FINAL REPORT for the RESEARCH PROJECT:
Towards a cartography of impact: Analysing the process of impact peer-review in the REF.

Richard Watermeyer
September 2016

1. Rationale

Understandings of what constitutes impact, and for that matter, *excellent* impact are notional and premised for the most part, on academic guesstimates of how research findings translate into 'real-world' outcomes. Notwithstanding, UK academics are challenged to conceive, mobilize and evidentially report on ways in which their research is *impactful* – in the REF context via narrative based impact case-studies. This they do, however, with only a speculative sense of the impact-value of their work; limited knowledge and competencies in making their research impactful or exploiting its impact potential; and limited skills in reconnoitring, recording, evaluating and articulating their impact activity.

Academics' navigation through the unfamiliarity and perceived opacity of impact assessment is furthermore compounded by limited, if not, a total lack, of foresight and penetration into the way their impact claims will be interpreted and assessed by academic peer-reviewers and user-group assessors. Whilst numerous studies have explored the nature of the peer-review process (Armstrong 1997; Bornmann & Daniel 2005; Cole, Cole & Simon 1981) these have almost exclusively focused on how academics assess the quality of research. Significantly less attention has been paid to other iterations of peer-review (Lamont 2010). Unsurprisingly, therefore, intelligence related to the kinds of criteria and influences shaping the assessment of impact in the REF, is impoverished.

In responding to this evidential shortfall, this study explored the methodologies and 'pathways', consciously and unconsciously, co-opted and navigated by REF panelists in making assessments of the legitimacy, authority and value of impact/impact claims submitted by their disciplinary peers. Through a series of semi-structured interviews with members of disciplinary sub-panels (both academic and user-assessors) it considered the extent to which panelists' evaluations, confirm or deny, veer towards or bypass, unwittingly or not, impact peer-review as a meritocratic process (Musselin 2013) and simultaneously the multifarious factors that contribute to variations in panelists'/panels' deliberations and determinations of impact worth. The interview process followed an open-ended and inductive approach to 'ethnographic' data collection, allowing for an uninhibited and non-proscriptive elucidation and mapping of the kinds of criteria (stated and tacit, individual and collective) used by, and influences affecting, panelists in the process of impact evaluation.

The research was conceptually guided by previous studies of academic evaluation (cf. Whitley 1984; Merton 1968, 1973, 1988, 1996), particularly the work of Michèle Lamont (2009) – also Lamont, Mallard and Guetzkow (2006) – and borrowed from sociological studies of science and technology, theories of actor-network (Latour 2005) and translation (Callon 1986). Concurrently, theories of 'distinction' (Bourdieu 2010); 'credibility' (Latour & Woolgar 1979); and 'expertise' (Collins & Evans 2002) were used in (de)constructing an epistemology of impact.

The study would focus on:

1. the extent to which panelists were distracted and/or motivated by evanescent criteria and an evaluation of the impact case-study as a document less an assessment of the veracity and significance of impact claims expounding the virtues of the underlying research i.e. to what extent does the elegance, erudition, cultural capital and cogency of the impact narrative (and its author) subvert panelists' interrogation of the supporting evidence?
2. how idiosyncracies of taste and variation in panelists' intuitive, prejudicial and protectionist sense of excellence and what counts, and a sense of 'marking their territory' (cf. Shapin & Shaffer 2017; Jasanoff 1990), were managed and the extent to which panelists were able to achieve what Weber (1978) calls 'rational legitimacy' and adhere to an impersonal and consistent or universalistic standard in the evaluation of impact case-studies?
3. the prevalence and observed effect of unequal professional and dialogical capital, verbal virtuosity, and power dynamics between panelists in terms of advocating and prioritising a higher award for some impact case-studies over others. In this instance, especially, the researcher explored the extent to which academic reviewers were scaffolded, influenced or led by the expertise of user-group assessors as potentially more knowledgeable of, and familiar with, a diversity of research impacts. Conversely, it was considered whether/and with what frequency/effect user-group assessors were subjugated by the intellectual capital of academic panelists and their greater knowledge and awareness of the quality of the underlying research. To what extent were academic and user-group assessors harmonised, and in such terms, to what extent did the underlying research, and evaluators' proximity/affinity with it, influence a final impact assessment?
4. the extent to which panelists practice homophily in rewarding those impact case-studies with the greatest personal resonance and/or where claims of impact were more easily or immediately discernable.
5. how the expertise and connoisseurship expected of, and distinguishing, panelists as peer-reviewers was reconciled with, or compromised by, their inexperience in evaluating the impact of research.
6. case-studies perceived to be most obstructive to consensus-making. Which disciplines featured as 'problem fields' (Lamont 2010) and which produced the most efficient consensus?

2. Executive summary

Participants' articulation of their experience of assessing impact in the context of REF2014 was characterized by the extent to which they had felt confident or concerned during and after the exercise. A variety of factors were highlighted as either contributing to or eroding their confidence as impact assessors and therefore also what they understood as the success and efficacy of the evaluation process. These are illustrated in Tables 1 and 2.

Key to what interviewees identified as a growing confidence in the process of impact evaluation in the REF was the sense of the exercise being a collective endeavor and a shared responsibility undertaken, by many for the first time, in good spirit. Positive relationships particularly between academic panel members and user assessors were the reported norm and viewed as prerequisite to an explicitly collaborative

undertaking. In addition to the creation of a positive working environment, interviewees also reported that the guiding framework and criteria for impact evaluation provided by HEFCE and furthermore, an iterative process of impact evaluation, greatly aided their self-belief, particularly in the context of the credibility of their evaluative scores. A calibration process that allowed provisional scores submitted at a sub-panel level to be scrutinized and commented upon at a main panel level was also seen to contribute to a sense that the scores ultimately arrived at were the result of an extensive and robust process of collective deliberation. However, as will be discussed, the manner with which sub-panel decisions were scrutinized by main panel members was also viewed as a type of surveillance or being called to account that meant the scoring of individual impact submissions might be predicated less on the intrinsic value of a case study and instead a demand for consistency in the distribution of claims of impact excellence across sub-disciplines.

