regarding political, or social value judgements, it is not within the delegated authority of panels (as a group) to allow such inclinations to overshadow the type of specialized judgement expected from evaluators. However, in the evaluation of Impact the required level of expertise contingent on experience does not often pre-exist the experience and is dependent on the very practice of the evaluation (Derrick 2018). The danger of evaluating societal impact, therefore, exists when personal conceptions about the criterion incorporating political, economic, or social considerations are deemed as more reliable yardsticks for evaluation. This results in the evaluators reaching beyond their prescribed level of delegated authority. Furthermore, if evaluators are unable to act within a prescribed degree of expertise and exercise the necessary norm of disinterestedness in light of exogenous influences (Nedeva et al. 2014), then this challenges the reliability of peer review as the choice of evaluative tool.

2.2 The political and social debate around impact in the UK

As the UK's national assessment exercise, the 2014 Research Excellence Framework (REF2014) stated three main purposes, each of which resonates with aspects of audit cultures and social contract ideals: informing the selective allocation of funding based on research excellence (*quality—promotion of research performance*); providing benchmarked information and establishing reputational yardsticks (*benchmarking*); and providing accountability for public investment in research and evidence of benefits (*accountability*) (Research Excellence Framework). Though, as Wilsdon notes, as the exercise has evolved, two additional effects have been associated with performing the exercise: influences on research cultures and behaviours, and as a management framework for research activities (Wilsdon 2016a). This has resulted in the REF2014 playing a major role in the creation and proliferation of an audit culture (Dahler-Larsen 2006, 2011; De Rickje et al. 2016) in the UK, as well as a more publicly transparent benchmarking of research's value to national interests and ambitions.

The inclusion of the societal impact criterion (Impact) within the REF exercise was driven by a politicized 'impact agenda'—an ambition by the UK government 'for the UK [to have] a reputation not only for outstanding scientific and technological discovery, but also as a world leader in turning that knowledge into new products and services' (Treasury 2004: paragraph 1.1). Tying in with audit ideals, the government argued that 'the HE [higher education] sector can and should do more to ensure that its excellent research achieves its full potential impact' (HEFCE 2009: 4, paragraph 10). In response, The Higher Education Funding Council for England (HEFCE)—developers of the REF along with the four other UK Higher Education Funding Bodies—compelled UK Higher Education Institutions 'to showcase the impact of [their] research'[1] during the REF so that the research community could advocate their value to both the Government and society. Such a response broadly resembles an argument put forward by Kearnes and Wienroth who, beyond audit culture, view 'the evaluation and assessment of research [as] ... part of a broader set of political struggles concerning the place of science in public life' (Kearnes and Wienroth 2011: 169). The need to showcase impact thus becomes evident as a method to advocate the role and value of research to government and society alongside other contenders of public money; and the impact agenda, through the eyes of academics, becomes interrelated to societal perceptions of research value and worth.

3. Methods

3.1 The in-vitro approach

The methodological design utilized a series of interviews with evaluators of the UK's REF2014 responsible for the assessment of the 'Impact' criterion. This design, known as the 'in-vitro approach' (Derrick 2018), was developed as a way of compensating for the difficulty of observing peer review panel processes directly, including the REF2014. This approach juxtaposes and compares evaluator responses during two sets of interviews; one conducted with evaluators before the evaluation begins (pre-evaluation), and then another after the evaluation is complete (post-evaluation). This approach captures pre-conceived ideas and opinions that evaluators had in the pre-evaluation interviews then maps the evolution of these ideas through in-group interplay (using recollections captured in the post-evaluation interviews). In particular, using the pre- and post- evaluation responses, reasonable hypotheses could be drawn about group debates, as well as the interaction of raw, baseline views expressed prior to the evaluation process were resolved in committee, and then tested and confirmed by the post-evaluation interviews. In this way, the design adopts a social constructivist approach without interfering with the process itself and provides a robust research design for studying peer-review panel dynamics, in lieu of direct observations.

3.2 The REF2014's impact criterion

The REF2014 was chosen as the focus of this study due to the inclusion of the world's first, formal, dedicated assessment of societal impact (Impact). The criteria informed 20 per cent of a submitting Higher Education Institute's (HEI) overall assessment for a specified Unit of Assessment (UoA). To aid in the preparation of submissions, as well as the evaluation, Impact was defined as '... an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (HEFCE 2011). The assessment process was governed by one of the four overarching Panels and their subsequent sub-panels, which reviewed a specific UoA. In this study, we used Main Panel A and its subpanels to investigate the evaluation of impact (societal impact).

3.3 Recruitment

Interview participants were sourced purposefully from REF2014's Main Panel A, which covers six Sub-panels: (1) Clinical Medicine; (2) Public Health, Health Services, and Primary Care; (3) Allied Health Professions, Dentistry, Nursing, and Pharmacy; (4) Psychology, Psychiatry, and Neuroscience; (5) Biological Sciences; and (6) Agriculture, Veterinary, and Food Sciences. This panel was selected in order to allow the exploration of issues related specifically to impact evaluation (rather than issues about the characterization of the criterion itself) in fields in which there is already a myriad of recognized ways in which this type of research influences society.

Initially 215 evaluators were identified and invited to participate in the projects. This invitation included output only, impact only, and output and impact evaluators (whilst output only evaluators did not assess impact, they were invited to act as a control group). Two sets of semi-structured interviews were conducted with willing participants. Pre-evaluation interviews were conducted by GD before the REF evaluation process started, between January–March 2014 ($n = 62$, including six output only assessors; 47 output and impact assessors; and nine impact only assessors; 28.8 per cent response rate). Post-evaluation interviews were conducted with the same sample of interviewees by GS and GD between December 2014 and April 2015 ($n = 57$, including six output only assessors; 44 output and impact assessors; and seven impact only assessors with five declined invitations).

