UNIVERSITY OF BATH

University of Bath

Evaluating 'impact' in the UK's Research Excellence Framework (REF): Liminality, looseness and new modalities of scholarly distinction.

Abstract
Little is known about the process of evaluating the economic and societal impact of research undertaken in university settings. In this paper, we explore the accounts of senior academics and user-assessors, populating disciplinary sub-panels spanning the humanities and social sciences, convened to judge and 'score' the impact claims of researchers from UK universities as a new component of research evaluation within the specific context of the UK's performance based research funding system (PBRFS), the Research Excellence Framework (REF). We perceive from their accounts the emergence of a new and liminal space in the production of scholarly 'distinction' that is unlike archetypal modalities of academic excellence. Analogously, we identify an emotional and intellectual vulnerability in the review process and the loosening of the structures reviewers traditionally call upon in making value-determinations that simultaneously facilitate their role as impact evaluators and create new modalities in scholarly distinction.

Dr Richard Watermeyer
Reader in Education
Department of Education
University of Bath
Claverton Down
Bath, BA2 7AY
r.p.watermeyer@bath.ac.uk
+44 (0)1225 383278

Dr Jennifer Chubb
Department of Education
University of York
Heslington
York, YO10 5DD
Jennifer.chubb@york.ac.uk

Introduction

This article reports on what we believe to be the first empirical study to qualitatively investigate a formal system of impact evaluation – perhaps uniquely undertaken by both academic and non-academic/user assessors – for researchers operating in UK universities in the specific contexts of social science and arts and humanities disciplines. It offers a critical sociological analysis of the transformational effects of this performance management 'innovation' on the research praxis of such disciplinary communities. It concurrently exposes how the intensification of a policy demand for scientists' accountability (as receivers of public funds) produces deviations, or what will be observed as 'new distinctions' in scholarly conduct that are normalized or otherwise perhaps made tacit by the pervasiveness and persistence of an audit culture that has come to characterize the UK's higher education sector.

Inclusion of the economic and societal 'impact' of research as a performance indicator and evaluation criterion within the UK performance based research funding system (PBRFS), the Research Excellence Framework (REF) 2014, is credited as a game-changer for researchers working in universities, influencing how they think about and conduct research (Chubb and Watermeyer 2017; Watermeyer 2012, 2016). As a twenty percent component of assessment in what was the first, and to-date only REF in 2014, researchers' depositions of impact – formalized in narrative case-studies – represented high value items and levers of both research esteem (and positionality in performance league tables) and Quality Related (QR) research monies.

The REF, successor to the Research Assessment Exercise (RAE), which had previously been administered on six occasions between 1986 and 2008, is used to highlight areas of excellence and benchmarks the quality of previously funded research. However, unlike the previous RAE, REF2014 signalled a first attempt to assess the impact of academic research beyond scholarly communities. The REF's principal architect and administrator the Higher Education Funding Council for England (HEFCE) defined impact 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" which "includes, but is not limited to, an effect on, change or benefit to a range of contexts" (REF 2012: 48). In REF 2014 'eligible' researchers – typically academics with permanent or open-ended teaching and research contracts or rather those selectively submitted by their institutions (cf. Watermeyer and Olssen 2016) – were required to produce impact case studies that included an impact summary, details of the underpinning research, references, details and sources to corroborate the impact. HEFCE demanded that impact case studies would need to be underpinned by research of 2* quality (internationally recognized but below the benchmark for QR funding) undertaken within a census period of 1st January 1993 - 31st December 2013, while the impacts would be restricted to 1st January 2008 - 31st July 2013 timeline. In addition to the case study was an expectation from HEFCE that those returning to the Unit of Assessment (UOA) would produce an impact template and summary that might reference for instance, commercialisation, policy and service improvement and the enhancement of creative output and health achievements. Crucially, in REF 2014, the assessment of research impact (and outputs) would not rely solely on (biblio)metric analysis. Instead, the efficacy of peer-review was upheld and continues to be for the purpose of its next iteration in 2021 where the contribution of metrics are seen to 'support, not supplant, expert judgement' (Wilsdon et al. 2015: viii).