A cause for concern among panel members related to a sense of lost or leaking credibility of evaluative judgements caused by the unavoidable effect of disciplinary bias; a conservationist instinct for the sub-discipline; and a sense of the impossibility of complete impartiality and/or non-advocacy for the sub-discipline. Furthermore, whilst interviewees reflected on the importance of the HEFCE criteria in making what were felt to be credible impact assessments, they also pointed to a number of grey areas and/or difficulties in the interpretation of supporting evidence in impact submissions; determinations of what counts as impact and especially impact deserving of a four star grading. Interviewees were united in speaking of the huge labour and time demands of the REF evaluation process and signposted issues of resource and capacity – particularly numbers of people assigned to review any one impact submission at the sub-panel level – and this despite the perceived assurances provided by the calibration of scores between sub-panel and main-panel. Finally, interviewees spoke often of the difficulty they encountered in penetrating a kind of impact discourse and the kinds of elaborations made by researchers in their attempts to 'sell' their impact narratives.

Table 1. Factors stimulating confidence of panel assessors

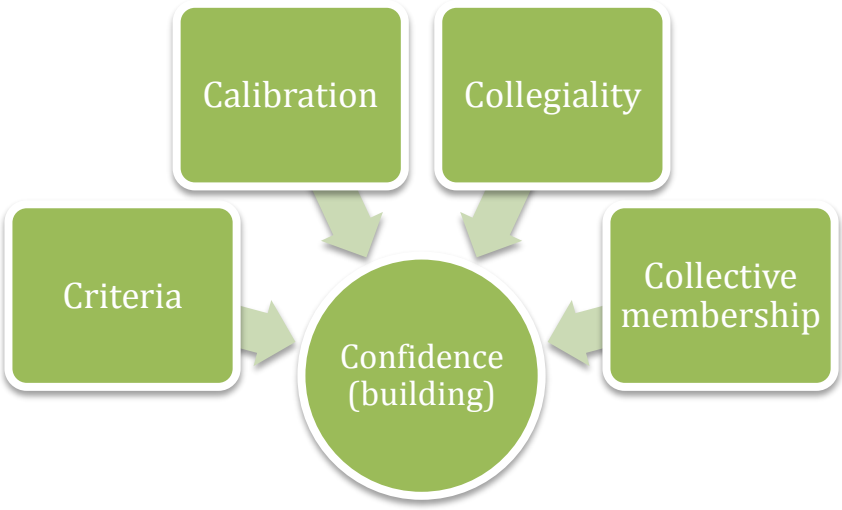
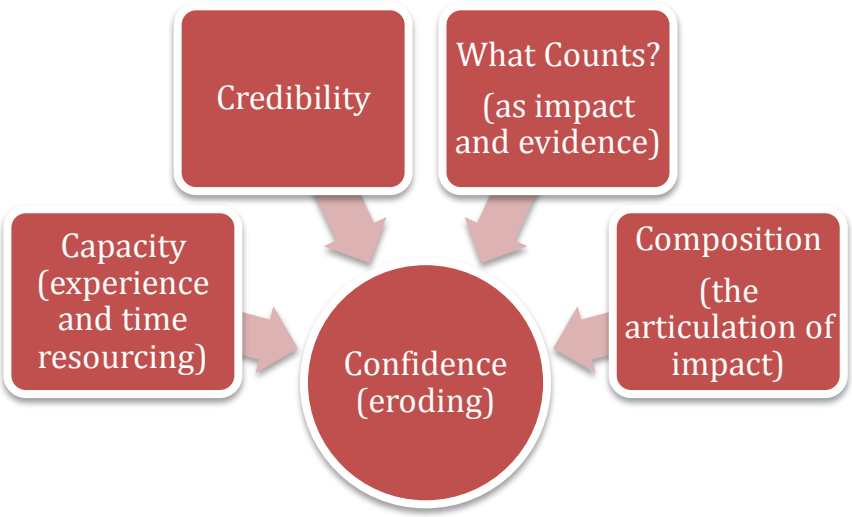


Table 2. Factors eroding confidence of panel assessors



Interviewees' accounts of evaluating impact in REF2014 reveals an emotional, personal and of course professional investment; and of course why shouldn't it? This was a massive undertaking and huge responsibility undertaken against a background of some not inconsiderable controversy and uncertainty. Their retold experiences, therefore, unsurprisingly, gravitate towards, indeed appear dominated by a sense of self-justification and striving towards confirmation of the efficacy of the process of impact evaluation and their role within it.

Reported herein are thus what interviewees reported as a growing confidence in the task of impact evaluation, greatly aided by a sense of a united effort and united front. The successes of impact evaluation are attributed to the cultivation of positive behavioural conditions, specifically the building of a collegial, respectful and fluent relationship between panel members, particularly academic and user assessors. However, whilst much of the success of the process was attributed to the successful integration and collaborative fluency of sub-panel and main-panel members, the process was seen to suffer interruption and provoke doubt where issues related to structure; meaning; and resources intervened.

3. Historical Context

The UK's Research Excellence Framework (REF) – successor to the Research Assessment Exercise (RAE) – is the route to which determinations of excellence in research are made and the means by which, the UK Government through its arm length funding agencies, such as the Higher Education Funding Council for England (HEFCE), distributes somewhere in the region of £1.6 billion of what is termed Quality Research (QR) Funding. The REF operates as one part of a dual funding system for research undertaken in UK universities, the other being that of an open funding competition administered by the Research Councils UK (RCUK), comprising seven discrete disciplinary funders such as for instance the Economic and Social Research Council (ESRC). With the REF and former RAE – the latter begun in 1986 and thereafter repeated at 4-6 year intervals – the UK higher education sector has been at the vanguard in developing a system of performance based funding now

imitated across the world and increasingly adopted by other national higher education sectors.

In the wake of the 2008 RAE and the events of global economic downturn those in the driving seat of higher education policy in the UK, influenced in large part by the paymasters of Her Majesty's Treasury, began to think through a response to pressures for increased accountability in the research funding landscape. With a dramatic shrinkage in the public purse and the advent of what would become and arguably until very recently remain as a politics of austerity, justification in the allocation of QR funds, would it was felt, require more robust defence.

Consequently, in strengthening the claims for public expenditure on academic research, university researchers would find themselves obliged to evidence a return on such investment. Two methods for confirming the efficacy of public funding for research were conceived. One would be that researchers in application for research grant funding made available by RCUK, would have to provide prospective accounts of the impact of their research plans. Researchers would have to articulate the ways with which they would produce impact in the form of narrative accounts, or pathways to impact statements that would and continue to feature as mandatory requirement of funding applications. A veritable smorgasbord of social and economic impacts were conceived and used to help researchers think through what might count in the specific context of their own research. These impact pathways were in the context of RCUK funding applications linked to or seen to lead on from what researchers perceived as the academic impact of their research. In other words RCUK bifurcated impact into the effect a researcher might have on his/her own professional community and the impacts of research received by non-academic or public constituencies. Crucially the assessment of a research proposal submitted to RCUK would have to consider impact as the effect upon both academic and public constituencies. The second method for confirming the efficacy of public funding for research was tied to QR funding and therefore the evaluation of research in the context of the REF.

