Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014

Gabrielle N. Samuel* and Gemma E. Derrick

Health Economics Research Group (HERG), Brunel University London, Uxbridge UB8 3PH, UK
Corresponding author. Email: Gabrielle.samuel@brunel.ac.uk

The relative newness of 'impact' as a criterion for research assessment has meant that there is yet to be an empirical study examining the process of its evaluation. This article is part of a broader study which is exploring the panel-based peer and end-user review process for societal impact evaluation using the UK's national research assessment exercise, the Research Excellence Framework (REF) 2014, as a case study. In particular, this article explores the different perceptions REF2014 evaluators had regarding societal impact, preceding their evaluation of this measure as part of REF2014. Data are drawn from 62 interviews with evaluators from the health-related Panel A and its subpanels, prior to the REF2014 exercise taking place. We show how going into the REF exercise, evaluators from Panel A had different perceptions about how to characterize impact and how to define impact realization in terms of research outcomes and the research process. We conclude by discussing the implications of our findings for future impact evaluation frameworks, as well as postulating a series of hypotheses about the ways in which evaluators' different perceptions going into an impact assessment could potentially influence the evaluation of impact submissions. Using REF2014 as a case study, these hypotheses will be tested in interviews with REF2014 evaluators post-assessment.

Keywords: societal impact; Research Excellence Framework; qualitative research; research evaluation.

1. Introduction

In 2014, the UK Research Excellence Framework (REF) was the first mandatory national research assessment body, which is linked to funding allocation, to dedicate a formal proportion (20%) of its overall evaluation criteria to considerations of societal impact. Societal impact of research was assessed for REF ex post (after the event) and was defined as '... *an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia*' (Research Excellence Framework 2011: 26).

The REF2014 exercise for societal impact proceeded via panel-based peer and end-user review assessment. During such processes, a group of evaluators jointly deliberate and judge the merit of whatever is being assessed, with a final chair making the decision based on the common judgment of all reviewers (Olbrecht and Bornmann 2010). The peer review system is extensively used in academic research, and as such it has been widely studied, with numerous scholars pointing to both potential and observed risks. For example, the individual peer review process, which occurs during publishing, has been the focus of a breadth of studies, with many scholars pointing to issues of bias (Sandstorm 2009; Bornmann and Mungra 2011; Lee et al. 2013). A number of studies have also examined the process of panel-based peer review (Langfeldt 2006), with an increasing body of literature exploring this process during grant application decision-making (Bornmann 2011). In particular, some of this work has centred on

the analysis of the social conditions that lead panelists to view their judgments as fair, and to believe they are able to identify the best and less good proposals (Lamont and Huutoniemi 2011). In essence, scholars studying panel-based peer review are interested in the decision-making process—how decisions are reached and how group dynamics influence the common judgment that is made (Klein and Olbrecht 2011). Work has also focused on exploring the advantages and disadvantages of such assessment approaches (Klein and Olbrecht 2011). Advantages have been noted as: evaluators being able to motivate each other; being able to jointly re-evaluate their arguments; to weigh up the arguments against one another; and to make distinctions between important and less important arguments (Olbrecht and Bornmann 2010; van Arensbergen et al. 2014). On the other hand, group interaction has been argued to result in poorer decision-making, because shared responsibility creates a situation in which everyone withdraws and no one really endeavours (Levi 2007; van Arensbergen et al. 2014). Moreover, it has been suggested that some of the alleged effects of individual peer review, such as cronyism, self-interest, and cognitive particularism, may influence the way panels are set up (Lamont and Huutoniemi 2011).

The newness of societal impact as a criterion of assessment means that little is known about the process of peer review for such a measure. Indeed, the newness of the impact criterion suggests a need to understand more clearly how evaluative decisions are made about this measure during the peer review process. This is particularly true given the current contention about how societal impact should be defined, evidenced, demonstrated, and measured (Buxton and Hanney 2008; Mostert et al. 2010; Spaapen and van Drooge 2011; de Jong et al. 2011; Bornmann 2013; Bornmann and Marx 2014).

As the first mandatory, formal, ex post, peer and end-user review assessment of societal impact linked to funding allocation, REF2014 provides a good model for analysis of societal impact evaluation, and indeed a world first case study in which to explore societal impact in the context of the process of evaluation. In light of this, this article is part of a broader study which is exploring societal impact evaluation specifically for research submitted for assessment to the REF2014 health-related fields. This larger study draws on interviews with evaluators both prior to, and following, their assessment of impact. In the context of this larger study, this particular article draws on 62 interviews conducted with REF2014 evaluators prior to them engaging in the review process, and prior to their assessment of societal impact for the REF2014, in order to explore their perceptions about impact. At this time, interviewees were aware of the significance of their upcoming role as an impact evaluator for REF2014, and had access to the REF2014 guidelines for evaluating impact, but had not yet embarked on the process of assessment. This article therefore describes the baseline different opinions of evaluators about how they characterized impact in terms of assessment prior to the REF exercise taking place. (In a sister paper also reporting on the pre-evaluation interviews, we report evaluators' baseline opinions about how they valued societal impact with respect to scientific impact (Derrick and Samuel 2015)). These pre-evaluation interviews therefore provide information about any different characterizations about impact that would be brought into the evaluation discussions, and inevitably contribute to the development of the committee culture (Olbrecht et al. 2007), thereby potentially influencing the evaluation of impact submissions. In a following paper, we will explore how these different perceptions unfold and influence the impact assessment peer and end-user review process.

Below, we provide a brief literature review of societal impact evaluation research, as well as describe the REF2014 guidelines for health-related fields (Panel A) as they relate to societal impact. We then present our methods and discuss our findings. Finally, we consider our findings and hypothesize about their potential influence on the peer review process when evaluators consider impact as a measure during the REF2014. This will be tested post-assessment, during the second stage of interviews with the evaluators, the results of which are not discussed in this article.

2. Literature review

2.1 Societal impact evaluation

There is an increasing expectation for research institutions and funding bodies to demonstrate how the research they fund offers demonstrable benefits to society. Many UK research funding bodies have already introduced measures to assess ex ante (before the event) societal impact, most commonly recognized as the 'pathways to impact' statement that appears as a constituent on many of the UK grant application forms (Research Councils UK). The Higher Education Funding Council for England's (HEFCE's; which conducts the REF on behalf of all four UK funding councils) decision to formally incorporate ex post societal impact as a criterion of research assessment represents a further response to this expectation, and has sparked great interest in this area. Scholars have been informally developing and appropriating a variety of tools for the assessment of the societal impact of research, many of which informed the development of REF2014's impact evaluation guidelines. These tools vary in their characteristics and criterion of assessment, and, as such, contention about the definition of societal impact, its characteristics, and how this reflects different modes of evaluation still remains (Bozeman and Boardman 2009; Holbrook and Frodeman 2011; Spaapen and van Drooge 2011).