All interviewees were provided with a participant information sheet and informed and/or written consent was obtained prior to commencement of the interviews. Ethics approval was granted on 22 November 2013 from the Brunel University Research Ethics Committee (2014/4).

3.4 Guidelines for assessing impact

Panels assessed Impact by reviewing four-page case studies submitted by each university, as well as an impact template that described the wider university strategy of facilitating the translation of research into impact. The structure of the four-page case studies was tightly controlled by a template supplied by HEFCE, where universities must nominate pieces of underpinning research and then proceed to explain how this research has had an impact. This underpinning research must be considered to have reached a threshold of no less than two stars in quality ('quality that is recognized internationally in terms of originality, significance and rigour').

Evaluation guidelines for the health and medical panels stipulated that impact could be 'achieved from within a wide variety of research contexts and resulting from a wide diversity of approaches', and there was 'no pre-formed view of the ideal context or approach' towards impact: (33). Contributions to health and welfare; society, culture and creativity; economy and/or commerce; public policy and services; production; environment and practitioners and services; and international development, were viewed legitimately, and with equal weighting (Research Excellence Framework 2012). Indeed, the REF guidelines acknowledged that 'impacts can be manifested in a wide variety of ways including, but not limited to: the many types of beneficiary (individuals, organizations, communities, regions and other entities); impacts on products, processes, behaviors, policies, practices; and avoidance of harm or the waste of resources' (27).

To guide the panel's assessment of impact, evaluators have been asked to make an overall judgement of impact against two criteria: significance and reach. Significance is defined as the 'intensity of the influence or effect', whereas Reach is described as 'the spread or breadth of influence or effect on relevant constituencies'. The criterion of 'reach' is not restricted to purely geographical terms, nor in the number or context of particular beneficiaries, but instead on the spread or breadth to which the potential constituencies have been affected. The assessment of impact is awarded either one of five star profiles, where the lowest rating (Unclassified) is where '... the impact has little to no reach or significance, or was ineligible, or not underpinned by excellent research produced by the significant unit', and the highest (four stars) is where the impact '... is outstanding in terms of its reach or significance'.

The marking scale incorporated five star profiles, where the lowest rating (0 – Unclassified) is where '... the impact has little to no reach or significance, or was ineligible, or not underpinned by excellent research produced by the significant unit', and the highest (four stars) is where the impact '... is outstanding in terms of its reach or significance' (Research Excellence 2011). Evaluators were instructed to consider the evidence presented within the case studies during their assessment of impact.

3.5 Interviews

Interviews were conducted via the telephone, skype, or face-to-face; lasted between 35 minutes and 2 hours; and were recorded and transcribed for analysis. A more in-depth structure of the interview schedule (pre- and post-evaluation interviews), along with a discussion of the methodology's rigorous qualitative framework has been discussed previously (Derrick and Samuel 2016a,b; Samuel and Derrick 2015). In the interests of confidentiality, all participant information was coded and entered into NVivo (qualitative analysis software package) for analysis. The codes used in the results below relate to the participant's panel (Main panel = P0; Sub panel 1 = P1, and so forth) and their evaluation responsibilities (Outputs and Impact (OutImp); Impact only (Imp); or Output only (Out)).

3.6 Analysis

An in-depth discussion of the analysis of the pre- and post- evaluation interviews has previously been provided (Derrick and Samuel 2016a,b; Samuel and Derrick 2015). In brief, the analysis used an inductive approach to grounded theory. Such approaches use an exploratory style methodology and 'coding' techniques, to allow concepts, themes, and ideas to emerge from the data (Charmaz 2006). Duplicate coding by both the first and second author was cross-checked to ensure reliability of data. Whilst we note, and have discussed elsewhere (Derrick and Samuel 2016c; Derrick 2018; Derrick and Samuel 2017; Samuel and Derrick 2015), differences between evaluator's views and decision-making about societal

impact, the common themes expressed in this particular article represent the consensus of all evaluators, including academic and user evaluators; including evaluators from different subpanels, including the main panel; including different seniority levels; and including those evaluators who considered their own research to have had an impact.

4. Results

4.1 Pre evaluation interviews

4.1.1 The need to demonstrate public value in science

Even before the assessment of societal impact had commenced, evaluators had expressed their awareness of their delegated authority and their role participating in a process that had the opportunity to 'need…to demonstrate the value of science' (P3Out2). However, implicit in these declarations was a personal 'need' for the Impact evaluation outcomes to be used to publicly demonstrate this value. This 'need' served two purposes. The first was around ideas of 'accountability' (P2OutImp1). Originating in the realms of the audit culture, and resonating with new notions of a social contract and the reported purpose of the REF2014, for these evaluators, researchers had a 'responsibility' (P4OutImp6) to communicate research and explain its benefits to society: 'we do really … have an obligation to society' (P0OutImp1). This morality extended from an ethical belief that the public had a 'right to know' (P4Out1) how taxpayers' money is distributed (Clegg Smith et al. 2009). As P0OutImp6 pointed out: 'much of what we do … . is paid for by the public in the street, so we ought to think very carefully about the messages we give back to them … to me that's a very moral part of the impact agenda'.