In the previous RAE2008, evaluation conducted by senior academic peer-reviewers arranged over a large number of disciplinary panels, focused on academic outputs, journal articles, books and other indicators of research excellence and the overall research environment and its conduciveness to producing world leading research. In the REF, impact or rather more specifically the societal and economic impact of research constituted a new measure of assessment. Prior to its formalization as a core component of the REF exercise, impact assessment was piloted across a number of selected universities and subject disciplines. The results of the pilots confirmed that assessing impact was both viable and necessary. Impact would feature alongside the assessment of research outputs and research environment in what would become REF2014 (Watermeyer 2014).

If previous murmurings of an impact agenda had set off alarm-bells among members of the UK academic community news of impact being an official component of REF provoked in some quarters panic and in varying measure a vitriolic response,

especially from those who construed in impact a disciplinary enemy and further infringement upon their research lives. For those in less applied disciplines or for whom a public interface was more arms-length than 'arms-embraced', news of impact in the REF was met with particular concern (Chubb and Watermeyer 2017; Watermeyer 2012, 2016). How could impact be claimed by researchers, where the nature of their research was not conducive to public interface and where any sense of a user community was at best vague or ill-defined? Furthermore, at what disadvantage would be those communities for whom the emergence of impact occurred with a more elongated lead-in time or lengthier gestation period? Cue, the assertion of many who would cite the peripatetic and frequently serendipitous process of scientific discovery, largely at odds with a system of performance evaluation, that demanded more in the way of 'over-night' results (cf. Collini 2012; Ladyman 2009; Moriarty 2011).

A number of other concerns were raised in the context of what would count as impact evidence, or rather what would be the best or most compelling forms of impact evidence? How would public engagement feature? Would it constitute a form of impact in itself or a route to impact (cf. Watermeyer and Lewis 2017)? These kinds of questions as others would accentuate with the production of guidance from HEFCE pertaining to, for want of a better word, the rules of impact in the REF. Firstly, it was formally announced that impact would constitute 20% of the overall value of the REF, with the main chunk of the REF focusing in on research outputs at 65% and research environment statements at 15%. A variety of calculations premised on this percentage split, meant that impact not only constituted a significant part of the REF evaluation but concurrently was of significant financial worth. Some calculated that a successful impact submission could be worth anything up to £375,000 or more within the REF period (Dunleavy 2012; Watermeyer and Hedgecoe 2016). The value of impact in the REF, where organized as narrative impact case studies and an impact template, was determined in other ways, such as for instance in comparison to the value attributed to research outputs. Consequently, it was reported that where in a 0-4* scale of excellence, where 4* represented the highest achievement, one 4* impact case study had the same monetary return to a submitting institution as would seven 4* research outputs. Overall, therefore, the significance of getting impact in the REF right was deemed to be huge (Watermeyer 2014).

Significant financial outlay was consequently committed by many universities in preparing for impact as a substantive component of REF2014. These would include the recruitment of copy-writers and science-writers, those who could help academics or even fully translate their claims to impact in ways that were jargon-light, easily accessible, easily understood and fundamentally as easy as might be, evaluated. Crucially, because, the assessment of impact in REF2014 would be undertaken not only by senior academic peer-reviewers but user-assessors, those for instance working outside of academia but able to make authoritative adjudications as to the extent of a piece of research's 'public' contribution. Universities in some instances also ran mock-versions of impact assessment, in a lite-bite fashion mirroring the actual process to be undertaken in the REF itself (cf. Watermeyer and Hedgecoe 2016). Simulated internal review exercises like these, which included

panel members of previous RAEs and what would be REF2014 itself, provided an opportunity to road-test the kinds of criterion for impact stipulated by HEFCE as the REF and impact in the REF architect. Such criterion was informed by what some called an 'impossiblist language' of impact evaluation (Dunleavy 2012); a reference to notions of impact 'reach'; 'significance' and 'transformative potential' – qualifiers which it was felt lacked the kinds of clarity and precision that would otherwise facilitate more in the way of a smooth process of evaluation (cf. Watermeyer 2014). Which gets to the nub of the research this report is designed to articulate, which focused not on second-guessing the REF's evaluation process for impact but understanding it first-hand through direct consultation with those responsible for its undertaking.

4. Methodology

Whilst previous recent research has provided ethnographic observations of the process undertaken in simulations of impact evaluations (cf. Watermeyer and Hedgecoe 2016), the confidentiality associated with the evaluation conducted in REF panels, meant that any direct observation of the 'real' version of impact review was strictly off limits. What would, however, be possible though potentially difficult to negotiate would be access to evaluators outside of the formal REF context and as transpired, 'after-the-event'. Being as it were, beyond the REF, however, would provide a benefit of time-lapse between panel members' involvement in the REF and them being asked to comment on the experience. In other words, a period to allow reflection on the lived experience. Of course, this would also mean that certain details associated with the experience of the exercise might be obscured by fading memory.

The project began by the identification of potential participants. An original aim had been to interview Chairs and potentially Deputy-Chairs of the sub-panels, populating all of the Main Panels A-D and therefore those with responsibility for overseeing the review of impact across the main disciplines of:

- Medical, health, biological, agricultural, veterinary and food sciences
- Physical, Mathematical, Computer Sciences and Engineering
- Social Sciences
- Arts and Humanities

However, at the time of negotiating access, it became known that a similar research project focused on the experiences of panel reviewers across the sub-panels of Main Panels A and B had been in progress and was beginning to report (cf. Derrick and Samuel 2016; Samuel and Derrick 2015). Without wanting to end-up with a duplicate study, it was decided to focus the study in two alternative ways. One was to provide a more intensive focus on the experiences of evaluators with membership of two sub-panels within Main Panel C, specifically the sub-panels of Education and Sociology. A more in-depth focus across two panels it was felt would provide less of the headline account that might be provided by Chairs/Deputy Chairs and a fuller range of the experiences of all panel members, including, importantly, user-assessors. In addition to interviewing the majority of the members of these panels,

interviews were also conducted with Chairs/Deputy Chairs across a further four of the sub-panels in Main Panel D. In total, thirty-two in-depth interviews were conducted, which lasted anywhere from forty-five minutes to eighty minutes; the majority lasting about an hour.