The payback framework, developed initially for the health sciences, was one of the initial research evaluation tools that incorporated both academic outputs and societal impact as a criterion for assessment. This framework, which was developed during the 1990s (Buxton and Hanney 1996), has been used to assess a number of health science funding programs, including those in the UK (Wooding et al. 2005), as well as internationally (Bernstein et al. 2006; Kwan et al. 2007; Oortwijn et al. 2008; Scott et al. 2011; Donovan et al. 2014). It uses an outcome-based retrospective, narrative, case study approach to assess a series of five outcome categories of individual paybacks from research. These include: knowledge production; research targeting, capacity building, and absorption; informing policy and development (which is interpreted very broadly and does not just refer to national policies); health benefits; and broader economic benefits (Hanney et al. 2004; Donovan and Hanney 2011). The framework encompasses a somewhat laborious and intensive process, and uses a multitude of methods to collect and triangulate data, including interviews, questionnaires, bibliographic analyses, and document analyses (Hanney et al. 2007; Kalucy et al. 2009). Scholars have also noted that it can be difficult to 'capture' impact (Martin 2011, 2012; Scott et al. 2011); that attribution concerns make it difficult to determine the exact contributions of research versus other factors, in achieving the impact (Hanney 2005; Buxton and Hanney 2008; Frank and Nason 2009; Kalucy et al. 2009; Buxton 2011; Scott et al. 2011); that there is a long time lag between research application and societal impact (Hanney 2005; Frank and Nason 2009; Kalucy et al. 2009; Buxton 2011; Scott et al. 2011), which, as Frank notes can take anywhere from 2 to 30 years from basic discovery to effective therapy (Frank and Nason 2009); and that there are a variety of different conceptualizations of impact (Bornmann 2013). However, the framework's multidimensional categorization of benefits has been emphasized (Hanney 2005), as well as its ability to help focus analysis, organize the assessment of impact on non-academic audiences, and provide consistency for presenting case studies (Hanney 2005; Klautzer et al. 2011). The narrative, case study approach to societal impact evaluation is therefore now considered best practice (Donovan and Hanney 2011) and has been adopted in a number of other similar outcome-based impact evaluation models (Bornmann 2013).

Indeed, the narrative case study approach was a key component to the evaluation of impact in Australia's national assessment exercise, the Research Quality Framework (RQF). Whilst the RQF was never implemented due to a change in government, the assessment model was fully developed. This was by no means a smooth process—the definition of impact, and how it would be evaluated, was an incredibly contested issue (Donovan 2008). Though, finally, a five-point impact

rating scale was developed geared towards end-user interaction in order to emphasize the need for activities and mechanisms likely to enhance the chance of research being utilized. Towards the lower end of the scale, impact was characterized by an engagement with end-users; and at the higher end of the scale, impact was more outcome-based and characterized by the adoption of research for society's benefit (Donovan 2008). This scale compliments the payback framework, which recognized such end-user interactions in terms of 'interfaces', though stopped short of formally assessing them as 'impact' (Hanney 2005).

Alongside the development of the above models for societal impact evaluation, Spaapen and van Drooge proposed the Social Impact Assessment Methods for research and funding instruments through the study of Productive Interactions (SIAMPI) (Spaapen and van Drooge 2011). Similar to other methods of evaluating impact (Scoble et al. 2010), the SIAMPI model recognizes the fact that scientific research is not the sole contributor to societal impact, with interactions between researchers and stakeholders being important prerequisites (Molas-Gallart and Tang 2011; Spaapen and van Drooge 2011). In contrast, however, rather than assessing impact outcomes, the central theme of SIAMPI is to explore the process of impact, by capturing 'productive interactions' between researchers and stakeholders, and assessing their value (Spaapen and van Drooge 2011). Spaapen and van Drooge specifically define productive interactions as '*exchanges between researchers and stakeholders in which knowledge is produced and valued that is both scientifically robust and socially relevant... The interaction is productive when it leads to efforts by stakeholders to somehow use or apply research results... Social impacts... are [then] behavioural changes that happen because of this knowledge*' (Molas-Gallart and Tang 2011; Spaapen and van Drooge 2011).

Spaapen and van Drooge argue that focusing on 'productive interactions' help circumnavigate the time lag and attribution issues commonly associated with outcome-based modes of impact evaluation (Molas-Gallart and Tang 2011; Spaapen and van Drooge 2011; de Jong et al. 2014). They emphasize that assessing the process of impact allows an evaluation which is '*closer to the actual process that the researcher is able to influence, that is closer to the actual practice of the researcher doing research and interacting with stakeholders*' (Spaapen and van Drooge 2011; 216), and that by using productive interactions as an indicator helps anticipate societal impact that may not yet have occurred at the time of evaluation (de Jong et al. 2014).

2.2 Research Excellence Framework 2014

Based on a report by RAND Europe (Grant et al. 2009), which recommended the RQF case study approach as the most suitable assessment tool, REF2014 evaluated societal

impact via narrative case studies. The structure of these case studies, which were four pages in length, was tightly controlled by the impact template supplied by HEFCE. Within these templates, institutions had to nominate pieces of underpinning research conducted at their institutions—for example, reports in the grey literature, or academic journal articles—and explain how this research had had an 'impact' on society. The underpinning research had to have been considered to have reached a threshold of no less than 'two stars' in research quality. That is, research that has a '*quality that is recognised internationally in terms of originality, significance and rigour*' (Research Excellence Framework 2012a: 46). This underpinning research must have been produced during the time frame 1 January 1993–31 December 2013; however, the described impact must have occurred between 1 January 2008 and 31 July 2013.

The case studies were assessed by various panels. There were four overarching Main Panels (Main Panel A–D) which were loosely divided into fields of research, and further divided into a number of subpanels responsible for evaluating submissions from more specific fields of research, or units of assessment. The responsibility of each Main Panel was to provide overarching advice and guidance to subpanels. Main Panels included, among others, the Chair and Deputy Chair of each of the subpanels and a number of international and UK-based evaluators, as well as evaluators who were assigned by HEFCE as academic experts (AE), or research users (stakeholder) (UE) evaluators. The UEs were evaluators predominantly from outside the academic sector and who represented the private, public, or charitable sectors that either use university-generated research, or commission or collaborate with university-based researchers. However, evaluators could be AEs, and also have a significant level of experience working outside of academia.

Recognizing the newness of the criteria and its assessment, it was key that the REF2014 guidelines defined impact broadly. For health-related Panel A, impact could be '*achieved from within a wide variety of research contexts and resulting from a wide diversity of approaches*', and there was '*no pre-formed view of the ideal context or approach*' towards impact (Research Excellence Framework 2012b: 33). Moreover, different types of 'impact' were recognized, which could be viewed legitimately, and with equal weighting for the REF2014 assessment of the case studies. These included contributions to: health and welfare; society, culture, and creativity; economy and/or commerce; public policy and services; production; environment and practitioners and services; and international development (Research Excellence Framework 2012b). Indeed, the REF2014 guidelines acknowledged that '*impacts can be manifested in a wide variety of ways including, but not limited to: the many types of beneficiary (individuals, organisations, communities, regions and other entities); impacts on products, processes, behaviours, policies, practices; and*

avoidance of harm or the waste of resources' (p. 27). These spanned, for example, those impacts related to interactions with stakeholders and the public, such as evidence of influence on health policy and/or advisory committees, and increased public awareness, understanding, and engagement; through to changes in policy, document changes to working guidelines, or the creation of spin out companies; and also those more 'final' outcomes, such as improvements in well-being, patient outcomes, employment figures, and the development of new products. Furthermore, the impact template stipulated the importance of research institutions' approaches '*to interacting with nonacademic users, beneficiaries or audiences and to achieving impacts from its research*' (p. 33), though such interactions were not assessed formally as 'impact', and aspects of productive interactions, such as public engagement, were explicitly stated not to be used as evidence of impact in the REF evaluation guidelines. Rather, the REF2014 guidelines stipulated that impact should be assessed against two outcome-based criteria: significance and reach. Significance was defined as the '*intensity of the influence or effect*'; whereas reach was described as '*the spread or breadth of influence or effect on relevant constituencies*' (Research Excellence Framework 2012b). The assessment of impact was to be awarded either one of five star profiles, where the lowest rating (0—Unclassified) was where '*…the impact has little to no reach or significance, or was ineligible, or not underpinned by excellent research produced by the significant unit*', and the highest (four stars) was where the impact '*… is outstanding in terms of its reach or significance*' (Research Excellence Framework 2011: 44).