The second need, revolving around a 'justification' (P1OutImp3) for research, was an extension of 'accountability' in that it was also linked to allocation of research funds. However, beyond merely being answerable to the public for received funding, in this account, evaluators also perceived a need to 'justify' funding received. Evaluators believed that researchers must show the return from government investment in research in defence of the government's initial decision to fund research. This belief in its political role extends beyond an execution of expertise that is independent of external, non-academic influences. This belief also resonated with narratives of audit culture, accountability and value, as well as with the REF's aim to 'evidence [the] benefits' of research (Research Excellence Framework). Underpinning this, was the perception that the Government only funded research in order to receive some sort of (economic) benefit: 'the reason that government funds research is because they want to improve money and the economy, that's the only reason' (P1PoutImp2). Similar to the language used by HEFCE themselves,[2] evaluators perceived that 'showcasing' the value of research conducted within universities during the societal impact assessment provided an opportunity to evidence these benefits: 'I can see why it is useful for HEFCE to have these cases to show to government that what they have been funding for all these years has had some impact' (P1OutImp7).

The showcasing of impact, however, went beyond REF's stipulated requirement to evidence research benefit; it became an 'advocating mechanism' (Haynes et al. 2011) to demonstrate the worthiness and value of research receiving public funding over and above other contenders for public money and ultimately echoing a political motivation by evaluators. This mechanism also acted as an advocacy tool to validate further funding as a reward for what had already been achieved, and as such embodied the continued political struggle fought by the research community to ensure that science maintains a dominant place in public life and government funding (Kearnes and Wienroth 2011). Thus, aware of the limited resources available to universities, evaluators expressed a need to 'argue' for continued funding: 'resources are getting smaller and smaller - we have to make a reasoned argument as to why we should be funded' (P1OutImp3). Interviewees commonly spoke about needing to 'persuade people' ('we need to persuade people that our money is being well spent … that we are actually getting health gain' (P0OutImp2)), and to 'convince politicians' of the value of research ('convince the politicians … that universities are important for health of the country' (P1OutImp2)). They also spoke about a need to 'orientate' society (POP1OutImp1) and 'change the perspective of the external world' (P3OutImp8) to view research more favourably, so that society would also support science investment, and they, too, could champion scientists' interests and help with the research sector's political struggle to accrue continued funding (Kearnes and Wienroth 2011). As P4OutImp4 pointed out, 'this is essentially an exercise in convincing the treasury to give us more money'. P5OutImp4 concurred: 'translating [research] into impact is very good at drawing money down from treasury if nothing else'. The need to convince the government about the benefit of science research was viewed by evaluators as paramount in terms of UK science and investment, and yet this desire is beyond an acceptable degree of delegated authority permitted through peer review.

The consequences of not grandiosing the societal impact of research during the evaluation exercise was perceived as detrimental in terms of funding allocation: 'unless we tell the public the good news of what we've funded, we get less money coming in' (P1OutImp4). As one assessor, P4Imp2, stressed, 'if we come out with the wrong answer that may affect international, national and other sorts of investments' (P4Imp2). This terminology ('the wrong answer') serves a crucial purpose to highlight the strength of the established preconceptions about the impact evaluation prior to embarking on the process, and it was these beliefs that were brought into the assessment process. The use of terminology ('wrong answer') by participants also indicates a perception that evaluation outcomes are determined by guideposts other than an appropriate level of expertise, since by nature, peer review (if done with an appropriate degree of expertise) does not produce wrong answers.

4.2 Post evaluation interviews

4.2.1 The political purpose of societal impact

During post evaluation interviews, evaluators' views about the political purpose of the impact evaluation process had not changed. In fact, similar to pre-evaluation, in the post-evaluation interviews, evaluators' narratives about the rationale of assessment still reflected those stipulated by the developers of the REF exercise. In this way, evaluators continued to articulate the evaluation process as an accountability exercise to evidence research's benefit to the government, and communicate and educate the public about the emerging societal benefits of research ('I think it would be very good…to give lots of good examples of impact of research … the general public are not aware of this. It's a good story, it's a British story and it's worth counting').

Evaluators also continued to perceive the impact case studies, and their subsequent evaluation exercise, as an advocacy mechanism in justifying science research beyond their prescribed authority. As articulated in the pre evaluation interviews, this was constructed as

a need to reassure the government that the funding already provided to the research sector had produced societal benefit ('from the point of view of government who are giving money to universities and from maybe people giving money to charities, it's reassuring them … that what they have contributed to is making a difference' (P1OutImp2)), as well as a need to influence future government funding allocations, so that science funding maintains a central role in public spending ('we wanted to get money from the treasury for higher education' (P0OutImp4)). In order to achieve this, a strong perceived need remained from pre-evaluation to post-evaluation to prevent loss of public support for research; support seen as critical to help promote the importance of science research investment to the government ahead of other rivals (Kearnes and Wienroth 2011): 'everyone talks about working on a cure for this and that … and medicine never gets transformed. And that translates into a risk of losing public in the battle for the right place for science in what we do as a culture' (P4OutImp2).

Below the lack of traditional socialized yardsticks available for evaluators is argued as a reason why perceptions about the societal impact evaluation process as an accountability and advocacy exercise was adopted by this peer-review group. The adoption of such a 'political' yardsticks 'steered' group decision-making away from forms of expertise evaluation drivers and towards a 'showcasing' of societal impact: 'it was made very clear, at least informally that we couldn't be seen to have the impact fail … this was very political' (P5OutImp1)). This 'political steer' to showcase impact was exhibited in a pattern of generous marking: 'that's a danger, if universities came out saying that research has virtually no impact … so … the impact assessment has to give high scores, otherwise government wouldn't be funding research' (P3OutImp10). The group's adoption of this generous marking strategy is explored below.