Gaining access to interview participants was not always straightforward. While many of the members of the Education sub-panel were familiar to or known by the researcher and consequently provided for a relatively uncomplicated access negotiation, the same was self-evidently not true of other sub-panels, particularly those in Main Panel D. A good number of requests for interview faced a non-response. For those that did engage with the project a number of assurances were made, such as for instance total anonymity and non-referencing of sub-panel membership. Furthermore, all requests for interview included official HEFCE guidance of what sub-panel members were entitled to disclose. Where interviews were generally characterized as conversational, interviewees did as might be expected, make mention of certain aspects of their experience that could not be reported upon for reasons of confidentiality and were subsequently redacted from analysis.

Interviews followed a semi-structured design which adhered, though not restrictively, to a predetermined set of questions. Indeed, interviewees were intentionally given full reign and ample opportunity to provide expansive answers, sometimes,which deviated from 'the script' and/or explored other avenues and aspects of their REF memories. An interview schedule was designed that asked respondents to reflect upon their overall personal experience of evaluating impact in the REF context and what they had learnt; what aspects they had found challenging or had struggled with; what aspects they found to have been effective; and what things they would be inclined to change or do differently.

Audio recordings were made of all the interviews, which were subsequently transcribed by a professional transcription agency. Interview transcripts were subsequently poured over and thematically analyzed. Findings are presented within this report as aspects of the experience of conducting impact evaluation in REF2014 bifurcated into two overarching categories: experiences that fostered a confidence in the process being undertaken and aspects of the experience of impact evaluation that were seen to be problematic of cause for concern.

The study adhered to principles of ethical practice in educational research (BERA 2011). Prior to the onset of fieldwork, research plans were submitted to the scrutiny of a Research Ethics Committee at the researcher's host institution. Furthermore, permission for the research was sought and granted by the Higher Education Funding Council for England (HEFCE). All research participants were reminded of the voluntary nature of the research and their right to withdraw at any point. All research participants provided informed consent prior to their participation.

Fundamentally, the research project was intended as a social study of the process of impact evaluation and therefore an exploration of the various factors, social and

systemic that enabled or potentially inhibited the REF panel members as impact evaluators.

5. Findings

5.1 Behavioural conditions

5.1.1 *Confidence*

Interviewees reported increasing confidence in the task of evaluating impact and a clearer sense of what to look for and how to score with repetition. Interviewees routinely expressed impact evaluation in REF2014 as the best form of training and kind of learning on the job that would be invaluable to their potential contribution to impact evaluation in future iterations of the REF. Only one or two among interviewees expressed doubt in terms of quite what they had learnt from the process and whether indeed they were any more expert as impact assessors now than at the outset of the REF.

5.1.2 *Collegiality*

A collegial spirit and certainly a sense of equal participation among all evaluators – academic panel member and user-assessors – was perceived by all research participants as pivotal to an effective evaluation process. A positive finding was that integration of all sub-panel members was typically smooth and formed the basis of a successful joint effort:

> So I went into the process thinking people would more or less act as individuals and there'd be some moderation between them of course and that they'd be a team in the sense that sometimes you'd have to pass things, you know, there'd be some specialisation of course by virtue of your expertise and sometimes you'd have to pass things between people as well. All of that was true but it was much, much more. So I think what I was heartened by was that there was a tremendous collegiality, which included sharing interpretations, you know, debates about what constituted different star ratings and discussion of relative strengths and weaknesses and the importance of different things, discussions about the extent to which evidence could be indirect.

Furthermore, the role of the sub-panel Chair was understood by all those interviewed as integral not only to the successful moderation of dialogue and ensuring equal participation but in fostering and securing a collegial spirit among each group:

> Basically he led the group, so he was our leader, so he was helpful in resolving any disagreements in discussions, he'd help make people feel part of the group. [He] was brilliant. He really made an effort to make sure all people coming in knew who everybody was, everybody knew other's roles, and you really felt part of the team. So I felt he made the process really pleasant, you know, for all those taking part, and you know, he arranged little socials, and they went off to see a Shakespeare play and things like that, he was just … good to keep the team together because it's a bit of a

> gruelling process. So … I think it's important to have a strong leader. Keep everybody on track, keep everybody happy … and … be somebody that everybody can trust to kind of make the final decision where there are disagreements. Or, you know, help people come to that consensus.

5.1.3. *Collaboration*

The collaborative dimension of assessing impact in the REF where it involved active discussion between evaluators was perceived as both a strength of the evaluation process and also something that distinguished the evaluation of impact from the evaluation of outputs:

> I think what's interesting about impact is it was much more collaborative. Assessing outputs is a pretty solitary exercise but particularly when there's agreement and all you're doing is entering numbers and figures on a spreadsheet. In the majority of cases there is agreement and you just move on, whereas it was in the impact case studies there was much more discussion. I was pleased with that because it was the new element of the exercise and having that ownership through dialogue conversations, coming to collaborative decisions seems very appropriate, not least because it means that it gives you confidence in the outcomes I think.

5.1.4. *Collective ownership and being held to account*

Throughout the interviews, interviewees repeatedly articulated a sense of collective ownership and shared responsibility in ensuring the credibility of the exercise and the final impact scores, reinforcing a view of the importance of a collegial and collaborative evaluative process:

> For the whole exercise it was absolutely crystal clear that the responsibility for the final decision rested with the sub-panel as a whole and had to be endorsed by the main panels and then by the system in totality, as it were. So I don't think anybody would have felt their judgement stood alone without being part of a broader judgement.

While the vast majority of decision-making related to impact scores occurred at the sub-panel this was not in isolation to the guidance and/or instruction of the main disciplinary panel. Organizational structures such as the calibration of scoring between sub and main panels was consequently seen as instrumental to confirming accuracy and parity of and across sub-discipline scores and a way to reflect on scoring trends (at the sub-panel level) but perhaps only in so much as confirmation was born from comparison with the scoring trends of other sub-disciplines:

> We were sort of pressed to consider how this would look afterwards if, for example, if philosophy turned out to be the most impactful academic discipline in the country, how plausible would that be, in the media…

> The moderation with the other sub-panels enables you to play fair on a scale. You get the range, you know? Because what they tell you is, they,

> sort of, pose questions for you, really, they don't tell you exactly. But the questions might be, "Are you sure you've really got quite that many four stars because nobody else seems to have?" And it makes you go and have a little think about it. Have you really? . . . So the feedback loop coming from the main panel was important to me in informing my judgement about advice to give to my colleagues in the sub-panel.