With a QR fund of approximately £1.6billion – distributed by HEFCE on behalf of the UK central government – and various calculations suggesting that high quality evidences of impact in the

REF had a currency translation of seven high quality research outputs and a potential return of anything up to £350k, the prominence of impact as a 'positional good' (Hirsch 1977) was whilst controversial, difficult to dispute. In preparing impact narratives for REF 2014, many UK universities committed significant investment in providing support for impact including recruiting copy-writers and science-writers to help academics translate their claims to impact in ways that were jargon-light, easy to understand and therefore evaluation-friendly. Crucially, this was because the assessment of impact in REF 2014 would be undertaken not only by senior academic peer-reviewers but user-assessors; those working outside of academia able to make authoritative adjudications regarding the 'public' contribution of research yet potentially lacking the fluency of their academic counterparts' subject expertise. Universities in some instances also ran mock-versions of impact assessment mirroring the process to be undertaken in the REF itself (Watermeyer and Hedgecoe 2016). Simulated internal review exercises like these, which included panel members with previous experience of the RAE and those who would become members of REF 2014 panels, provided an opportunity to road-test the kinds of criteria for impact stipulated by HEFCE. Such criteria were informed by what some designated an 'impossiblist language' of impact evaluation; a reference to notions of impact 'reach'; 'significance' and 'transformative potential' – qualifiers which it was felt lacked the precision that might otherwise facilitate a more efficacious process of evaluation (Watermeyer 2016).

Despite such efforts, knowledge of impact appraisal remained at best thin and continues to be in a post-REF 2014 landscape; due mainly to a high degree of associated confidentiality. There are instead only glimpses into the world of impact evaluation (Grant 2015) and a modest number of studies that have focused on impact metrics (Mirnezami, Beaudry and Larivière 2015, Wilsdon *et al*. 2015) and perceptions of academics required to show the impact of their research (Smith and Stewart 2016; Chubb and Watermeyer 2017). The majority of these studies have also tended to privilege the sciences (De Jong, Smit and van Drooge 2016) and medical and/or health care contexts (Buxton and Hanney 1994, 1996; Hanney *et al*. 2004). Whilst there are studies into processes of peer-review, these have tended to focus on the award of research grants or research outputs and not impact (Lamont 2009). At the same time, while a number of impact manuals have emerged (Bastow *et al*. 2014; Denicolo 2014) – accompanying academics' rush to be better responsive to the impact agenda – these have tended to concentrate on methods for the production of impact and not how it is experienced or evaluated. Understandably, the task of evaluating impact in REF 2014 would be, as our account reveals, met by panelists with trepidation and a sense of venturing into the unknown; exacerbated by a dearth of experience of most kinds of impact evaluation, economic included.

The research this article reports was intended as a response to such uncertainty and an attempt to locate a more comprehensive understanding of impact evaluation in the context of the REF as an evolving and internationally imitated PBRFS (Hicks 2012) and as viewed through the social sciences and humanities; disciplines for which evidence pertaining to the evaluation of impact is limited to a handful of studies (Donovan 2009; Samuel and Derrick 2015; Oancea 2009). The research reported herein is also explicitly a *social* study of the process of impact evaluation and consequently an exploration of the various factors – macro, meso and micro – that enabled or potentially inhibited REF panelists. Such a study is, we argue, significant not only for the fact that the inclusion of impact within the REF has been seen and felt by the UK academic community as a game changing event, but also a challenge to perceptions and praxis of what counts as excellence in research (Oancea and Furlong 2007; Smith, Ward and House 2011). This challenge is based on the premise that the significance of impact is not exclusively attributable to its

returning substantial economic dividends through QR. There is perhaps instead, another layer of significance which we suggest is the contribution of impact in the REF to opening a liminal, loose and to date largely unchartered space for the production of a new form of scholarly 'distinction'.

Distinction as a conceptual gateway to impact evaluation in REF2014

We borrow from and adapt a Bourdieusian (2010) notion of 'distinction' – whilst also acknowledging a cognate theory of 'fields' (cf. Bourdieu 1993) – in problematizing impact evaluation in the REF as an innovation of new public management, that disrupts and alters the configuration of excellence as the leitmotif of 'academic capitalism' (Slaughter and Rhodes 2004). Bourdieu conceives of distinction as the demonstration or indication of possessing social capital or the features that separate classes and reproduce inequalities. Displaying distinction typically means standing out or apart from the crowd and figures as an expression or exhibition of self, synchronous to an 'aesthetic disposition'; of 'pure taste'; and (discerning) judgment. In the specific context of academic research, distinction is found through the academic's contribution to knowledge and the recognition of such by his/her peers, and as mediated by title, role and relative position in an explicitly stratified and hierarchical 'donnish dominion' (Halsey 1992). The distinction of the scholar is also expressed in the tenets of an Enlightenment idealization of academia and Mertonian norms of critical autonomy, freedom, dispassion and detachment (Merton 1979). However, in the neoliberalized Academy, distinction and an aesthetic disposition has become corrupted and antagonistic to what Bourdieu calls a 'distance from necessity'. Consequently, distinction as the symbolization of cultural or intellectual capital, 'otherness' and order among scholars is in the contemporary milieu more frequently understood in terms of a proficiency in leveraging 'positional goods' such as competitive funding or a contribution to institutional ranking within equally competitive performance league tables. Thus 'REF impact' provides a lens through which to observe this evolution of scholarly distinction.