4.2.2 Showcasing societal impact through generous marking

Marking generously was a consistent theme discussed by all evaluators in the post-evaluation interviews ('we scored them high, I thought we were generous' (P5OutImp1)), including by the control group non-impact (output only) evaluators ('it is an assumption that everything ought to be three or four star' (P2Out1)). Therefore, the generous marking scheme evident in the expert review assessment of societal impact was not solely related to the newness of the criterion, or to evaluators' inexperience. Rather, it was related, according to evaluators, to both external and internal factors influencing the output and impact assessment processes. These included, for example, each discipline's vested interests in showcasing themselves ('the fact that they're a super panel supposedly overlooking, they're still in a cognate area, broadly speaking which doesn't want to do itself down' (P3OutImp2)). They also included a desire by evaluators to 'do the right thing', and not necessarily apply their technical or interactional expertise, but to lend encouragement to universities who may be less established:

> we would have said, it really feels like a two to me but these guys, they've only got two impact case studies and it's a really young and new university. It's spreading into this area and actually we want to encourage more of that kind of stuff, so let's give it a three (P3Imp2).

The freshness and novelty of the impact case studies made the 'stories' appealing, also promoting an air of celebration ('a celebration of science' (P0P2OutImp1))[3] that resulted in generous marking: 'I think when they'd finished it [the evaluators], there was a great sense of celebration really, because they suddenly realised what an amazing contribution UK was making in these different areas' (P0OutImp6). Beyond these factors, a whole variety of additional reasons were also given to account for the generous marking scheme for impact. These included, for example, the applied nature of the health/medical fields; universities submitting only the best impact cases; and the well-acknowledged funding difference between two-star (no funding gained) and three-star (funding achieved) grades. Encouragement to be generous also originated from international assessors who were 'very good by saying, don't be so bloody British and put yourself down; think about how good this is' (P2OutImp2). This suggests that for both evaluation exercises, external influences beyond the expertise of evaluators played a role in making judgements.

However, it was the newness of Impact, and the lack of clarity regarding its valuation which encouraged evaluators to adopt external, more political influences of accountability and advocacy as a reason to adopt a generous assessment strategy ('my expertise to be totally critical about a research paper is much, much more advanced than my expertise to assess whether an impact case study is a really great story or not' (P5OutImp3)). This was because evaluators were uncertain of how to score Impact, having neither the experience, expertise, or evaluation precedent, to use as a benchmark to assess case studies. As a result, evaluators desired a substitute benchmark to compensate for this lack of direct, technical expertise. As previous impact evaluation experience was not available to evaluators as prior experience of assessing ex-ante impact assessments were reduced to 'tick box' criteria (Derrick and Samuel 2016a), evaluators did draw upon their expertise as researchers, but not as direct peers, producing impact. Indeed, for this group of health and medical researchers, echoes of their medical training permeated their desire to 'do no harm', and therefore give an above average impact score ('rule of the hospital, first do no harm – I think that was an approach most of us probably had for the impact' (P3OutImp2)). In addition, evaluators were influenced by the external, political factors of accountability, and advocacy, which were explicit motivations for the REF2014 exercise. In these cases external, non-academic influences were adopted as evaluation benchmarks and evaluators' need to showcase impact became a generous yardstick against which to value impact case studies; 'we were reminded several times that this [the assessment exercise] was about making sure that British research was seen as good' (P2Imp2). In short, evaluators needed a purpose to assess societal impact, and permission to showcase the case studies gave the group the liberty required to apply a generous marking strategy.

4.2.3 Approaches to generous marking

Previous research has highlighted how the REF2014 evaluation outcomes for the Impact criterion were heavily skewed towards the upper (four-star) end of the scale, especially when compared to the evaluation of outputs (Derrick 2018; Manville et al. 2015). An examination of distribution of rates awarded to Impact from Main Panel A demonstrated this tendency towards generosity (Derrick 2018). The way in which the generous marking strategy was applied showed evidence of external, political influences on evaluators' behaviours within the evaluation process. In fact, panels were 'encouraged by the Main Panel to give people the benefit of the doubt' (P0P1OutImp1), with 'a general will to round up, rather than down' (P3OutImp2). P0OutImp1 explained how the message received from the Main Panel was to showcase impact: 'the major

message to me was we had to do a better job of. . .showing people what we're doing and how important it is'. Evaluators were encouraged to distinguish their assessment of societal impact ('not assessing in grant-mode') from their usual, more critical, style of marking often used, for example, during grant funding evaluations ('grant mode'). This conscious decision away from a behaviour more commonly found in approaches to assessing more traditional and socialized criteria ('grant mode') suggests a modus operandi beyond a previously accepted level of delegated authority for peer-review panels. As such, evaluators were prompted to grade impact highly (all case studies were potentially a four-star), unless reason was evident to lower the mark:

> 'the strategy I came away with was not to think you're in a grant-awarding mood . . . whatever you have in front of you is good unless proven otherwise. So it was, kind of, innocent rather than guilty' (P2OutImp2).

Such advocacy fuelled strategies resulted in evaluators becoming 'slightly more liberal in what constitutes evidence' (P1OutImp2) and a 'lot of stuff scored highly . . . all you had to do was trace a linear link between one particular study, and a change in a guideline' (P0OutImp4). 'Upstream' impacts – those closer to the research than to a final benefit to society—were valued as much as those downstream:

> upstream impacts were valued as highly as downstream impacts on the scale that we were using. . .as long as you met the criteria for impact and the evidence is there that your research was a contributing factor to that impact, you were pretty much in the four star, certainly . . . three-star range (P0P2OutImp1).