So much of the accuracy of the scores appears, therefore, to do with how the subject conformed to the expectations of 'disciplinary keepers' and/or those of public stakeholders. In a way, therefore, the calibration exercises between main and sub panels might be interpreted not only as a way to ensure harmonization of the scoring regime and the avoidance of conspicuous differences in scoring claims between sub-panels but a means to justify and even defend the 'honesty' of disciplinary claims to impact via adjustments of modesty. At a more mundane but no less important level, the 'swopping of scores' between sub-panels was seen to provide clarity and confidence among panel members in what 'should' be identifiable as excellent impact:

> You know, that I think helped a great deal in clarifying what it was we were looking for and how we would recognise the legitimate impact when we saw it.

From the accounts of interviewees, the calibration exercise employed in the assessment of impact in REF2014, therefore, provided almost a double check or check-on-the-check of impact claims, where the assessments of impact evaluators at the sub-panel level were being held to account by the main panel 'watch(wo)men'). Consensus in the context of impact evaluation scores was consequently reported as being required not just at the level of sub-discipline or sub-panel but at the overarching disciplinary level. However, it was claimed that this was not to necessarily quash differences, more that such differences might be justified:

> Had our results been out the main panel would have not accepted them. The ruling made was that there may be differences between sub-panels and that was okay, but we had to understand those differences. There had to be reasons for them. What we were all collectively on the lookout for were variations, which could not be accounted for and, therefore, could not necessarily be justified. It might just be a sub-panel trying to pull a fast one which, obviously, would not have been acceptable. So there were variations but we had to understand those variations.

What this then was seen to produce was a system of constant posting, monitoring, reflection, action and potential moderation of scores, designed to eradicate the risk of significant deviation and ensure consistency in sub-panel scores:

> What you had was a process where people were uploading, downloading, uploading, downloading. Each member who put their stuff in, it went into a transparent space where sub-panel Chairs could monitor, could then both

> support colleagues who might not be getting through the workload very quickly, or could see outliers and could probe and say you might just be being a bit mean there, possibly. Unless you think your particular area is a bit dodgy or not too strong.

While such a system of check and control might be fairly rationalised on the basis of critical reflection and alignment it may also, however, overly influence and potentially even adversely manipulate and/or misrepresent scores. A clear hierarchy in the terms of evaluative expertise between main and sub panels might mean for instance that what some sub panel members might recognise and intuit as impact excellence would be rejected by main panel members and result in the generation of anomalies. There is none more so explicit example of this in the testimony of interviewees than one who spoke of the out and out rejection by one very senior member of a Main Panel of public engagement as a form of impact. This in the accounts appears to have much to do with either sub-panel members' interpretive license or strict adherence to HEFCE's REF impact criteria.

5.2 Successes and challenges in building the academic/user interface

While user assessors and academic panel members shared much in the way of a common experience, there were some aspects of their experience that differed or were more intense. These related mainly to reflections on being socialized and integrated into the REF process; limitations of time and the time/labour intensiveness of impact evaluation and the REF evaluation process in general; a perceived absence of a theory of change within case studies; and being inherently stricter in their valuations of impact.

The extent of user-assessors being socialized and integrated into the REF process differed across the sub-panels. While some like SPUA-1 discussed the importance of informal socialization to the formal business of evaluating the case studies on a collaborative basis with academic counterparts others like SPUA-5 spoke of a lack of informal introductions and limited sense of group identity, particularly that of the user-assessors themselves:

> I don't know how relevant it is but the social bit helped as well. After the meeting it was important that you would go for that meal or a pre-meal drink and just sit and chat with people. Some you knew but most of them you didn't and got a chance to chat and that definitely helped . . . that's a very important part because then you genuinely feel that you're part and parcel of what's taking place and you're genuinely an equal part and what you have to say is listened to so I found that bit to be very helpful and important.

> I think the one thing that would have helped perhaps the discussions and the relationships was if there'd been a bit of getting to know each other at the beginning. For example I didn't actually know any of the other people who were impact assessors on my panel and I think if we could have had a bit of getting - no matter how small – if we could have had a bit of an

> introductory session together to get to know each other and where we were coming from etc. we could have worked a bit more effectively to pilot the approach. So as impact assessors we didn't actually…or I'm not aware I mean I didn't I don't know whether the others did, if they knew each other, any of them knew each other. We didn't actually have any identity as a group and I think it would have been useful for perhaps impact assessors themselves to have discussed some of the things that they were grappling with on impact rather than only with the other members of the panel that they paired up with.

While the time commitment made by panel members is well documented, there was among user assessors a sense of surprise and in part disbelief at quite how demanding impact evaluation in REF2014 was. Indeed for some, the task was found to be unmanageable where unlike the majority of academic panel members whose focus was exclusively on the REF, they continued to deal with the demands of their everyday professional roles. Indeed for some, their role as impact-assessors proved to be so demanding that it culminated in the cessation of their participation:

> My overwhelming view of the REF, that it was a really, it was a huge exercise, an incredibly demanding exercise for those of us that, that weren't working in the university sector and hadn't been able to amend our own timetables and job responsibilities to allow us the time to concentrate on the REF. I really found it very, very difficult for that, because I was trying to do lots of other things at exactly the same time.

5.3 Issues related to assessing the impact environment and recognizing the serendipitous nature of impact generation

A near consensus of opinion dictated that the impact template was an unhelpful aspect of the impact evaluation process and should be collapsed into the overall reporting of research environment. Many felt that the impact template was not only difficult to judge where evidence supporting many of the claims asserted was scant, if not entirely absent, but that it encouraged impact authors to preference and indulge in an overtly 'headline-grabbing' narrative style that was similarly difficult to unpick and confidently scrutinize. Furthermore, interviewees spoke of the distortive effect in the scoring of impact templates in terms of determinations of overall research performance.