The broadly conceived notion of societal impact assessment developed by the REF orchestrators aimed to allow the exercise to be refined over time, a factor openly acknowledged by HEFCE. Even so, the move to incorporate a societal impact assessment into the REF2014 has been criticized on the basis that there are still too many issues related to the nature of 'societal impact' and how it will be evaluated, which still need ironing out. Scholars have argued that societal impact is often indirect, partial, opaque, and long-term leading to issues of attribution and time lag (Martin 2011; Penfield et al. 2014); and that there are still major conceptual problems regarding how impact should be defined and assessed (Frank and Nason 2009; Brewer 2011; Martin 2011; Bornmann 2013). Moreover, as Brewer noted, impact '*varies over time and can change, positively or negatively, at the one-point snapshot whenever it is measured*' (Brewer 2011:256).

3. Methods

3.1 Recruitment

HEFCE was informed and supportive of the research project as long as it did not interfere or breach the confidentiality agreement of the REF2014 evaluators and the

Table 1. The number of interviews conducted with REF2014 Main Panel A and its six subpanels

Subpanel name	No. of subpanel members	No. of AEs (% of AE's) on subpanel	No. of UEs (% of UEs) on subpanel	No. of participants (% of participants) from subpanel*
Main panel	19	14 (73.7)	5 (26.3)	8 (42.1)
Subpanel 1—Clinical Medicine	39	32 (82.0)	7 (18.0)	10 (25.6)
Subpanel 2—Public Health, Health services, and Primary care	27	23 (85.1)	4 (14.9)	13 (48.1)
Subpanel 3—Allied Health Professions, Dentistry, Nursing, and Pharmacy	51	42 (82.3)	9 (17.7)	14 (27.5)
Subpanel 4—Psychology, Psychiatry, and Neuroscience	35	28 (80.0)	7 (20.0)	9 (25.7)
Subpanel 5—Biological Sciences	35	30 (85.7)	5 (14.3)	6 (17.1)
Subpanel 6—Agriculture, Veterinary, and Food Science	29	16 (55.1)	13 (4.9)	4 (13.8)
Total	235	185 (78.7)	50 (23.2)	62 (28.8)

Note: Note that two of the participants sat on two different subpanels.

evaluation process. Interview questions were provided to HEFCE for review prior to the interviews taking place. This coordination meant that all interviewees felt comfortable and adequately informed about the aims and objectives of the research project.

Interview participants were sourced purposefully from Main Panel A, which covers six subpanels: (1) Clinical Medicine; (2) Public Health, Health Services, and Primary Care; (3) Allied Health Professions, Dentistry, Nursing, and Pharmacy; (4) Psychology, Psychiatry, and Neuroscience; (5) Biological Sciences; and (6) Agriculture, Veterinary, and Food Sciences. The number of evaluators assigned to each subpanel under Main Panel A ranged from 51 (Allied Health Professions, Dentistry, Nursing, and Pharmacy) to 27 (Public Health, Health Services, and Primary Care). The Main Panel A included 19 evaluators. A number of evaluators (n = 20) were also represented on more than one subpanel. Each subpanel was composed of both AE and UE (see Table 1).

Within these panels, evaluators were responsible for evaluating outputs only, impact only, or both outputs and impact. This research did not take into account whether evaluators were also responsible for the 'Environment' criterion of the REF2014. This information, therefore, was not collected.

A total of 215 evaluators were identified and invited to participate in the projects. Invitations were originally sent via email, resulting in a total of 62 evaluators agreeing to participate in the interviews (28.8% response rate; see Table 1). All interviewees were provided with a participant information sheet and informed and/or written consent was obtained prior to commencement of the interviews. Ethics approval was granted on 22 November 2013 from the Brunel University Research Ethics Committee (2014/4), prior to the interviews taking place.

3.2 Interviews

Interviews were conducted via the telephone, skype, or face-to-face, and were recorded, and transcribed for analysis. Interviews lasted between 1 and 2 h, were semi-structured, and were conducted by Gemma Derrick during January to March 2014. In line with the study's aims and objectives, all interviews were completed before the REF2014 evaluation process started.

To ensure the interviewees' views about the definition and characterization of impact were not influenced by the interview discussion, the interview was opened with a broad question regarding the participant's definition of impact ('In your own words, please tell me how you would define research impact?'). Following on, the interview schedule incorporated a number of themes each comprising of one, main, overarching question, followed by a series of 'prompts' for further investigation. Importantly, the semi-structured nature of the schedule allowed the interview to flow as a natural discussion, rather than the interviewer introducing new concepts, which could inadvertently prompt a response. In this way, the interview was interviewee-led with participants driving the discussion, and cues about the ordering and structure of the interview were taken from the interviewee. The prompts were thus used to keep the interviewee on topic, while also serving as a method to explore emerging themes in more depth, and maximizing the strength of the qualitative approach adopted in this study. Interview themes were based around common issues currently discussed in the academic literature about the evaluation of research impact and peer review (previously described). Interview questions also drew on the participants' previous research and peer-review research evaluation experience, and the influence of research impact in these situations. Participants' past experience with impact was also

used as a prompt to explore their opinions about the importance of evaluating research impact, and its inclusion as a formal criterion in the REF2014.

In the interests of confidentiality, all participant information was coded and entered into NVivo (qualitative analysis software package) for analysis. The codes used in the results below relate to the participant's panel (Main panel = P0; subpanel 1 = P1 and so forth) and their evaluation responsibilities (outputs and impact (OutImp); impact only (Imp); or output only (Out)).

3.3 Analysis

Analysis of interview data used an inductive approach to grounded theory. Such approaches use an exploratory style methodology, allowing concepts and ideas to emerge from the data (Glaser and Strauss 1967; Charmaz 2006). This method also empirically grounds theorizing to data so that abstract conceptualizations can be developed from a close analysis of the data.

As such, the analysis of data was based on two interlinked rounds: overview analysis and detailed analysis (Strauss, 1987). Overview analysis consisted of memo-making and broad coding. Extensive memo-making was employed by the interviewer directly after each interview. This allowed for the interviewer to reflect and note the emergence of different themes for analysis, as well as to draw parallels between interviewees as the interviews progressed. Broad coding by both the first and second author was conducted by scanning the interview transcripts for relevant ideas and themes. Discussion between the two authors found no major disagreements in the emerging themes. Codes were compared with these themes from the memo-making, and three over-arching themes were then developed. These were: value (the value, or types of values, evaluators place on research impact); process (how evaluators view the research and impact process); and evaluation (evaluators' views related to how impact will be assessed). Themes were then used to inform detailed coding of the full transcripts during a further, second round of coding.

Detailed, line-by-line analysis of the interview transcripts was employed using the NVivo software. As outlined in grounded theory (Glaser and Strauss 1967; Charmaz 2006), this coding was carried out using the constant comparison method. This requires the comparison of codes to be constant, rigorous, and allow for developing and refining of conceptual categories and their properties. In addition, duplicate coding by both the first and second author was cross-checked to ensure reliability of data.