As an instance of academic peer-review the REF conforms to traditional academic notions of distinction, or what we might think of as 'established distinction'. However, in REF2014 the inclusion of impact could be seen to represent a challenge to established forms of academic distinction. We consequently propose, 'distinction' as a conceptual gateway facilitating insight into the method and personality of impact evaluation as undertaken by REF2014 panelists. Concurrently, we draw on Michele Lamont's (2009) work on de-contextualization of the self in the process of moving from one role to another in reference to transitions between everyday academic and panelist identities. We recognize that the process of de-contextualization is ephemeral and occurs only partially. While reviewers are temporarily stepping out of their role within academia and into a new role as peer reviewer, they bring with them the knowledge, understanding and experience associated with their discipline; the reason for their appointment. Applying this expertise in a new context requires emotional labour and an investment of self into the process of peer reviewing, setting the scene for both a consolidation of and disruption to the forms of distinction common in academia. At the same time, new self-concepts emerge in the transition towards becoming a peer-reviewer that are inherently vulnerable. This vulnerability in self-concept is we argue accentuated in the context of impact evaluation; yet it is not altogether a limitation but a strength and necessary condition of the process. Instead, we argue that impact evaluation in the terms of REF2014, and prospectively in future iterations of the REF, is inherently precarious and necessarily 'loose'. We argue that such 'looseness' typifies not only the agency of those undertaking impact evaluation but the structures intended to define it. Such looseness, for example in the interpretation and application of evaluation criteria and

underpinning evidence by panelists suggests that impact evaluation in the REF mobilizes alternative or liminal spaces for the curation of scholarly distinction.

In the following discussion we use structuration theory to elucidate 'modalities' (Giddens 1984) of impact evaluation and, therefore, aspects of agency and structure that contribute to impact as both a process and outcome of distinction. We propose three inter-connecting modalities that inform our discussion of impact evaluation as a liminal space of scholarly distinction. We begin by considering the generation of self-concept among REF panelists as a catalyst for shared and/or routinized practice. We then consider how evaluation criteria both informs and is challenged by the connoisseurship of panelists, before discussing how evidence submitted by researchers as confirmation of their impact claims was, perhaps unusually, an underused and expendable resource in making adjudications of excellence (Moore *et al*. 2017).

Methodology

Whilst recent research into impact as a REF requirement has provided ethnographic observations of the process undertaken in simulations of impact evaluations (Samuel and Derrick 2015; Watermeyer and Hedgecoe 2016), the explicit confidentiality of REF2014 meant that any direct observation of the 'real' version of impact review was never possible. What was possible, however, was access to REF evaluators 'after-the-event'. This approach would have benefits and drawbacks. While panelists would benefit from time to reflect on their experience of the REF, their accounts would be susceptible to the vagaries of memory.

An original aim had been to interview Chairs and potentially Deputy-Chairs of the sub-panels, populating all Main Panels A-D and therefore those with responsibility for overseeing the review of impact across the following disciplines:

- o Main Panel A. Medical, health, biological, agricultural, veterinary and food sciences
- o Main Panel B. Physical, Mathematical, Computer Sciences and Engineering
- o Main Panel C. Social Sciences
- o Main Panel D. Arts and Humanities

However, at the time of negotiating access we became aware of a similar research project targeting the sub-panels of Main Panels A and B. We subsequently decided to focus our study in two alternative ways. Firstly, we would focus through in-depth qualitative inquiry on the experiences of panelists within two of the sub-panels belonging to Main Panel C (with membership varying between approximately n=20-30 persons) capturing the perspectives of the Chair, Deputy Chair, main body of academic panelists and user-assessors. We would dovetail into this, additional interviews with Chairs and/or Deputy-Chairs across six other disciplinary sub-panels in Main Panel D. We anticipated that such an approach would allow for both fine-grained and deep analysis of multiple perspectives occurring from within a limited number of sub-panels in addition to more landscape or headline accounts of perspectives from a broader range of disciplinary contexts. Accounts from across the sub-panels (and two main panels) were surprisingly homogenous and we were unable to discern any significant difference in the way with which panelists' made sense of their experience of assessing impact, despite variation in the representation of sub-disciplines. In total, thirty-two in-depth interviews were conducted, which lasted anywhere between forty-five and eighty minutes.