> Impact case studies were far easier to asses than impact statements or templates as they were sometimes called, so impact statements or templates were it seemed to me highly dependent on word craft skills, you know, how well could you describe an idea, a concept, of supporting impact or concept of impact and then how you were supporting it more than what you actually perhaps did, you know. Very hard to check of course whether anyone's, you know, telling the truth or engaging in elaborate kind of embellishment. Whereas with case studies there's always some hope of checking, often it's eminently practical to do it if you wanted to. If you thought someone was

> over claiming you could check, if you thought that their evidence was iffy you could look at it, you know. So case studies were far easier to judge

One other interviewee spoke of how the impact template as a strategy document might not necessarily reflect excellent impact:

> One just kind of pragmatic thing was the kind of overall strategy template thing, was almost useless. I mean, you could just see it was well-written or it was badly-written… We ended up giving low scores for that, the department who then, we chose their case studies to be excellent, so then we wondered maybe it doesn't matter if they've got a bad strategy, if they're doing great on impact, and vice-versa, other ones who had really good strategies but actually the impact case studies seemed quite weak. And so that was, of all the whole, everything they actually I think that was the one thing that was hardest to get a grip on, and I would be sceptical about including that in the future.

This view reflected the opinions of many concerning the serendipitous nature of impact generation and that excellent impact is rarely something that can be necessarily planned:

> There were some case studies, which were pure serendipity in the sense of when I did this thing I didn't expect anything from it and then someone phoned me up one day and said, 'I want to do this with it'. And we took the decision that that was as good as a planned impact strategy from the moment you start to do the research because the rules didn't require that. The rules required the impact, not the planning.

5.4 Issues in the interpretation and application of evaluation criteria

The criteria for evaluating impact, formalized by HEFCE, was viewed by interviewees in both positive and negative terms and moreover something that was dynamic and evolving. Some even spoke of the criteria being made fit-for-purpose by the sub-panels themselves:

> I think you've got to recognise that the panel had quite an influence on the criteria. Because that's what you do the first two years of the REF panel's work is spent debating the criteria. And an awful lot of energy went into trying to get the impact criteria. Both contributing to the main panel or the overall criteria, and in particular, our education level criteria that related to that. And so, for example, through examples of what that looks like, there were heated debates in the panel about at the stage of setting the criteria. I think when it came to the assessment, people stuck to those criteria. And yes, of course in the end, it's still subject to the interpretation of those criteria.

Some spoke of the need for a wider range and more sophisticated set of evaluative criteria and indicators of impact that might allow for a more holistic and composite appraisal:

> If you're looking at a particular institution's outputs, you know, there may be a couple of hundred elements that you're looking at. Whereas on the impact, you know, you've got between five and ten at most and those are the elements that you're putting into it. And if there were certain kinds of criteria or indicators that we could look at that went beyond the kind of the standards, so you had a wider range of criteria. And you'd have a larger number of elements to add together to produce you profiles. So it's part of the thing that James Wilsdon's doing with his review of metrics and so on. That, you know, although we don't want metrics that will pre-empt the assessment, to find some kind of indicators and measures that we could use. Even qualitative indicators would be far better than having to rely on the relatively crude criteria that we had this time.

Indeed, some spoke of how the criteria resulted in a kind of selectivity in those impacts that were reported – ostensibly those most easy to evidence and concurrently those likely to receive higher evaluative scores – meaning that a lot of impacts failed to be reported:

> We felt that we weren't getting always a lot of the kind of impact that we knew was going on in the discipline but people hadn't been submitting to us. And I think it was that we were aware that institutions were perhaps playing it safe because this was the first time that impact had been included and so they wanted to include what they thought was the most obvious kind of impact and that would do well. And so we were, you know, aware that we weren't getting kind of a cross-section of the kind of impact that we knew went on in the discipline.

One major aspect of difference between user assessors and academic panel members' interpretation of the criteria focused on what the former identified as a failure of impact authors to engage in a theory of change, which they viewed as integral to any justification of impact and the evidence of impact causality:

> The people who I work with as partners and stakeholders around promoting research in our sector, we draw a lot on the work of Sandra Nutley. And that was the kind of thing I was expecting to see in the templates, that there would be a reference to some underpinning work on understanding how research gets used in a particular context or particular sectors. And then all the activities then would build on that, would be built around that theoretical underpinning and I think in all of the templates that I saw only one of them referred to something like that.

Finally, in the criteria of what counts as impact, the REF rules dictate that public engagement is in and of itself not a legitimate form of impact. However, this was

something that all interviewees commented upon as remaining a grey area and a focus of some considerable discussion which furthermore, reflected issues in disentangling public engagement as a form of or conduit for impact:

> That is something we discussed a lot, and I think there are a number of things that we thought were important and valuable activities, that didn't necessarily fit very well into the impact framework. I mean, one issue that often came up with public engagement was the extent to which it was tied to the research in the department, as opposed to generally sort of give PR to the discipline, so quite often - you know, you hear people on the radio or whatever talking about your discipline, they're not necessarily talking about their own research, but they're there because they're experts, and that was debated to and fro quite a lot . . . I think we tried to be sympathetic to people who were doing that kind of thing, but at the same time, you know, it was sometimes hard to recognise that within the rules.

5.5 Issues involving evidence

Interviewees routinely spoke of the challenge of interpreting the evidence put forward by researchers in justification of their impact claims in addition to the challenge of learning how to determine and trust its legitimacy and how to weight and assign value to different forms of evidence:

> The issue of evidencing in itself is difficult and that I think we were grappling with "What makes good evidencing of impacts?" So I would probably say that the evidencing end of it for me was the key bit that was problematic. So I was able to look at an impact case study and see that impact they'd claimed, whether it was appropriate, whether it matched with the research – because that's the other thing: was it really impact from research rather than just the esteemed person doing the project – looking at the quality of the research. All of that I think I felt much more comfortable with. It was the nitty-gritty of the evidence and knowing how to trust that evidence, and a sense of everybody – all institutions I think – also trying to work out what evidence they needed to give. So what kind of weighting do you give to a personal testimony over and above a letter of commendation; all those things, if it was solicited what did it mean? Did we know the person wasn't just a friend of the person of the case study? And then what do you make of a statement about a certain number of hits on a website or how you change the company and government's thinking about x. So it was that side of it I think probably more than anything for me that really was the challenge.