4. Results

Below, we describe the different perceptions of societal impact expressed during our 62 interviews with societal impact evaluators. In particular, we describe how the majority of evaluators perceived impact as an 'outcome' of research, though participants' views varied with regards to how much value they placed on different 'research outcomes' when it came to impact assessment. Possible implications of these perceptions in terms of assessment are discussed, whilst at the same time acknowledging that at the time of interviewing the evaluators had not yet commenced the process of impact assessment, and that their perceptions may change during this process. We also describe how, rather than viewing impact as an outcome, a small minority of interviewees viewed impact as a process and placed value on research activities which promoted research outcomes rather than the outcomes themselves. We note the similarities between these views and the SIAMPI model of assessment.

4.1 Impact as an outcome

The majority of interviewees defined societal impact as an 'outcome' (n = 58): '*[impact] is outcome focused*' (P4 Imp2). Outcome was mostly defined as a 'change' or a 'difference'. For example, it could be a change to health, such as to clinical practice ('*impact is demonstrating... that the work that we funded actually is leading to, or has already changed clinical practice*' (P1 OutImp4)), to public health or the health service ('*what has changed public health*' (P1 OutImp2)), or to patient benefit ('*mak[ing] a difference to patient care or outcomes for patients*' (P3 OutImp4)). More broadly, others defined impact as '*something that changes people's lives*' (P5 OutImp1) or something which has '*made a difference to the world*' (P5 OutImp5). Impact outcome was also described in more economic terms as a 'creation'—'*creating jobs, creating economic benefit to the country*' (P1 OutImp6).

Whereas many participants consistently identified impact as an outcome, beyond this, interviewees had different ideas about how they characterized this outcome, and how these different outcomes could be weighted against each other for the purpose of assessment. In particular, different opinions emerged regarding the point at which interviewees perceived the research process to end, and for 'impact' to commence, and therefore which outcomes could be perceived as 'impact': '*is there a point which it ceases being a research outcome and being impact, and where is that point?*' (P1 OutImp5). P1 OutImp5 illustrated the complexities of evaluating impact using the exemplar of vaccine development. This interviewee argued that each stage of vaccine development could be viewed as a 'different stage of impact'. Here, impact could be initially realized when the infectious agent is determined to be the cause of the disease. However, when developing a vaccine, other impacts can all be additionally considered as separate impacts, such as the first demonstration to be effective in clinical trials, and the 'roll out' of the vaccine into the health system.

There are different stages of impact ... there [are] various stages along the way where you start off with, we think the virus is the cause of [the disease], so if we can do something about the virus that has impact because it's demonstrating causality, which is something that was elusive. Then the next stage is can we make the vaccine to this? [If] vaccine development [i]s successful then the development of a successful vaccine is the next stage of that impact. Then there's the whole efficacy argument, right? Okay, we've got a vaccine. How good is it?...So, very rapidly the impact move[s] on from that point, from individual clinical trials to actually being rolled out into ... programs, and having broad public health benefits

The issue then arises as to which stage or stages should be evidenced as impact, and to what degree these should be scored against each other. Further confounding this issue was an alternative view of impact described by P1 OutImp5—one this interviewee commented would only be held if '*you were hardnosed about it*'. This was characterized solely by the realization of a health outcome: '*I guess if you were hardnosed about it, you would say "well, actually it's not impact until it's had an impact on the disease in people". You may have rolled out into the population, but until it actually starts to reduce the incidence, it hasn't had an impact*'.

This ambiguity about how and when to define impact was reflected in many of the views by evaluators. Many evaluators were, as described above, 'hardnosed', perceiving impact to only be achieved in the presence of a health, economic, or similar 'final' outcome. For others, final outcomes were graded most highly in terms of their impact, but these evaluators also recognized the role of other 'secondary' impacts, such as the 'inclusion in guidelines', 'patenting', and 'clinical trials'; through to 'drug development' or 'policy development'. To a lesser extent, all outcomes along the research process were considered alongside each other and given equal merit—these were the 'different stages of impact' discussed by P1 OutImp5 above.

The ambiguity of impact definitions was, in itself, a common theme expressed by all evaluators with very little consensus found about impact across all interviewees, including academic and user evaluators, including academics at different seniority levels, and including those evaluators who considered their own research to have had an impact.

4.1.1 Health or economic outcome as the measure of impact.
Many evaluators (n = 19) perceived research as having impact only after there had been a marked health, economic, or other similarly 'final' outcome. These evaluators recognized that there may be a number of stages involved in the research process prior to achieving this final outcome, but in terms of assessment, these stages did not count as impact—rather, they were seen as the means for 'creating' impact: 'I think, they're all part of creating impacts. They themselves are not impact'

(P0 OutImp3). P2 OutImp3 illustrated this using the example of research that had led to the development of an assessment tool both inside and outside the UK. For this interviewee, while this research had been 'picked up' and applied, the research was deemed too early a stage to 'press all the buttons' to be classified and assessed as impact. Instead it could only hold a 'promise' of health impact: It's early days for exactly how successful they will be in improving ultimate health outcome... and hence it's slightly early for the kind of pressing all of the buttons for an impact case study. But... they have been picked up... to different extents... they hold great promise. Likewise, P0P1 OutImp1 talks about the 'potential' of impact: 'so I can't claim that I have impact achievement at this point in time, but it certainly has potential impact'.

In terms of commercial impacts, other evaluators perceived filing patents as a far cry from impact. In fact, interviewee P1 OutImp4 highlighted how this exercise made no *significant* contribution to translating research into a 'final' outcome, and as such '*was not rated highly*': '*I would not rate highly in terms of impact somebody who told me that as a result of the work they had done, there were now three patents sitting somewhere. That hasn't translated in any way*'. Likewise, for P3 Imp2, filing a patent or creating a spin-off company was not itself an 'impact', rather it is 'what is done with them' that creates value:

It's not the filing of the patent which is important. It's what you do. Creating a spinoff company in and of itself, not important, not really. What's important is that [the] company does something. There's an application or a creation of value using that spinoff company as a means for doing that

A similar sentiment was observed for the inclusion of research into guidelines or policy documents, which were considered as not always bringing an additional 'outcome' in terms of health, economics, or other 'final' outcomes ('*a lot of policy doesn't have any impact at all, as far as I can tell*' (P3 Imp2). Indeed, these practices were viewed as an 'intermediate measure' that ultimately was not very 'inspiring' and did not have 'a terribly profound effect' (P2 OutImp8). Being 'mentioned in a parliamentary report' or 'sitting on a select committee' was likewise viewed with scepticism. These activities were characterized as being '*ephemeral*', by not having a long-lasting effect on society or on any change in terms of health or economics. In fact, these activities were defined as more to do with the esteem of the researchers themselves, rather than the actual outcome of the research:

I've got a little bit of an inherent scepticism about the impact case study which says 'we were mentioned by this parliamentary report, there was a select committee or an all party committee which ran a meeting about this', because on the one hand those were terribly flattering ... but on the other hand I think they tend to be pretty ephemeral (P2 OutImp8)

This distinction between the research and the researcher is interesting and is returned to in the final section of the results during the discussion about impact as part of the research process.

4.1.2 Impact as incremental stages. For some evaluators (n = 39), impact amounted to more than just the final outcome, it was characterized by '*identifying the thresholds in the movement towards that final product, and who contributes to those thresholds*' (P0 OutImp6). For these evaluators there was a gradation of impacts: '*I guess there is no single ultimate impact, there's a gradation*' (P1 OutImp4). However, in terms of assessing the impact, evaluators differed with regards to the amount of weight to place on these incremental stages.