This generous approach to assessment was compounded by the marking scale evaluators were provided (0–4 stars). This scale, rather than representing a bell curve was skewed such that once the threshold of a four-star had been reached, all case studies were marked in this top bracket, whether they were very good, 'impressive' (P0OutImp5), 'outstandingly impressive' (P0OutImp5), or 'the best of the bunch' (P0OutImp2). The evaluation thus became a reflection of the evaluators' desire to advocate for science's value, rather than of an expert review assessment designed to distinguish between different levels of quality: 'in the end the purpose was to celebrate the excellence of British research, not to discriminate between different types of research and its impact' (P0P2OutImp1). In this way, the evaluation was driven by a level of interactional expertise within the group to favour higher scoring for Impact.

4.2.4 Generous marking—making impact assessment 'doable'

The consequence of the generous marking strategy was that the inability (or reluctance) of the group to value different impacts differently, as would be expected from an evaluation (Dahler-Larsen 2011). As described above, if the impact case study could be classified as 'good', it had already reached full marks (four stars), and as such, did not require to be distinguished from any other case studies which were at a similar level, or better ('it wasn't particularly discriminatory because so many people were scored at the top level' (P1OutImp4)). Indeed, there was a feeling that all impacts could be put together into one category; meaning that the assessment process 'wasn't too hard' (P1OutImp4) because 'the impact didn't have to be necessarily all that spectacular . . . the people just had to prove or to demonstrate or tell a story about what impact it had had' (P3OutImp1). This approach removed many of the pre-evaluation concerns about having to make value judgements about different impacts (Derrick and Samuel 2016a; Samuel and Derrick 2015) since, as P5OutImp1 stated:

> when you look at a drug that has cured, say, a very rare disease compared with a drug that may have had a 30 percent complete response rate in a common disease, how do you compare those two? And in the end, I said, actually I don't have to compare those two; they're all going to score 4 stars (P5OutImp1).

Thus, in line with a post-evaluation report by RAND, which reported that evaluators perceived the impact component of REF to have 'gone well' (Manville et al. 2015), 58 per cent evaluators felt that impact was easy to assess for health and medical research, and evaluators spoke about the simplicity of their assessment of impact comparative to their original anticipation[4]: 'I think there was a lot of uncertainty . . . and then when you started doing it, it was surprisingly easy' (P4OutImp5). The perception was the process was 'doable' ('it's doable without too much variation across the panel' (P1OutImp4)) and as P0OutImp2 noted, even those who initially had reservations about how the assessment of impact would proceed given the newness and uncertainty surrounding the criteria, now agreed that assessment was indeed possible: 'even the most cynical of us who went into this process would agree now that you can to an extent make an assessment of impact'. The assessment process was therefore announced a success: 'the final report from the REF was almost overwhelming because it was a splendid thing, and it all worked' (P0OutImp4).

However, this perception of success by evaluators is blurred by their involvement in the evaluation process, and by their intention to advocate for science and not be driven by a robust deliberative assessment characterized by the interchange of various forms of expertise (Greene 2000). Rather, evaluators had 'lulled themselves' into believing it worked (P0OutImp4) when instead, due to the lack of expertise evaluators adopted an evaluation approach which was externally and politically motivated and not solely dependent on a level of expertise ordinarily used to provide legitimacy to expert review assessment. Thus, necessary value judgements to distinguish assessments of 1–4 star impacts were 'shelved' until post-evaluation ('I think these are really quite troubling things about impact . . . which perhaps everybody just had to shelve' (P3OutImp2)). In light of this, one evaluator described the exercise as a 'fass': 'the main panel chairman came in and said we'd underscored, so we needed to up scores, which in my opinion made a fass of the whole exercise' (P6OutImp2). And another interviewee commented: 'we've persuaded ourselves that this worked . . . but, no, I don't think we've demonstrated vigorously societal impact' (P0OutImp4). Indeed, the lack of expertise exchanged within the panel, resulted in shortcuts and overly pragmatic strategies being adopted beyond the level of delegated authority normally granted to peer-review panels.

4.2.5 Consequences of generosity associated with external, political considerations

Some evaluators were acutely aware of the generous marking strategy applied by others during the assessment. First, evaluators were concerned that, given the political connotations in terms of science investment of performing poorly in future evaluations, pressure may be placed on future evaluators to maintain, or even heighten, the generous marking approach seen in this assessment. As P3OutImp6 stated:

> if the government gives us more money or keeps the funding the same because the success of the REF, and the next time [impact] doesn't improve, then somebody is going to say well … these people haven't worked harder because that is just what they did last time.

Evaluators also pointed towards the devaluing of higher education institutions as an unintended effect of marking impact generously. The lack of distinction between star ratings applied to different impact case studies meant that good case studies performed equally to those which were more impressive, making it difficult to distinguish between different levels of institutional quality: 'it's a pretty serious business, this up-scoring … a lot of institutions that scored very well in the first round are devalued by mediocre institutions being up-scored' (P6OutImp2). Other evaluators were concerned that the public may look unfavourably on the lack of differentiation between good and excellent impact case studies, especially when four-star impact cases had little societal benefit ('I think if the public saw that we valued saving lives as roughly the same as having written it down in NICE, I think they might say, 'I'm not sure about that' (P2Imp2)).

Finally, P2OutImp1 also noted that whilst marking generously avoided the need to value some Impacts as better than others, it actually permeated perverse value judgements. This was because classifying a wide range of different impact qualities together under one or two-star ratings (three or four stars) suggested that each of those impacts on society were equally important. However, this was not always morally defensible:

> … my position was … you're making social value judgments anyway, but you're making really perverse ones. You're saying two kilograms of peanuts is the same as one baby's worth, and one baby is worth 500, 000 people in Pakistan, and that's just wrong.