Gaining access to interview participants was not straightforward. Many of those approached had concerns regarding potential attribution and required multiple assurances of our commitment to concealing their identity in all published accounts. Consequently, all interview extracts presented within our discussion are without attribution. Additionally, in our approach to potential interview participants we provided official HEFCE guidance of what sub-panel members were entitled to comment upon. Where interviews were generally characterized as conversational, interviewees did as might be expected and made mention of certain aspects of their experience that broke with HEFCE's guidelines. These were consequently omitted from analysis. The ethical integrity of the study was further guaranteed by subjection of research plans to the scrutiny of, and subsequent approval granted by an institutional social science research ethics committee, and adherence to principles of ethical research established by the British Education Research Association (BERA: 2011).

Interviews followed a semi-structured design, which adhered, though not restrictively, to a predetermined set of questions. The interview schedule probed respondents' personal experience of evaluating impact in the REF context: 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 agency. Transcripts were thematically analyzed using an inductive approach to identify emerging themes. The following discussion is thus organized into three sub-sections that consider modalities of impact evaluation, specifically aspects of agency and structure through which an understanding of impact evaluation as an explicitly 'loose' process and pathway to a 'new' scholarly distinction emerges.

Modality 1: Self-concept, collective identity and group-style

REF panelists took part in a process that demanded they cohere to a collective identity unlike and disassociated from their everyday contexts. They also, however, needed to maintain, project and preserve a sense of self-concept that differentiated themselves from, and conveyed to other panel members their intellectual capital and the legitimacy of their role and inclusion within the group as nominated experts. Their task was especially precarious – certainly at the outset of the process – where their credibility (and distinction) as experts was at risk of being undermined by nascent, tentative and potentially miscued judgements. Consequently, establishing a culture of trust among panelists was integral to their collective problem solving, deliberative dialogue, consensus making and no less importantly, the emergence of a 'group-style' (Eliasoph and Lichterman 2003).

In part, the viability of the sub-panel and its group style was related to the recruitment of the 'right' people. It was also seen to depend upon the skill of the sub-panel Chair as moderator and mediator of the evaluation process. Indeed, the contribution (and credibility) of the sub-panel Chair was explicitly mentioned by all our respondents as fundamental to the success of discussion and the cohesion and confidence among the group in making collective decisions:

> *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.*

The importance of the sub-panel operating as a co-operative unit engaged in an explicitly collaborative endeavor was frequently elicited and elaborated upon by respondents as an aspect of the process that had, for some at least, positively confounded their expectations, and for the majority of others, made for its success:

> 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 … there'd be some specialisation of course by virtue of your expertise and sometimes you'd have to pass things between people … All of that was true but it was much, much more. So I think what I was heartened by … a tremendous collegiality, which included sharing interpretations… debates about what constituted different star ratings and discussion of relative strengths and weaknesses and the importance of different things, like the extent to which evidence could be indirect.

The collaborative dimension of assessing impact in the REF where it involved active discussion between evaluators was also perceived as both a strength of the process and something that distinguished the evaluation of *impact* from the evaluation of academic *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 [with] the impact case studies there was much more discussion. I was pleased with that … [it] gives you confidence in the outcomes I think.

While the vast majority of decision-making on impact scores occurred at the sub-panel level, this was not in isolation from the main disciplinary panel. Organizational structures such as calibration of scoring between sub and main panels were seen as instrumental to confirming accuracy and parity across sub-discipline scores. It was also seen as a way to reflect on scoring trends at the sub-panel level:

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

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, but a means to defend disciplinary claims to impact via adjustments of modesty. At a more mundane but no less important level, the 'swapping of scores' between sub-panels was seen to provide clarity for panel members in what 'should' be identifiable as excellent impact:

> The moderation with the other sub-panels enables you to play fair on a scale. You get the range, you know? Because what they [HEFCE] tell you is, they, sort of, pose questions for you… [such as] "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…

The calibration dimension of impact evaluation in REF2014 was understood by respondents as producing an iterative system of constant posting, monitoring, reflection, action and potential moderation and re-moderation of scores, designed to eradicate the risk of significant deviation and achieve consistency in sub-panel scores. The award of final scores thus was the outcome of what amounted to a prolonged process of moderation, distillation and finessing:

> *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 calibration was intended to produce increased alignment between sub-panel scoring, the hierarchical nature of main and sub panels allowed the potential for members of the former to overrule members of the latter. The manner with which sub-panel decisions were scrutinized by main panel members 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.