Difficulties in making what were felt to be credible assessments of impact were also attributed, rather unsurprisingly, not only to a lack of expertise in handling claims of impact generation but the type and/or quality of the evidence. For some, the evidence upon which a value judgement would be made was intangible and was what was felt to differentiate the peer-review of impacts from outputs:

> If I'm looking at an output, a piece of research published in the public domain, I'm able to judge the methods, the quality of argument, the results, the literature behind it all, do you know what I mean? All of that is quite tangible. Your opinion and my opinion may differ because that's the nature of professional judgment but it means that we do have something in front of us that we are looking at that is tangible, that starts on page one and finishes on page twenty or whatever. Whereas I think with impact it is literally so many words of persuasive narrative broken up into two or three sections, four sections, which are inadequate in themselves to giving any kind of substance. Nothing you can hang your judgment on other than having known the area or having taken the time to go and try and solve it.

The quality of the evidence or rather evidence that made a clear and cogent link to excellent impact was a concern for others who perceived, despite a commitment to a 'fair' evaluative process and inability to avoid privileging certain forms of evidence, and thence impact and ultimately research:

> What we were trying to do was to be fair. I suppose one problem was it was obvious that some things were much easier to evidence with a very concrete piece of evidence. So a policy change could often be very clearly documented because you had a policy where it was changed and you could see the attribution; other things were less easy to evidence, and what we didn't want to do was to penalise people's impact because the nature of the impact made it harder to evidence. So how can you be fair? If some things are better evidenced just out there – changes in professional practice, sometimes they are hard to evidence and people were trying very hard to provide robust evidence. So partly I think we were all kind of tangling ourselves with a dilemma of not wanting to privilege particular kinds of evidence. In effect what that did was privilege certain kinds of impact.

> The reason why people put a lot of things on policy is because it's easy to document that So and So appeared in a policy document, and usually often policy develops our personal context so as long you can provide a supporting letter saying that, yes, this research had influenced this policy, you'd kind of squared that circle or, you know, you'd kind of closed the loop. So really the criteria privileged policy impacts and as in counted against ones that were more public engagement, and I just don't think there's any doubt in that. And I know that from the other side of the coin which is as somebody who worked in developing case studies that the evidence thing was much easier to prove in policy and the ones that could not show it got ditched, even though they may have had greater significant impact but it was just difficult to provide evidence for it.

Most interviewees spoke of evidence as something taken at face-value or as a last resort where there might be doubt in terms of the veracity of impact claims, where the sheer diversity of evidence types coupled with needs of evaluative efficiency and/or pressure to get the job done meant that evidence would be less than

frequently called upon and scrutinised as a source informing adjudications of impact scores:

> There were lots of links to websites and links to, you know, little videos and whatnot. And sometimes it was hard to judge the case study without looking at them. But we were told don't look at the corroborating evidence unless there's a problem. Because you can't look at it all, so don't look at any of it unless there's a real issue and you can't agree on that case and you need more information. Because unless you look at everybody's and look at everything it's not really a fair process which I'd agree with. I don't know how that was presented. People seemed to be basing, seemed to be expecting you to look up all this stuff and of course you don't have time, all that you can look at is what they've written, and so the idea of having the corroborating evidence, I suppose it's important that it's there if there is a problem, but it's not read unless there . . . unless I thought there was, you know … well I've written here, "It will only be accessed if there's considerable doubt about the claims made". So really we ignored it unless, you know, we were really concerned about the link between the research and the impact.

Where evidence was to be considered, interviewees expressed concern related to closely they would need to scrutinise the supporting evidence:

> One of the things that we felt was, and we had debates particularly with the users about this, was how far down the chain you had to go. So, for instance, let's take a policy example, if you did some research that fed into policy and the policy was put in place and the evidence was compelling I think most of us felt that that was evidence of impact. Some people wanted to argue that you then had to show that the policy made a difference and unless the policy impacted on the experiences or outcomes of learning then we couldn't say have impact. And we felt that was unfair because we felt that more rigid criteria were being applied to educational outputs than, say, would be in other areas. Because it's to do with educational learning you couldn't claim that a policy... if you hadn't closed an achievement gap the research hadn't had some impact. So how far down the chain do you have to show that what you did made the world a better place, and that's something that we never really fully resolved but clearly it would be unreasonable to expect the researcher to also show that a policy they influenced or that their contribution to a policy they influenced actually improved outcomes for the learners.

5.6 Over-counting the significance of impact.

While some interviewees spoke of the difficulties they had encountered in handling the supporting evidence of impact submissions, others spoke of what they perceived to be the potentially distortive effect of impact scores, where the value assigned to a case study could overly influence the overall research score of a submission:

> We were looking at, I mean, some kind of GPA rank ordering and it then appeared that there might have been a discrepancy between how you expected a particular institution to do and how they actually did. Whether positively or negatively. And when we looked into this, I men what was found was that it was to do with the weighting of the three elements. And this is something that we have subsequently raised with HEFE. That you know, impact counts for 20% but because of the way it's assessed you're looking at a relatively small number of units for any institution. So the distribution of scores for impact is very clumped. You know, that people are likely to be, you know, towards the four star end or towards the two star, whatever, rather than a more normal distribution that you might get from the outputs. So if an institution did very well on its case studies, all of its 20% will go into the three or four star. But if it did particularly badly for whatever reason all of that 20% would be going into the two star or one star and that meant that the impact was appearing to have a disproportionate influence on the overall profile. As you were combining profiles that had different shapes in the distribution. And so this is what was generating anomalies. If an institution had done very well or very badly on case studies that brought them right up or right down in the overall rank ordering.

Other interviewees expressed concern regarding an over-inflation in the significance attributed to impact scores:

> A worry I had at the time which has been borne out is worth having which is that the scores for impact would be taken to be as important and as valid and as reliable and as determining as the scores for outputs, that was my fear and that's been proved to be the case.

5.7 An issue of resourcing

Interviewees were able to offer a number of suggestions pertaining to what they felt would improve the process of impact evaluation in future assessment exercises. Recommendations for what to do different in the future coalesced around factors of resourcing or rather the way resources for impact evaluation – chief among which are evaluators – might be more effectively managed and distributed.