For a small number of evaluators (n = 6), outcomes such as 'policy change' or 'halting drug development' were viewed as impact to be assessed on par with 'final' outcomes. For example, P2 OutImp9 talked about how contributing research to guidelines is a form of impact: '*I work on a guideline steering group, I know that my research gets quoted. I know that it gets picked up. The study that I've worked on is regarded as one of the important studies... so I believe that that has an impact*' (P2 OutImp9). Referring to the development of a drug from basic research, P0 OutImp6 re-iterates the importance of recognizing early outcomes as impact. For this interviewee, discovering during the first stages of the research process that a particular protein is implicated in a disease is impact just as much as the final development of a drug. Early outcomes are part of realizing the development of a drug and therefore must be valued equally. Referring to the star scale of the impact evaluation system used in the REF2014 assessment that requires evaluators to rate the 'impact' from zero to four stars, this interviewee emphasized that as long as research had 'moved to another stage' it could be considered four-star impact:

> If ... you discover a new protein ... you can't expect ... to have impact in terms of an entirely new drug or entirely new diagnostic test if that protein is changed in a disease, because that takes years and many people to do that. What I would expect ... is that protein, to move on to another stage ... to turn it into something that might be practical, and so to me that's a four-star impact just as much as producing a drug at the end of the day..[..]..and we're going to have to make sure that they're valued just as much as the final product (P0 OutImp6)

For other evaluators (n = 33), outcomes achieved earlier in the research process were also considered impact, and therefore warranted inclusion in the REF2014 impact assessment. However, their inclusion was characterized by being '*second-level*' impacts or '*intermediates*' and therefore not as worthy as the 'final' outcome:

> [The] number of times that appears in NHS policy documents or department of health policy documents ... that's impact of a

kind, but I think that's a sort of second-level impact ... frontline is patient impact; the second line behind that would be having an impact on policy and practice documents (P3 OutImp1)

Similar to the idea of 'promise' drawn out by evaluators in the above section, these evaluators acknowledged that these intermediate practices could only potentially lead to 'final' outcomes, and assessment needed to reflect this. As P4 Imp1 stressed, intermediates only '*might*' lead to such final outcomes: '*things like NICE guidance, textbook references being made in key media, that might lead to a health outcome*' (P4 Imp1).

Consequently, for these evaluators, the final outcome, whether health, economic, or other, was still regarded as '*the most important thing*'—'*the biggie*'. Again, referring to the star scale of impact evaluation to be used in the REF2014, P1 Imp1 highlighted the different types of 'impact' with reference to comparing the most important four-star outcome of 'avoiding deaths' to two or three star outcomes, such as 'changing practice':

> The most important thing is how many deaths from cancer have we avoided? ... And then you can go down to saying, 'well, have we changed practice, have methods been developed?' Or various sorts of things that you come along the way we might call it one and two and three stars. But really ... the biggie we're looking for is reduction in mortality or increasing quality of life (P1 Imp1)

For the evaluators described above—evaluators who gave weight to impact occurring at various stages of the research process prior to the 'final' outcome—an important implication emerges in terms of societal impact evaluation: what might have earlier in the research process appeared to 'have an impact', may subsequently lead to a 'negative' or 'wrong' effect: '*there is lots of research that has an impact in the short-term and turns out to be wrong or misguided once the time has passed*' (P0 OutImp2). For example, in cases when a policy passed by a government turns out to be detrimental: '*because you could have an impact in terms of changing government policy but... that change in government policy actually has a detrimental effect because you got it wrong*' (P6 OutImp2). One interviewee drew on the example of the development of drugs, to exemplify this point. This participant stressed how, at first, despite being twice as expensive, a particular generation of drug appeared more beneficial than others, and was therefore incorporated into practice. A few years '*down the line*', this generation of drug was found to be without the originally claimed benefits and would have consequently cost the health service twice as much, with no parallel patient benefit:

> If it [a new generation drug] is discovered to [have] a significant outcome versus the cost to work, and then ... if you measure the impact at the end of the first round of publication, you say 'it's great; these things are going to be twice as good'. And if you measure them at the end of the second round of publication, you'd say, the impact is terrible; it's cost us over twice as much money for the last 10 years (P3 OutImp3)

This is unproblematic if impact is perceived as value neutral or, in other words, if no judgment is made during assessment about whether a specific impact harms or benefits society. However, many evaluators did not perceive impact as such. For them, impact was about '*making a positive change*' or '*improving*' society: '*for me… impact is when you do research, someone takes it up in the real world and makes a positive change*' (P2 OutImp2); '*impact for me is something is going to change, something is going to improve*' (P4 Imp2). Measuring impact therefore required the input of a value judgment: '*there's a value judgment involved in assessing positive impact, isn't there? It's just to assess the impact generally objectively seek some criteria and measure them. But whether that resulted in better health outcomes or better health behaviours or something, that's another issue*' (P4 Imp1). Value judgments such as these raise questions about defining 'positive' impact and making assessments about whether such undefined 'positive' outcomes should be valued over and above other outcomes as 'impact'.

Different implications emerged for those evaluators who viewed 'final' outcomes as the 'biggie' in terms of impact. First, in terms of basic and applied research, researchers earlier in the process would not be credited for delivering any final outcome if in consecutive evaluations an outcome from a specific piece of research can only be claimed and attributed to the applicants initially involved in the development of the research since, but not prior to, the previous assessment. In P2 Imp2's example of the breast-screening programme described above, if the health intervention was claimed initially as an impact outcome, only those researchers conducting research after the intervention could realistically claim any 'final' health outcome. Researchers therefore working at different stages of the research process will be scored differentially and potentially lower, for this impact. Second, an issue of time lag exists. This issue has been extensively discussed in the literature (Spaapen and van Drooge 2011; Penfield et al. 2014) and relates to the length of time it often takes for health outcomes to be realized. For example, P2 Imp2 noted that while there had been an extension of a breast-screening programme, and this is no doubt being '*felt*' because the programme had changed significantly; it will be '*ages*' until data emerge to tell whether this particular change in the programme has achieved the desired health outcome, and therefore impact:

> *Sometimes it takes ages to demonstrate the outcome. A good example might be the extension of the breast-screening program down to 57 and up to 73. The recent work that's being done will demonstrate what impact that has, what outcome that has on women, but not for ages; but it's absolutely certain that even now there's impact being felt because the program has changed significantly (P2 Imp2)*

In these instances, by the time the health outcome of a particular piece of research had been realized, it would be too late to claim impact in terms of the 'final' outcome. Evaluators viewing impact solely in terms of 'final' outcomes may therefore exclude recognition of impact on the grounds that '*it's the old, "it's too early to say" thing*' (P4 Imp1); or alternatively it could promote a downgrading of those case studies by other evaluators who viewed the final outcome as the 'biggie'. There was, however, scope in the existing REF2014 impact evaluation guidelines for exceptional cases of research to be extended further back than 20 years (the current threshold for the underpinning research), and also the possibility that subsequent impacts could be submitted for future evaluations. However, this raises questions of whether the submission of multiple, smaller impacts rather than the more long-term consideration of the entire 'impact journey', could significantly disadvantage applicants in terms of the overall impact assessment score obtained.

Overall, the views expressed above favour a characterization of impact as an outcome over a process involving a number of individual impact events. This aligns with the context of the REF2014 guidelines for impact evaluation, and is also similar to the way impact is represented in the payback framework for research evaluation. We cannot draw too many conclusions about the implications of holding different views about the characterizations of impact discussed above because these views reflect evaluators' beliefs prior to, and not during, their assessment of societal impact, their implications can only relate to these pre-evaluation views, and not to the review exercise of societal impact itself, during which opinions are likely to change. They are, however, worthy of a mention because, as part of a larger study which is exploring the review process for societal impact evaluation using REF2014 as a case study, these implications can be revisited in greater detail later, and compared with the findings from the interviews with REF2014 evaluators following the review process. In this way we can explore how any such implications may have been addressed during the review exercise.