Notwithstanding, respondents noted how confidence, collegiality and co-ownership emerged organically within their sub panels overriding some of the potential pitfalls of a new process, and undeterred by the moderation between sub and main panels. One possible explanation for the apparent success reported here is the existing disciplinary and experience-based commonality between members. When considering that all had worked in the same field over the long-term, it is not surprising that panel members had productive discussions and while the process was new many of the debates within it were likely to be well rehearsed. However, this is not to suggest total uniformity or homogeneity among the experience of panelists. While user-assessors and academic panel members shared common experience, accounts of socialization and integration into the REF process differed. While some discussed their experiences of informal socialization and its importance to the formal business of evaluating the case studies on a collaborative basis with academic counterparts, others cited the lack of a group identity among user-assessors, which might have been remedied with a greater focus on team integration and building:

> *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 ... 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.*

Also among user-assessors came a sense of surprise at the time-intensive nature of impact evaluation. For some the task was found to be unmanageable because, unlike most academic panel members whose focus was exclusively on the REF, they continued to deal with the demands of their everyday professional roles. For these their self-concept as impact evaluators

was arguably further strained by the frequency of their de-contextualization and their continuous oscillation between their everyday roles and their contribution as user-assessors:

> *My overwhelming view of the REF, that it was a really…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.*

Despite the majority reporting good collaboration, in cases where integration was less positively reported, user-assessors appear to have been at a disadvantage. Improving communication between user assessors and additional support for those not dedicating their full time to REF could be one avenue for improvement in the process. While the organic emergence of a framework for assessment based on confidence, collegiality, co-ownership and experience appears to be successful, more could be done on the basis of these accounts to integrate and socialize user-assessors into the process; thereby stabilizing their self-concept as impact evaluators. For academic panel members the road to self-concept was arguably more straightforward where their commitment to the REF was for the most part all-encompassing and prolonged (for many an entire year) and uninterrupted by the same kind of repeated transitioning to and from their everyday academic roles.

Modality 2: 'Looseness' in the interpretation and application of evaluation criteria
The criteria for evaluating impact – formalized by HEFCE – were viewed by respondents in both positive and negative terms and as dynamic and evolving. Some even spoke of the criteria being made fit-for-purpose by panelists themselves:

> *I think you've got to recognise that the panel had quite an influence on the criteria. And an awful lot of energy went into trying to get the impact criteria. And so, for example, through examples of what that looks like, there were heated debates in the panel 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.*

In some respects a positive picture emerges of evaluation criteria shaped by experienced panelists, but in others it suggests a need for improving the current impact evaluation measurements. 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 representative appraisal:

> *. . . certain kinds of criteria or indicators went beyond the kind of standards, so you had a wider range of criteria. And you'd have a larger number of elements to add together to produce your profiles. So it's part of the thing that James Wilsdon's[1] doing with his review of metrics … although we don't want metrics that will pre-empt the assessment, we need to find some kind of indicators and measures that we could use. Even*

---

[1] cf. Wilsdon, J. et al 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*, HEFCE.

qualitative indicators would be far better than having to rely on the relatively crude criteria that we had this time.

One major aspect of difference between user assessors' and academic panelists' interpretation of the criteria focused on what the former identified and the latter missed as a failure among case study authors to assume a theoretical standpoint, which they viewed as integral to any justification of impact and evidence of impact causality:

The people who I work with as partners and stakeholders around promoting research in our sector draw on theory. And that was the kind of thing I was expecting to see, 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.

A focusing on theory by user-assessors and conversely a lack of by case study authors and academic panellists, reveals another break with established modalities of scholarly distinction, where in the convention of academic research empirical findings are bound with and explained by theory yet in the context of impact, theory is absent and seemingly abandoned.

Finally, in the criteria of what counts as impact, the REF2014 rules dictated that public engagement was not a legitimate descriptor of impact. However, many respondents were unconvinced and spoke of their ambivalence in interpreting and differentiating public engagement as a form of impact or its conduit:

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 to 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.

Diverging interpretations and applications of the impact value of public engagement among panelists reinforces a view of the looseness and liminality of impact evaluation in REF2014. Furthermore, ambivalence related to the legitimacy of public engagement as a form of impact highlights the challenge of self-conceptualization for case study authors as impactful researchers where their claims of impact are modeled on public engagement. Those claiming impact through public engagement would seem, therefore, to have been vulnerable to a lack of consensus and/or indecision among panelists and lack of fit with HEFCE's formal evaluation criteria. This does not, however, have appeared to deterred REF2014 case study authors' use of public engagement as impact. Watermeyer and Lewis (2017) for instance report that public engagement featured within 4871 impact case studies from a possible total of 6637.