These unintended effects of non-expertise driven societal impact evaluation and the effect on the legitimacy of the evaluation and its outcomes need to be considered in the development of future evaluation frameworks which utilize peer review. A broader exploration of how future frameworks assessing societal impact by peer review can be improved are included in Derrick (2018).

5. Discussion

Using an in-depth, qualitative analysis of pre-conceived conceptions and their behaviour during the evaluation process, this article offers an empirical analysis of how external political considerations, infiltrate the assessment of societal impact. It examined how, for the assessment of societal impact, the political and academic roles for peer-review clash and the role of expertise in defining an acceptable degree of delegated authority for assessing societal impact. It found that for peer-review panels where the relation between the expertise available to the panel and the criterion are more diffuse and less direct, that panels adopted evaluation shortcuts that echoed external, more political influences expressed by panellists prior to the evaluation. Such influences included the desire to advocate on the behalf of research, and generously mark to promote its larger societal value. In particular, the motivation to 'showcase' societal impact as a means to justify continued and, in some cases, increased public financial support for research. This motivation expressed by evaluators prior to their assessment of the impact criterion, manifested itself in a 'political steer' that resulted in a pattern of generous marking and desire 'to do no harm' (P3OutImp2). This article does not suggest that evaluators approached the assessment with anything less than a balanced and critical assessment of the case studies as best they could given the lack of technical expertise available, but that this absence resulted in assessment strategies that were heavily influenced by the political and economic value associated with the outcomes. The drawing on such exogenous political factors during the evaluation exemplifies the socially/politically-constructed nature of the evaluation, building on work described by Derrick (2018).

For the REF's impact criterion, evaluation was always going to be difficult, given the fact that the criteria remained undefined at the point of evaluation; the lack of precedent about how to assess these new measures; and the inexperience of evaluators to do so (Derrick and Samuel 2016a; Huutoniemi et al. 2010; Samuel and Derrick 2015). In this regard, the panel's decision to adopt a strategy of generous marking may have been adopted as a reasonable strategy to navigating the complexity of societal impact, or else the result of other, more explicit panel direction. In the case of the evaluation of interdisciplinary and ground-breaking research, similar difficulties have been noted (Collins 2010a; Huutoniemi 2010; Laudel 2006; Porter and Rossini 1985), and some of these difficulties are related to epistemic differences and others to the lack of a direct 'peer' with the technical expertise sufficient to make assessments. In these cases panel peer review experienced difficulties maintaining accountability, as evaluators struggled to identify previous experiences or benchmarks to guide their evaluation (Huutoniemi et al. 2010; Langfeldt 2006; Luukkonen 2012a; Samuel and Derrick 2015). However unlike Impact, for interdisciplinary research the value of the interactional expertise used to generate their assessment processes were sufficient as they were based on socialized norms of 'excellence' in academia that transcend disciplinary boundaries.

In this way, assessing societal impact using peer review is different as the difficulties commonly associated with peer-review processes were compounded by external, political influences available to evaluators to sway their decision-making in lieu of direct expertise. Therefore, the 'impact accountability' and acceptable degree of delegated authority and expectation of expertise (Huutoniemi 2012b) became difficult to maintain. The essence of the peer/expert review system lies with the belief that evaluators have expertise in the field to evaluate the research, and will conduct an objective assessment with integrity aligned with the evaluations' purpose and rationale (Langfeldt 2006; Molas-Gallart 2012; Research Councils 2006). If evaluators are viewed as lacking the expertise to conduct the assessment, or of being un-objective, then the process is called into question, and its integrity is compromised. The evaluator's approach of marking generously in order to 'showcase' the impact can be understood in the face of uncertain criteria (Derrick and Samuel 2016b), but doing so sets a dangerous precedent for the robust and reflexive assessment of societal impact in future evaluation frameworks.[5] Applying a soft touch approach to societal impact evaluation in REF2014 promotes a skewed value of its 'excellence'. This risks how the research sector and policymakers form their opinion of what constitutes good societal impact, the yardsticks used to assess it, and the legitimacy of any assessment outcomes. Thus, if the evaluation outcomes are questioned so too are the processes and experts used to produce them.

The challenge for future assessment procedures is to balance the forces of the state with the desires of the academy to ensure standards in universities (Dill and Beerkens 2013). With the assessment approach towards societal impact tending towards the reflection of the 'political state' rather than that of 'expert opinion' it is

recommended that in the initial implementation of a societal impact criterion, that guidelines acknowledge the role external, political influences play in infiltrating expert review process. This role needs to be explicitly stated to allow for engagement and debate about the range of external factors, which can enter the decision-making process, and how such factors can be incorporated, accommodated, and/or weighted during assessment. In addition, policies must debate whether considerations of such influences are within the remit of the delegated authority peer-review panels have for societal impact.

Notes

1. <http://impact.ref.ac.uk/CaseStudies/FAQ.aspx> accessed 1 December 2017.
2. <http://impact.ref.ac.uk/CaseStudies/FAQ.aspx> accessed 1 December 2017.
3. This was compounded by the fact that evaluators were instructed to take the impact case studies at face value, with evaluators instructed not to verify claims made within the document ('as per the rules … one had to just take that [the impact case study] at face value' (P3OutImp2)).
4. However, some impact evaluators, as well as the output only evaluators, did continue to hold on to their original concerns ('the easiest thing is the immediate effects on the scientific field … trying to assess the wider impact on society or humanity as a whole, is a bit more tricky' (P4out1)).
5. This includes the REF2021 where the value of the Impact criterion has been increased from 20 per cent to 25 per cent.