4.2 Valuing the process of impact

A number of interviewees (n = 18) recognized that impact was contingent on social processes. They perceived that the possibility of impact being realized was more related to a range of social factors, than adequately reflecting the nature of the research, or the efforts of the researchers themselves. This concept is built from the observation that the societal impact of science is not value-free and neutral and that science does not have an impact based solely on its particular capabilities. Rather, scientific research, its shaping and development, and its application to society, is related to multilayered social factors (Brown and Webster 2004). These social factors, which are unrelated to the scientific nature of the research or

researchers themselves, can represent a facilitator or a barrier to impact being achieved.

Indeed, some interviewees were aware that outcomes of research were dependent on a whole range of 'uncontrollable', 'outside factors' (or '*forces*') that needed to be overcome: '*there are forces out there that try to inhibit development as well as encourage it*' (P0 OutImp6) (n = 40). Some '*forces*' were viewed as a barrier to impact, whilst others were seen as 'aiding' it. These forces, which have been described in relation to evidence-informed health policymaking ([Haynes et al. 2011](#)) include facilitating factors such as serendipity, along with more impeding factors, such as whether the research is 'fashionable': '*often its not to do with the quality of the research. It's a whole lot of other things about workplace cultures and what is the kind of fashionable thing of the time*' (P3 OutImp5); ready to be accepted: '*it take s a long time for things like that to be accepted*' (P1 OutImp2)); desired: '*you could do excellent research that wasn't impacted because stakeholder groups didn't want to take it up*' (P4 Imp1); or financially viable: '*say a pharmaceutical company does buy a product . . . and then they decide not to develop it. That won't be because of scientific reasons . . . but maybe just for financial reasons*' (P0 OutImp6).

A number of evaluators also acknowledged that in order to overcome many of the 'barriers', and 'push research towards impact', researchers needed to engage with stakeholders—in fact generating an outcome was viewed by some evaluators as dependent on overcoming barriers to impact via engaging with stakeholders (n = 36). These engagements were discussed in terms of 'building relationships'. Rather than the research itself, these relationships—be them with industry, policy makers, and/or patient groups—were perceived to be a necessary link between research and impact:

> Unfortunately in the real world, things often turn on whether or not you've got someone's ear in high places—whether that would be the trust chief executive or the public health agency chief executive or the permanent secretary—a lot of impact still is built around personal relationships (P2 OutImp4)

Whilst many evaluators recognized the importance of social processes in determining outcomes, and the necessity to 'push' research forward, this was only sometimes taken into account when considering how to value impact realization. P4 OutImp6 was one such interviewee who considered this. This participant highlighted the importance of recognizing a policy change as '*impact*', even in instances when the change has '*not been taken notice of*'. For this evaluator, being noticed is a separate social matter, and therefore the outcome still deserves recognition as an impact. With a similar sentiment, P6 OutImp2 was concerned that if a case study presented a specific drug which never made it to market, other evaluators would think they were having the 'wool pulled over their eyes' in terms of assessing impact. However, this

may 'not be your fault', according to this interviewee, because it might have more to do with commercial decisions, than scientific issues:

> Let's say . . . you are now making a new drug . . . [the] company's bought it, but the company decides to keep it on the shelf because they don't want to market it right now. How do you measure that? They [other evaluators] will just say you are just pulling the wool over my eyes; it didn't get to the market . . . [but] it's not your fault—there is nothing wrong with the discovery, and there is nothing wrong with the pathway you've taken. It's just that a commercial decision has been taken by the company (P6 OutImp2)

For this interviewee, then, such external factors need to be considered when evaluating impact.

For a small proportion of interviewees (n = 4), rather than any research outcome, it was the interactions researchers formed with stakeholders during the impact journey to promote their research, which were weighted heavily when evaluating impact. Impact was therefore best shown by the efforts of researchers trying to 'get their research out there', and it was these interactions which these interviewees believed should be assessed as impact: '*the question we should be asking is whether enough effort has gone into that in the past and levering research into its next stage*' (P0 OutImp6). This interviewee continued: '*the most important part of the impact assessment is . . . the journey really, from the discovery to that next end stage and . . . I'm trying to add value to that journey and look for how the university or researcher use that opportunity to get to that next stage*' (P0 OutImp6).

For these evaluators, impact was perceived not so much as a 'noun', but rather as a 'verb': '*if you think of impact as a verb rather than a noun, I think it's a lot easier to analyse*' (P0 OutImp4). Impact was something that a researcher could do to push research forward, rather than the end-point or sole outcome of a research process:

> Impact is the relationships you build. It is the dialog that you have that makes you ask research questions that are subtly different from the ones you would have asked if you hadn't linked with whether it's policymakers, whether it's citizens, whether it's industry at the beginning. So impact is not something that you have right at the end. Impact is a relationship and that attitude of mind that you have throughout the research process (P0 OutImp4)

By incorporating stakeholder interactions as a criterion of impact assessment, these evaluators felt that the researchers would be rewarded, as opposed to the research. Activities such as, sitting on a committee, which was characterized earlier as '*terribly flattering*' but '*ephemeral*' (P2 OutImp8), were therefore valued by such evaluators. As mentioned above, no distinct differences were found in the views expressed about impact expressed by the academic evaluators and the user evaluators. Rather than being based on the REF model of impact assessment, such views about impact resonate with the SIAMPI model

of evaluation, which argues the importance of evaluating productive interactions between academics and stakeholders as a criterion of assessment for societal impact (Spaapen and van Drooge 2011).

5. Discussion and conclusion

This article reports on a series of 62 interviews with REF2014 evaluators of impact prior to undertaking its formal assessment. The findings highlight the range of views evaluators had about how impact should be characterized during assessment. We have shown how the majority of evaluators viewed impact as an 'outcome', and that many valued the 'final biggies' as the most important aspect of impact. We have discussed some of the potential implications of these opinions.

At the time of interview, participants were aware of the importance of their forthcoming role as an impact evaluator for REF2014, and had access to the REF2014 guidelines for evaluating impact, but had not yet embarked on the process of assessment. These findings are therefore important as they describe the baseline different opinions of evaluators about how they characterized impact in terms of assessment prior to REF2014 taking place. This is useful as it provides policymakers embarking on other future impact evaluation frameworks, both nationally and internationally, with a rich description of the different types of conceptualizations impact evaluators may bring to formal assessment frameworks before embarking on assessment, and therefore has implications for the peer and end-user review process by which impact is assessed, as well as how future guidelines for impact evaluation are formulated. In particular, it reminds policymakers of the importance of starting debate early about the definition of impact, at the initial stages of evaluation framework development, in order to tease out evaluators' specific issues and concerns. Such concerns have the potential to influence how evaluators implement evaluation guidelines (Benda and Engels 2011), as well as the culture of the evaluation committee which has been shown to impact on peer review outcomes (Kerr et al. 1996).

Further, whilst we have noted that we cannot draw conclusions at this stage about how evaluators' perceptions will effect the impact assessment exercise due to the likely shift in opinions during the process, as baseline opinions, they do serve to highlight a number of hypotheses to now be tested about the process of impact evaluation.