Modality 3: Variability in the application of evidence

Respondents routinely spoke of the challenge of interpreting the evidence put forward by case study authors; how to determine and trust its legitimacy and how to weight and assign value to different forms:

> *The issue of evidencing in itself is difficult and that I think we were grappling with "what makes good evidencing of impacts?"… So I was able to look at an impact case study and see the impact they'd claimed, whether it was appropriate, whether it matched with 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, 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 most often to the type and quality of the evidence, rather than inexperience in processing impact claims. For some, the evidence upon which a value judgement would be made was intangible and what many felt differentiated 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, 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' evaluation process, an inability to avoid privileging certain forms of evidence; and hence 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 … 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.

Reflected here is the sense of an uneven playing field and some research being awarded a high impact score on the basis not of claims of impact but the strength of the underlying evidence. However, most interviewees spoke of evidence as a last resort and consulted only when case study authors' claims of impact gave them cause for doubt. Respondents claimed this 'light-touch' approach was due both to the sheer diversity and volume of evidence types and the time pressures they faced in getting the job done:

> *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 … because unless you look at everybody's … it's not really a fair process which I'd agree with… all that you can look at is what they've written … [about the corroborating evidence] "It will only be accessed if there's considerable doubt about the claims made". So really we ignored it…*

Where evidence was considered, respondents expressed concern in how closely it would need to be scrutinized. As such, they intimated how their experience of evaluating impact involved a method of distinction making unlike established repertoires connected with the evaluation of outputs and the interrogation of tangible evidence:

> *… we had debates particularly with the users about…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 we felt that was unfair … So how far down the chain do you have to go 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 actually improved [say] outcomes for the learners.*

Discussion

In these accounts panelists' efforts in determining what counts as impact are shown to be influenced (and potentially manipulated) by impact evaluation as 'emotional work', conducted in a liminal space that excludes the kinds of resources – theory and evidence (and at times, criteria) – and places at risk the forms of capital traditionally applied by academic researchers in the appraisal of scholarly distinction. These accounts concordantly reveal panelists' vulnerability in attempting to operationalize self-concept; evaluation criteria; and underlying evidence of impact.

An idea of self-concept, as Michele Lamont argues, is integral to the success of peer-review processes and especially in these accounts made to seem problematic, where not all panelists interviewed were united in support of an impact agenda or successfully harmonized into a group style. Any doubt in the efficacy of their undertaking may have had a significant influence (if only ever tacit) on their behaviour as panel members and what Callon (1998) designates as

'performative effects'. Yet, given the closed-shop nature of REF sub-panels and a high degree of associated confidentiality, this will never be firmly established. Nevertheless, we are able to envisage through these accounts, complementary with what Lamont in her study of funding panels observed as the social fragility of academic peer-review processes and the struggle of the academic field in such context. Correspondingly, we identify the significance of the role of the sub-panel Chair in moderating dialogue and facilitating and scaffolding self-concept and group-style; in mitigating against the formation of sub-panel oligarchies and therefore ensuring evenness and equality in dialogue; and in mediating between the sub-panel and main-panel. We also perceive the necessity for trust and the need for panelists to be socialized into the process, allowing them to develop rapport, mutual respect and a sense of being at ease with each other; particularly in the face of underlying ambiguity and ambivalence and variation in their professional contexts and levels of involvement. Yet such ambiguity or what we have termed 'looseness' in the interpretation and application of the resources that panelists drew upon, namely criteria and evidence, is a crucial aspect of impact in the REF that demarcates a shift from established to new modalities of scholarly distinction.

We would argue on the basis of these accounts, that looseness in the interpretation and application of impact criteria is prerequisite to a nascent and evolving evaluation focus. Such looseness may not necessarily facilitate rapid consensus but will provide for a richer and more composite perspective. It also, however, places an additional onus on the evaluator 'to alternate between different social worlds and translate between them' (Collins and Evans 2002: 262). Moreover, these accounts indicate the potential risk associated with respondents' gravitation towards an empathic response in the evaluation of impact case studies. Indeed, they illustrate the challenge faced by panelists in marshaling too great a drift towards empathic reasoning and intuitive deduction when accommodating the perceived weaknesses of pre-established evaluation criteria. For instance, we observe the struggle of panelists in reconciling their disagreement and transgression from HEFCE's formalization of public engagement in the terms of impact in the REF. This is also a struggle identified in a recent review of REF2014 (Stern 2016), which advocates for a broader interpretation of public engagement as impact and thus casts further open a net for new forms of scholarly distinction.