Acknowledgements

This research was funded by a Future Research Leaders Fellowship awarded by the UK's Economic and Social Research Council (ESRC: Grant reference: ES/K008897/1)

References

Higher Education Funding Council for England (HEFCE) (2011) '*Research Excellence Framework. Assessment framework and guidance on submissions*', Higher Education Funding Council for England (HEFCE), United Kingdom.

British Academy 'Peer review: the challenges for the humanities and social sciences, A British Academy Report' (September 2007).

Boden, M. (1990) '*Peer Review: A Report to the Advisory Board for the Research Councils from the Working Group on Peer Review*', Advisory Board for the Research Councils.

Charmaz, K. (2006) *Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: Sage.

Chubin, D. E. and Hackett, E. J. (1990) *Peerless Science: Peer Review and US Science Policy*. Albany, NY: State University of New York Press.

—— (1994) 'Grants Peer Review in Theory and Practice', *Evaluation Review*, 18/1: 20–30.

Clegg Smith, K., Friedman Singer, R. and Edsall Kromm, E. (2009) 'Study Centered on One U.S. Comprehensive Cancer Center Getting Cancer Research into the News: A Communication Case', *Science Communication*, 32/2: 202–31.

Collins, H. (2004) 'Interactional Expertise as a Third Kind of Knowledge', *Phenomenology and the Cognitive Sciences*, 3: 125–43.

—— (2010a) 'Interdisciplinary Peer Review and Interactional Expertise', *Sociologica*, 4/3: 1–5.

—— (2010b) *Tacit and Explicit Knowledge*. Chicago and London: University of Chicago Press.

—— and Evans, R. (2007) *Rethinking Expertise*. Chicago and London: University of Chicago Press.

Dahler-Larsen, P. (2006) 'Evaluation after Disenchantment: Five Issues Shaping the Role of Evaluation in Society', in I., Shaw, J. C. Greene and M. M. Mark (eds) *The Sage Handbook of Evaluation*, pp. 141–60. London, UK: Sage.

—— (2011) *The Evaluation Society*. Stanford, CA: Stanford University Press.

De Rickje, S., Wouters, P. F., Rushforth, A. D. et al. (2016) 'Evaluation Practices and Effects of Indicator Use—a Literature Review', *Research Evaluation*, 25/2: 161–9.

Demeritt, D. (2010) 'Harnessing Science and Securing Societal Impacts from Publicly Funded Research: Reflections on UK Science Policy', *Environment and Planning A*, 42/3: 515–23.

Derrick, G. E. and Samuel, G. (2016) 'The Evaluation Scale: Exploring Decisions About Societal Impact in Peer Review Panels', *Minerva*, 54/1: 75–97.

—— and —— (2017) 'The future of societal impact assessment using peer review: Pre-evaluation training and IRR considerations', *Palgrave Communications*, 2017: 17040.

—— (2018) *The Evaluators' Eye: Impact Assessment and Academic Peer Review*. London, UK: Palgrave Macmillan.

Dill, D. and Beerkens, M. (2013) 'Designing the Framework Conditions for Assuring Academic Standards: Lessons Learned about Professional, Market, and Government Regulation of Academic Quality', *Higher Education*, 65/3: 341–57.

Donovan, C. (2011) 'State of the Art in Assessing Research Impact: Introduction to a Special Issue', *Research Evaluation*, 20/3: 175–9.

Greene, J. C. (2000) 'Challenges in Practicing Deliberative Democratic Evaluation', *New Directions for Evaluation*, 2000/85: 13–26.

Guena, A. and Martin, B. (2003) 'University Research Evaluation and Funding: an International Comparison', *Minerva*, 41: 277–304.

Haynes, A. S., Derrick, G. E., Chapman, S. et al. (2011) 'From "our world" to the "real world": Exploring the views and behaviour of policy-influential Australian public health researchers', *Social Science & Medicine*, 72 /7: 1047–55.

HEFCE (2009) '*Research Excellence Framework: Second consultation on the assessment and funding of research*', Higher Education Funding Council for England.

—— 'REF2014: Panel Criteria and Working Methods', <http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf> accessed 1 December 2017.

—— (2011) 'Assessment framework and guidance on submissions', Research Excellence Framework 2014.

Hill, S. (2016) 'Assessing (for) Impact: Future Assessment of the Societal Impact of Research', *Palgrave Communications*, 2016: 16073. doi: 10.1057/palcomms.2016.73.

Hojat, M., Gonnella, J. S. and Caelleigh, A. S. (2003) 'Impartial Judgment by the "Gatekeepers" of Science: Fallibility and Accountability in the Peer Review Process', *Advances in Health Sciences Education*, 8: 75–96.

Huutoniemi, K. (2012a) 'Communicating and Compromising on Disciplinary Expertise in the Peer Review of Research Proposals', *Social Studies of Science*, 42/6: 897–921.

—— (2012b) 'Interdisciplinary Accountability in the Evaluation of Research Proposals', PhD Thesis, University of Helsinki.

——, Thompson Klein, J., Bruun, H. et al. (2010) 'Analyzing Interdisciplinarity: Typology and Indicators', *Research Policy*, 39/1: 79–88.

—— (2010) 'Evaluating interdisciplinary research', in R. Frodeman (ed.) *The Oxford Handbook of Interdisciplinarity*, vol. 10, pp. 309–20. Oxford, UK: Oxford University Press.