One key question, relates to whether evaluators change opinions about impact characterization during the process, and what the reasons are for any such change. This will shed light on how the process has influenced evaluators' views and beliefs about impact and how the potential benefits or drawbacks of this can be applied to any future evaluations of impact. This will be explored in the next stage of research that will include post-evaluation interviews. This will be done by combining the results obtained from the pre-evaluation interviews described above, with results obtained by the second, post-evaluation interviews. The unfolding of this question will also have an influence on the further hypotheses to be tested, discussed below.

A second important question, relates to what extent evaluators' varying opinions about impact characterization will influence their interpretation of the REF2014 impact criteria of 'significance' and 'reach'. The way in which interviewees characterize impact will reflect their research and experiences, and we hypothesize that therefore such views will at least play some role in defining what characteristics impact evaluators look for when using the 'significance' and 'reach' criteria. For example, we hypothesize that interviewees varying conceptualizations of an 'impact outcome' will be reflected in how the 'significance' and 'reach' criteria are applied to the case studies—with different evaluators regarding the necessity for certain 'signposts' or 'indicators' for case studies to receive a high star rating. Similarly, we hypothesize that those interviewees, whose views of impact aligned more with the SIAMPI model of assessment, may score impact case studies in favour of a process—that is, not only on the basis of significance and reach, but also in consideration of how such significance and reach was achieved. These practices of looking for specific signposts within the data that reflect evaluators' own conceptualizations of impact may also be applied to other conceptualizations of impact, for instance, whether an impact needs to have a benefit, or just have the potential for a benefit.

The third question would be to see the extent to which evaluators' viewpoints develop differently during peer review panel deliberations. It is well established that peer review is not a socially dis-embedded process in which reviewers apply a set of objective criteria to assessment, rather, it is widely influenced by group interactions and social factors (Glaser and Laudel 2005; Bornmann 2008). In addition, it is a process of social interaction, with some scholars going as far as arguing that decisions are, in fact, socially constructed (Glaser and Laudel 2005). van Arensbergen has pointed to differences in, for example, status and expertise of the panelists, which play important roles in this type of interaction (van Arensbergen et al. 2014). And in their analysis of committee peer review, Lamont and Huutoniemi have demonstrated that it is customary rules followed by panelists which guide panel deliberations, and that these rules are never formally taught, but are learnt through professional socialization (Lamont and Huutoniemi 2011). These authors argue that social conditions brought about by such rules lead panelists to build consensus with other evaluators, and to perceive the process as fair (Lamont and Huutoniemi 2011). Alongside these studies, Langfeldt has shown that guidelines given to panels during grant peer review had little effect on the criteria the evaluators emphasized as key for assessment, rather panels based their decisions on

factors not supposed to influence judgment (Langfeldt 2001). Other research has also shown how group dynamics influence final judgments on peer review panels (Klein and Olbrecht 2011). The question thus remains as to how bringing people together from differing backgrounds influences what type of consensus is reached about the value of impact. This will be particularly interesting to see in terms of the academic- and user-evaluators, despite, no differences being found between their views expressed prior to assessment commencement.

Finally, in the case of formally evaluating impact under the REF2014, the newness of the impact criterion and the continued debate about its definition, raises questions about the actual practice of evaluating societal impact. Specifically, with the range of views about impact discussed in this article, including most evaluators viewing impact in terms of the 'final' outcome, questions are raised about how social interactions and evaluator roles will inform and determine the decision-making process required to reach a consensus of opinion. Indeed, by applying such considerations, as well as considerations of how the panel members interact, our final hypothesis is that the emerging dominant definition of impact used to guide its evaluation will reflect those most prominent opinions about impact expressed by the evaluators prior to the evaluation process and discussed here in this article. Indeed, Orbrecht and Bornmann (2010: 302) have argued this very point in terms of peer review assessment—that '*if the reviewers have different opinions, pressure for a consensus could result in acceptance of the majority position without adequate consideration of deviating opinions*' (Olbrecht and Bornmann 2010). They have argued that such occurrences are more likely in circumstances similar to REF2014, where no formal guidelines for the judgment process exist; the group works together as a panel over a long time; and when the panel is under stress to evaluate a large number of applications within a short period of time (Olbrecht and Bornmann 2010). This will therefore be one of the key themes of exploration in the next stage of this larger research project, which will include re-interviewing evaluators post-assessment. It is hoped that by exploring and comparing the different views of evaluators both pre- and post-evaluation, important insights into how societal impact can be effectively evaluated by peer- and user-based review in the future will be learnt.

Funding

References

Benda, W.G.G. and Engels, T.C.C. (2011) 'The Predictive Validity of Peer Review: A Selective Review of the Judgmental Forecasting Qualities of Peers, and Implications for Innovation in Science', *International Journal of Forecasting*, 27: 166–82.

Bernstein, A., Hicks, V., Borbey, P. and Campbell, T. 'A Framework to Measure the Impact of Investments in Health Research', *OECD Blue Sky II Forum Conference Proceedings, Ottawa, Ontario, Canada*, 25 September 2006.

Bornmann, L. (2008) 'Scientific Peer Review: An Analysis of the Peer Review Process from the Perspective of Sociology of Science Theories', *Human Architecture: Journal of the Sociology of Self-Knowledge*, 6: Article 3.

——. (2011) 'Scientific Peer Review', *Annual Review of Information Science and Technology*, 45: 197–245.

——. (2013) 'What is Societal Impact of Research and How Can it be Assessed? A Literature Survey', *Journal of the American Society for Information Science and Technology*, 64: 217–33.

Bornmann, L. and Marx, W. (2014) 'How Should the Societal Impact of Research be Generated and Measured? A Proposal for a Simple and Practicable Approach to Allow Interdisciplinary Comparisons', *Scientometrics*, 98: 211–19.

Bornmann, L. and Mungra, P. (2011) 'Improving Peer Review in Scholarly Journals', *European Science Editing*, 37: 41–3.

Bozeman, B. and Boardman, C. (2009) 'Broad Impacts and Narrow Perspectives: Passing the Buck on Science and Social Impacts', *Social Epistemology: A Journal of Knowledge, Culture and Policy*, 23: 183–98.

Brewer, J. D. (2011) 'The Impact of Impact', *Research Evaluation*, 20/3: 255–56.

Brown, N. and Webster, A. (2004) *New Medical Technologies and Society*. Cambridge: Polity Press.

Buxton, M. (2011) 'The Payback of 'Payback': Challenges in Assessing Research Impact', *Research Evaluation*, 20: 259–60.

Buxton, M. and Hanney, S. (1996) 'How Can Payback from Health Services Research be Assessed?', *Journal of Health Service Research and Policy*, 1: 35–43.

Buxton, M. J. and Hanney, S. (2008) 'Developing and Applying the Payback Framework to Assess the Socioeconomic Impact of Health Research', *Medicina Clinica*, 131: 36–41.

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

de Jong, S., Barker, K., Cox, D., Sveinsdottir, T. and van den Besselaar, P. (2014) 'Understanding Societal Impact Through Productive Interactions: ICT Research as a Case', *Research Evaluation*, 23: 89–102.

de Jong, S., van Arensbergen, P., Daemen, F., van der Meulen, B. and van den Besselaar, P. (2011) 'Evaluation of Research in Context: An Approach and Two Cases', *Research Evaluation*, 20: 61–72.

Derrick, G. and Samuel, G. (2015) 'The Societal Impact-Focused and Quality-Focused Evaluator: Balancing Evaluator Views about the Assessment of the Societal Impact of Health and Medical Research', *Social Science and Medicine*, In press.

Donovan, C. (2008) 'The Australian Research Quality Framework: A Live Experiment in Capturing the Social, Economic, Environmental and Cultural Returns of Publicly Funded Research', *New Directions for Evaluation*, 118: 47–60.