What, however, we failed to identify was any significant variation or divergence in panelists' meditations on impact evaluation that might be attributed to their disciplinary affiliation and outlook. This relative absence of dissonance may be explained, at least in part, by the limited number of (sub)disciplinary panels consulted across the two Main Panels – eight from a potential of twenty-one – although we intentionally captured within our sample, membership of sub-disciplines engaged in different types of research and producing different kinds of impacts. We would, however, confidently surmise that a conceptualization of modalities of impact (and interpretation of impact as a new form of scholarly distinction) would likely diversify with inclusion of a more encompassing and disparate disciplinary demographic. Indeed, it would surely be naïve to assume that all disciplinary perspectives would cohere around a common logic or *weltanschauung* of impact as scholarly distinction. Instead, we postulate that these accounts reflect an experience of impact that is perhaps specific or especially germane to social science and arts and humanities disciplines and which, furthermore, emerges from our own socially constructionist approach to understanding the impact phenomenon. To these disciplines, we associate a great number of researchers – recognizable within our respondent sample – inherently predisposed or rather, less averse and sympathetic in the context of their research foci, undertakings and non-positivist approach, to empathic reasoning, intuitive deduction and

social immersion (and public engagedness) that are revealed here respectively as behavioural traits and value priorities of impact evaluators. Moreover, these disciplinary 'tribes' (Becher and Trowler 2001) may be well-suited to an appreciation of impact narrative, where narrative is a conspicuous aspect of their everyday research praxis. Analogously, the transitioning (and associated liminality and looseness) of their self-concept between their everyday roles and that of being impact peer-reviewers may be more fluid, flexible and easily accommodated than those for whom interpretation and analysis of narrative is less of a common or everyday occurrence. Consequently, the theorization of impact as a new form of scholarly distinction is in such terms paradoxically informed by panelists' disciplinary practice yet concurrently divorced from the application and generation of theory and the contestation of evidence that is the universal fulcrum of all fields of research. Our supposition accordingly is that disciplines of a non-positivist orientation – those for instance aligned with phenomenological and hermeneutic methodologies – may be most resilient and best equipped to tackle the multiple vulnerabilities experienced through and engendered by impact as a new form of scholarly distinction. Arguably thus, through impact, scholars of the social sciences and the arts and humanities are themselves made further distinct.

Notwithstanding, we see in these accounts, panelists' efforts in adjudicating the impact of research in the REF being shaped by the fragility of self-concept; the inadequacy of criteria; the inconstant and inconsistent use of evidence; the absence of theory; and yet the moderation of panel Chairs; and the intervention of Main Panels. We also recognize a paradigm of performance evaluation that is driven predominantly by an emotional less empirical response to the interpretation and attribution of excellence – perhaps even an unconscious resistance to rationalizing impact through theory and evidence – and which signposts, therefore, a break with established modalities of scholarly distinction. This is a fracture, however, potentially more easily mediated by those employing non-positivist research methodologies and is, therefore, deserving of prolonged scrutiny, particularly in mitigating the threat of any kind of disciplinary bias. There is much, we might argue for the social sciences and arts and humanities to gain from impact, despite earlier concerns (Watermeyer 2012), and moreover and more worryingly, to *game*. With the intensification of a policy commitment to impact in the REF and the recent formalization by HEFCE of an increase of an impact weighting: up from 20% weighting in REF 2014 to 25% in REF 2021, opportunities for these two tribes as purveyors of a new scholarly distinction appear abundant.

Conclusion
In a political milieu characterized by the relegation of expertise and veneration of non-truth the significance of impact in the REF may be interpreted by academics as a sign of the further intensification of their struggle both for self-justification and self-determination. Coterminously, it may also signal the further deterioration of their self-concept as guided by an established version of scholarly distinction. Furthermore, reports of ambivalence from our respondents concerning the efficacy and integrality of evidence as a resource for critical deliberation suggests that impact as a new form of scholarly distinction is far more a matter of taste – however norm-referenced – than an ability to interrogate or theorize 'the facts'. In the era of academia's 'competitive accountability' (Watermeyer and Tomlinson 2018) and 'performative turn' (Lucas 2006), impact may thus be construed as the ultimate indicator of travel between an *ancien régime* and its prescriptions of excellence in scholarly life and what we propose as new and evolving modalities of scholarly distinction.

References

Bastow, S., Dunleavy, P. and Tinkler, J. (2014). *The Impact of the social sciences: how academics and their research make a difference*. London: Sage.