While many felt that an over-reliance or additional recruitment even over-representation of user-assessors might either tilt the focus of the REF towards impact evaluation or even potentially bias the impact evaluation process, some felt that the more numerous involvement of user assessors would ameliorate the process. Such a recommendation was also, however, made with cognizance of the difficulties entailed in both recruiting and maintaining the involvement of user assessors were most would not be co-opted on the same basis as their academic counterparts:

> You know, one thing we've found was that it would have been useful if we'd had more impact assessors. We could've had more people involved in

> assessing each case study and each element. But that in doing the assessment we felt it was important that the user voice and the academic voice were equally balanced. And that it shouldn't be, you know, one user against half a dozen academics. And so that meant that we were kind of limited in the way in which we could organise that kind of assessment. In a way that we weren't quite so limited in terms of assessment of the environment which was exclusively done by the academics. You could have a much greater involvement of a larger number of people in that. So I think, you know, if there were more user assessors that would be a very useful thing but I think we recognised that it's more difficult to get user involvement into this. And also it's difficult to maintain that involvement over a long enough period.

Others, among the cohort of academic panel members suggested that the evaluation of impact might be boosted by an increased allocation of reviewers, both academic and user-assessor, to case studies:

> Given I said I only marked maybe about ten, I could have easily marked 40 in detail, or 35 or something like that. So my personal view is it would have been preferable for each impact case study and template to be assessed by, say, three or four academics and three or four users, and then it was marked by six and eight people together, and then you could get, you know, either agreed scores or average scores or so on, and then have a much better discussion about – you could still ask everybody to read everything, but at least you wouldn't have the odd outliers and so on there.

One further interviewee discussed a sense of disproportionality in the amount of time given to case studies versus the time dedicated to the assessment of outputs:

> I think I had over 300 or so publications to assess. So, you know, in terms of time, I spent a huge amount more time – if you've got a book out of your 300 – or not a book, you got 40 books or something to assess, you know, you could spend weeks doing a publication assessment. But the impact assessment, you know, was a day. I mean, it was hardly anything. Because, as I said, I only had to look in detail at – or was only asked to look in detail and provide scores and discuss on maybe a dozen, about ten or – I don't know, 10 or 12 or something, but somewhere between eight and ten, or maybe it's 11. So it wasn't a lot, so it was disproportionate that we didn't spend enough time. And I think if we had, it would have been more robust, with greater confidence in it.

6. Conclusion

This research has yielded fresh insight into the process of impact evaluation of research new to the UK's system of performance based research assessment, the REF. Specifically, it has elucidated the perspectives, and what amounts almost to an apologetic or self-rationalizing discourse of academic and user-assessor evaluators located within social science and art and humanities based disciplines on the

legitimacy and efficacy of the impact evaluation process. More specifically the research considers what they found to work and what is needed to work more successfully in making credible determinations of the economic and societal contribution of academic research.

The study reveals the importance of certain co-dependent behavioural conditions or characteristics deemed essential to successful impact evaluation or more specifically the confidence of evaluators (in the credibility both of the process and their value determinations/impact scores) as harnessed through collegial and collaborative enterprise. A 'safety-net' of collective ownership and responsibility is seen to be especially significant in the context of impact evaluation as both an original and controversial dimension of research assessment.

The study also reveals issues related to difficulties in the interpretation and application of evaluative criteria; issues of (im)partiality and the (non)containment of disciplinary bias and a conservationist (or protectionist) reflex among sub-panel evaluators; the use or neglect of supporting evidence; issues of resource, capacity and expertise; fears of the over-valuing of impact and its distortive affects; and concerns regarding the proscriptive nature of impact definitions not least as they relate to public engagement.

---

7. References

Armstrong, J. S. (1997) Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics*, 3:63–84.

BERA (2011) Ethical guidelines for educational research. Available from: https://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-Ethical-Guidelines-2011.pdf?noredirect=1

Bornmann, L. & Daniel, HD. Scientometrics (2005) 65: 391. https://doi.org/10.1007/s11192-005-0281-4

Bourdieu, P. (2010) *Distinction: A social critique of the judgement of taste*. London and New York: Routledge.

Callon, M. (ed.) (1998) *The Laws of the Markets*, London: Blackwell.

Chubb, J. & Watermeyer, R. (2017) Artifice or integrity in the marketization of research? Investigating the moral economy of impact statements within research funding proposals in the UK and Australia. *Studies in Higher Education*, 42(12), 2360-2372.

Cole, S. Cole, J.R. and Simon, G.A. (1981) Chance and consensus in peer review. Science, 20;214(4523):881-6

Collini, S. 2012. *What are Universities For?* London: Penguin.

Collins, H. and Evans, R. (2002) The third wave of science studies: Studies of expertise and experience. *Social Studies of Science,* 32(2), 235-96.

Derrick, G.E., Samuel, G.N. (2016) The evaluation scale: exploring decisions about societal impact in peer review panels. *Minerva*, 54(1), 75-97.

Dunleavy, P. (2012) "REF Advice Note 1. Understanding HEFCE's Definition of Impact." LSE Impact of Social Sciences Blog. http://blogs.lse.ac.uk/impactofsocialsciences/2012/10/22/ dunleavy-ref-advice-1/.

Jasanoff, S. (1990) *The fifth branch: Science advisers as policy makers*. Cambridge, MA: Harvard University Press.

Ladyman, J. (2009) Against Impact. *Oxford Magazine* 294: 4–5

Lamont, M. (2009). *How professors think. Inside the curious world of academic judgement*. Cambridge, MA: Harvard University Press.

Lamont, M., Mallard, G. and Guetzkow, J. 2006. Beyond blind faith: Overcoming the obstacles to interdisciplinary evaluation." *Research Evaluation.* 15(1), 43-55.

Latour, B. (2005) *Reassembling the social*. Oxford: Oxford University Press.

Latour, B. and Woolgar, S. (1979) *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Moriarty, P. (2011) Science as a public good in J. Holmwood (Ed.) A manifesto for the public university, pp. 56–73. London: Bloomsbury Academic.

Musselin, C. (2013) How peer review empowers the academic profession and university managers: Changes in relationships between the state, universities and the professoriate. *Research Policy* 42, 1165–1173

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), 229-241.

Shapin, S. and Schaffer, S. (2017) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton, NJ: Princeton University Press.

Watermeyer, R. (2016) Impact in the REF: Issues and obstacles. *Studies in Higher Education.* 41(2), 199-214.

Watermeyer R. (2012) From engagement to impact? Articulating the public value of academic research. *Tertiary Education and Management*. 18(2), 115-130.

Watermeyer, R., and Hedgecoe, A. (2016) Selling 'impact': peer reviewer projections of what is needed and what counts in REF impact case studies. *Journal of Education Policy,* 31(5), 651-665.

Weber, M. (1978) *Economy and society: An outline of interpretive sociology, Volumes 1 & 2.* University of California Press: Berkeley.