Donovan, C., Butler, L., Butt, A. J., Jones, T. H. and Hanney, S. R. (2014) 'Evaluation of the Impact of National Breast Cancer Foundation-Funded Research', *The Medical Journal of Australia*, 200: 14–18.

Donovan, C. and Hanney, S. (2011) 'The 'Payback Framework' explained', *Research Evaluation*, 20: 181–83.

Frank, C. and Nason, E. (2009) 'Health Research: Measuring the Social, Health and Economic Benefits', *CMAJ*, 180: 528–34.

Glaser, B. G. and Strauss, A. (1967) *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine.

Glaser, J. and Laudel, G. (2005) 'Advantages and Dangers of 'Remote' Peer Evaluation', *Research Evaluation*, 14: 186–98.

Grant, J., Brutscher, P., Kirk, S., Butler, L. and Wooding, S. (2009) *Capturing Research Impacts: A Review of International Practice*. Europe: RAND.

Hanney, S. (2005) 'Developing and Applying a Framework for Assessing the Payback from Medical Research', *New Technology and National Health Systems*. Kiel Institute's International Research Conference, the Institute for World Economics.

Hanney, S., Buxton, M., Green, C., Coulson, D. and Raftery, J. (2007) 'An Assessment of the Impact of the NHS Health Technology Assessment Programme', *Health Technology Assessment*, 11: iii–iv, ix-xi, 1-180.

Hanney, S. R., Grant, J., Wooding, S. and Buxton, M. J. (2004) 'Proposed Methods for Reviewing the Outcomes of Health Research: The Impact of Funding by the UK's 'Arthritis Research Campaign'', *Health Research Policy and Systems/BioMed Central*, 2: 4.

Haynes, A. S., Gillespie, J. A., Derrick, G. E., Hall, W. D., Redman, S., Chapman, S. and Sturk, H. (2011) 'Galvanizers, Guides, Champions, and Shields: The Many Ways That Policymakers Use Public Health Researchers', *The Milbank Quarterly*, 89: 564–98.

Holbrook, J. B. and Frodeman, R. (2011) ''Peer Review and the Ex Ante Assessment of Societal Impacts', *Research Evaluation*, 20: 239–46.

Kalucy, E. C., Jackson-Bowers, E., McIntyre, E. and Reed, R. (2009) 'The Feasibility of Determining the Impact of Primary Health Care Research Projects Using the Payback Framework', *Health Research Policy and Systems/BioMed Central*, 7: 11–4505-7-11.

Kerr, N. L., MacCoun, R. J. and Kramer, G. P. (1996) 'Bias in Judgement: Comparing Individuals and Groups', *Psychological Review*, 103: 687–719.

Klautzer, L., Hanney, S., Nason, E., Rubin, J., Grant, J. and Wooding, S. (2011) 'Assessing Policy and Practice Impacts of Social Science Research: The Application of the Payback Framework to Assess the Future of Work Programme', *Research Evaluation*, 20: 201–9.

Klein, T. and Olbrecht, M. (2011) 'Triangulation of Qualitative and Quantitative Methods in Panel Peer Review Research', *International Journal for Cross-Disciplinary Subjects in Education (IJCDSE)*, 2: 342–48.

Kwan, P., Johnston, J., Fung, A.Y.K., Chong, D.S.Y., Collins, R. A. and Lo, S. V. (2007) 'A Systematic Evaluation of Payback of Publicly Funded Health and Health Services Research in Hong Kong', *BMC Health Services Research*, 7: 121.

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

Langfeldt, L. (2001) 'The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome', *Social Studies of Science*, 31: 820–41.

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

Lee, C. J., Sugimoto, C. R., Zhang, G. and Cronin, B. (2013) 'Bias in Peer Review', *Journal of the American Society for Information Science and Technology*, 64: 2–17.

Levi, D. (2007) *Group Dynamics for Teams*. Los Angeles: Sage Publications.

Martin, B. R. (2011) 'The Research Excellence Framework and the 'Impact Agenda': are we Creating a Frankenstein monster', *Research Evaluation*, 20: 247–54.

——. (2012) 'The Evolution of Science Policy and Innovation Studies', *Research Policy*, 441: 1219–239.

Molas-Gallart, J. and Tang, P. (2011a) 'Tracing 'Productive Interactions' to Identify Social Impacts: An Example from the Social Sciences', *Research Evaluation*, 20: 219–26.

Mostert, S. P., Ellenbroek, S.P.H., Meijer, I., van Ark, G. and Klasen, E. C. (2010) 'Societal Output and Use of Research Performed by Health Research Groups', *Health Research Policy and System /BioMed Central*, 8: 30–4505-8-30.

Olbrecht, M., Tibelius, K. and D'Aloisio, G. (2007) 'Examining the Value Added by Committee Discussion in the Review of Applications for Research Awards', *Research Evaluation*, 16: 79–91.

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

Oortwijn, W. J., Hanney, S. R., Ligtvoet, A., Hoorens, S., Wooding, S., Grant, J., Buxton, M. J. and Bouter, L. M. (2008) 'Assessing the Impact of Health Technology Assessment in The Netherlands', *International Journal of Technology Assessment in Health Care*, 24: 259–69.

Penfield, T., Baker, M. J., Scoble, R. and Wykes, M. C. (2014) 'Assessment, Evaluations, and Definitions of Research Impact: A Review', *Research Evaluation*, 23: 21–32.

Research Councils UK. *Pathways to Impact*. <http://www.rcuk.ac.uk/ke/impacts/>.

Research Excellence Framework. (2011) *Assessment Framework and Guidance on Submissions*. REF 02.2011. <http://www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf>.

——. (2012a) *Panel Criteria and Working Methods*. <http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf>.

——. (2012b) *Main Panel A Criteria: REF01.2012*. <http://www.ref.ac.uk/pubs/2012-01/>.

Sandstorm, U. (2009) 'Cognitive Bias in Peer Review: A New Approach', *Proceedings of ISSI 2009 - 12th international conference of the international society for scientometrics and informetrics*, vol 2, 2th International Conference of the International-Society-for-Scientometrics-and-Informetrics. Rio de Janeiro, Brazil. 14–17 July. <http://www.forskningspolitik.se/DataFile.asp?FileID=172>.

Scoble, R., Dickson, K., Hanney, S. and Rodgers, G. J. (2010) 'Institutional Strategies for capturing Socio-Economic Impact of Academic Research', *Journal of Higher Education Policy and Management*, 32: 499–510.

Scott, J. E., Blasinsky, M., Dufour, M., Mandai, R. J. and Philogene, G. S. (2011) 'An Evaluation of the Mind–Body Interactions and Health Program: Assessing the Impact of an NIH Program using the Payback Framework', *Research Evaluation*, 20: 185–92.

Spaapen, J. and van Drooge, L. (2011) 'Introducing 'Productive Interactions' in Social Impact Assessment', *Research Evaluation*, 20: 211–18.

Strauss, A. L. (1987) *Qualitative Analysis For Social Scientists*. Cambridge: Cambridge University Press.

van Arensbergen, P., van der Weijden, I. and Van De Besselaar, P. (2014) 'The Selection of Talent as a Group Process. A Literature Review on the Social Dynamics of Decision Making in Grant Panels', *Research Evaluation*, 23: 298–311.

Wooding, S., Hanney, S., Buxton, M. and Grant, J. (2005) 'Payback Arising from Research Funding: Evaluation of the Arthritis Research Campaign', *Rheumatology (Oxford, England)*, 44: 1145–156.