Becher, T. and Trowler, P. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines*. Buckingham: OUP.

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

Bourdieu, P. (1993). *The Field of Cultural Production*. Cambridge, UK: Polity Press

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

Buxton, M. and Hanney, S. (1994). *Assessing Payback from Department of Health Research and Development: Preliminary Report. Volume 1: The Main Report*. HERG Research Report, No. 19. Uxbridge: HERG, Brunel University.

Buxton, M. and Hanney, S. (1996). How can payback from health services research be assessed? *Journal of Health Service Research and Policy*, 1(1), 35–43.

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.

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.

De Jong, S.P. L., Smit, J. and van Drooge, L. (2016). Scientists' response to societal impact policies: A policy paradox. *Science and Public Policy.* 43(1), 102-114.

Denicolo, P. (ed.) (2014). *Achieving impact in research*. London: Sage.

Donovan, C. (2009). Gradgrinding the social sciences: The politics of metrics of political science. *Political Studies Review*, *7*(1), 73-83.

Eliasoph, E. and Lichterman, P. (2003). Culture in interaction. *The American Journal of Sociology*, 108(4), 735-794.

Giddens, A. (1984). *The constitution of society*. Cambridge: Polity Press.

Grant, J. (2015). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*. King's College London and Digital Science.

Grant J., Measuring the Benefits from Research, 2006 RAND Europe (Published online) <http://www.rand.org/pubs/research_briefs/2007/RAND_RB9202.pdf> accessed 16 Nov 2017

Halsey, A.H. (1992). *Decline of donnish dominion: The British academic profession in the twentieth century*. Oxford: OUP.

Hanney, S., Grant. J., Wooding, S. and Buxton, M. (2004). Proposed methods for reviewing the outcomes of research: the impact of funding by the UK's Arthritis Research Campaign. *Health Research Policy and Systems*, 2, 4.

Hicks, D. (2012) Performance-based university research funding systems. Research Policy, 41(2), 251–261.

Hirsch, F. (1977). *The Social Limits to Growth.* London: Routledge & Kegan Paul.

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

Merton, R. K. (1979). *The Sociology of Science: Theoretical and Empirical Investigations.* Chicago: University of Chicago Press.

Moore, S., Neylon, C., Paul Eve, M., Paul O'Donnell, D., and Pattinson, D. (2017). "Excellence R Us": university research and the fetishisation of excellence. *Palgrave Communications* doi:10.1057/palcomms.2016.105

Mirnezami, S.R., Beaudry, C. and Larivière, V. (2015). What determines researchers' scientific impact? A case study of Quebec researchers. *Science and Public Policy*. 43(2), 262-274.

Oancea, A. (2009). Performative accountability and the UK Research Assessment Exercise. In: ACCESS: Critical Perspectives on Communication, Cultural & Policy Studies (NZ), 27, 1 & 2.

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

Penfield T., Baker M., Scoble R & Wykes M., Assessment, evaluations, and definitions of research impact: A review, *Research Evaluation*, 23(1), 21–32.

REF (2012). Assessment Framework and Guidance on Submissions. Updated. London: HEFCE.

Samuel, G. and Derrick, G. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*, *24*(3), 229-241.

Slaughter, S. and Rhoades, G. (2004). *Academic capitalism and the new economy*. Maryland: John Hopkins University Press.

Smith, K.E., and Stewart, E. (2016). We need to talk about impact: Why social policy academics need to engage with the UK's research impact agenda. *Journal of Social Policy* DOI: https://doi.org/10.1017/S0047279416000283

Smith, S., Ward, V. and House, A. (2011). 'Impact' in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. *Research Policy*, 40(10), 1369-1379.

Stern, N. (2016). Building on success and learning from experience: An independent review of the Research Excellence Framework. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/541338/ind-16-9-ref-stern-review.pdf

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 A 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*.*

Watermeyer, R., and Olssen, M. (2016). Excellence and exclusion: The individual costs of institutional competitiveness. *Minerva*, 54(2), 201-218.

Watermeyer, R., and Tomlinson, M. (2018). The marketization of pedagogy and problem of 'competitive accountability' in E. Medland, R. Watermeyer, A. Hosein, I.Kinchin and S. Lygo-Baker eds. *Pedagogical peculiarities: Conversations at the Edge of University Teaching and Learning.* Rotterdam: Brill/Sense Publishers.

Wilsdon, J., et al. (2015). The metric tide: Report of the independent review of the role of metrics in research assessment and management. London: HEFCE.