Jasanoff, S. (1990) *The Fifth Branch: Science Advisors as Policymakers*. Cambridge, MA, London, England: Harvard University Press.

—— (2003a) '(No?) Accounting for expertise', *Science and Public Policy*, 30/3: 157–62.

—— (2003b) 'Technologies of Humility: Citizen Participation in Governing Science', *Minerva*, 41: 223–44.

—— (2012) *Science and Public Reason*. London and New York: Routledge, Taylor & Francis Group.

Kearnes, M. and Wienroth, M. (2011) 'Tools of the Trade: UK Research Intermediaries and the Politics of Impacts', *Minerva*, 49: 153–74.

Klein, T. (1990) 'The interdisciplinary Process', in P. H., Birnbaum-More, F. A. Rossini and D. R. Baldwin (eds) *International Research Management:*

Studies in interdisciplinary methods from business, government and academic, pp. 20–30. Oxford, UK: Oxford University Press.

Lamont, M. (2009) *How Professors Think: Inside the Curious World of Academic Judgement*. Cambridge, MA: Harvard University Press.

Lamont, M. and Huutoniemi, K. (2011a) 'Comparing Customary Rules of Fairness: Evaluative Practices in Various Types of Peer Review Panels'. In C., Camic N., Gross and M., Lamont (eds) *Social Knowledge in the Making*. Chicago, USA: University of Chicago Press.

—— and —— (2011b) 'Opening the Black Box of Evaluation: How Quality is Recognized by Peer Review Panels', *Bulletin SAGW*, 2: 47–49.

Langfeldt, L. (2006) 'The Policy Challenges of Peer Review: Managing Bias, Conflict of Interests and Interdisciplinary Assessments', *Research Evaluation*, 15/1: 31–41.

Laudel, G. and Glaser, J. (2014) 'Beyond breakthrough research: Epistemic properties of research andtheir consequences for research funding', *Research Policy*, 43: 1204–16.

—— (2006) 'Conclave in the Tower of Babel: How Peers Review Interdisciplinary Research Proposals', *Research Evaluation*, 15/1: 57–68.

—— and Origgi, G. (2006) 'Introduction to a Special Issue on the Assessment of Interdisciplinary Research', *Research Evaluation*, 15/1: 2–4.

Luukkonen, T. (2012a) 'Conservatism and Risk-Taking in Peer Review: Emerging ERC Practices', 1, 48–60.

—— (2012b) 'Conservatism and risk-taking in peer review: Emerging ERC practices', *Research Evaluation*, 21/1: 48–60.

Manville, C., Guthrie, S., Henham, M-L. et al. (2015a) *'Assessing Impact Submissions for REF2014: An Evaluation'*, Cambridge, UK: RAND Europe.

Molas-Gallart, J. (2012) 'Research Governance and the Role of Evaluation', *A Comparative Study', American Journal of Evaluation*, 33/4: 583–98.

Nedeva, M., Baker, K. and Osman, S. A. (2014) 'Policy pressures and the changing organisation of university research', in C. Musselin and P. Teixeira (eds) *Reforming Higher Education. Higher Education Dynamics*, vol 41. Dordrecht: Springer.

Olbrecht, M., and Bornmann, L. (2010) 'Panel Peer Review of Grant Applications: What Do We Know from Research in Social Psychology on Judgment and Decision-Making in Groups?', *Research Evaluation*, 19/4: 293–304.

Penfield, T., Baker, M. J., Scoble, R. et al. (2013) 'Assessment, Evaluations, and Definitions of Research Impact: A Review', *Research Evaluation*, 23/1: 21–32.

Porter, A. L. and Rossini, F. A. (1985) 'Peer Review of Interdisciplinary Research Proposals', *Science, Technology, & Human Values*, 10/3: 33–8.

Research Councils, U. K. (2006) 'Report of the Research Councils UK Efficiency and Effectiveness of Peer Review Project'. <www.rcuk.ac.uk/documents/documents/rcukprreport-pdf>

Research Excellence Framework (2011) 'Assessment framework and guidance on submissions', REF 02.2011.

—— 'REF2014 Impact Case Studies', <http://impact.ref.ac.uk/CaseStudies/FAQ.aspx> accessed 1 December 2017.

—— (2012) 'Panel criteria and working methods'. <http://www.ref.ac.uk/2014/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf>

Samuel, G. N. and Derrick, G. E. (2015) 'Societal Impact Evaluation: Exploring Evaluator Perceptions of the Characterization of Impact under the REF2014', *Research Evaluation*, 24/3: 229–41.

Shore, C. (2008) 'Audit Culture and Illiberal Governance', *Anthropological Theory*, 8/3: 278–98.

Smith, R. (2001) 'Measuring the Social Impact of Research - Difficult But Necessary', *British Medical Journal*, 323: 528.

Strathern, M. (2000) 'The Tyranny of Transparency', *British Educational Research Journal*, 26/3: 309–21.

Treasury, H. M. (2004) *Science and Innovation Investment Framework 2004-2014*. London: HM Treasury.

Watermeyer, R. (2016) 'Impact in the REF: Issues and Obstacles', *Studies in Higher Education*, 41/2: 199–214.

Wilsdon, J. (2016a) 'Time for a Stern, hard look at the REF', *WONKE*. <http://www.wonkhe.com/blogs/time-for-a-stern-hard-look-at-the-ref> accessed 12 November 2017.

——(2016b) *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. London, UK: SAGE.

Zuckerman, H. and Merton, R. (1971) 'Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System', *Minerva*, 9 /1: 66–100.