palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 5 Apr 2016 | Accepted 25 Apr 2017 | Published 23 May 2017

DOI: 10.1057/palcomms.2017.40

OPEN

# The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability

Gemma Derrick[1] and Gabrielle Samuel[1]

**ABSTRACT** There are strong political reasons underpinning the desire to achieve a high level of inter-reviewer reliability (IRR) within peer review panels. Achieving a high level of IRR is synonymous with an efficient review system, and the wider perception of a fair evaluation process. Therefore, there is an arguable role for a more structured approach to the peer review process during a time when evaluators are effectively novices in practice with the criterion, such as with societal impact. This article explores the consequences of a structured peer review process that aimed to increase inter-reviewer reliability within panels charged with assessing societal impact. Using a series of interviews from evaluators from the UK's Research Excellence Framework conducted before (pre-evaluation) and then again after the completion of the process (post-evaluation), it explores evaluators' perceptions about how one tool of a structured evaluation process, pre-evaluation training, influenced their approaches to achieving a consensus within the peer review panel. Building on lessons learnt from studies on achieving inter-reviewer reliability and from consensus building with peer review groups, this article debates the benefits of structured peer review processes in cases when the evaluators are unsure of the criterion (as was the case with the Impact criterion), and therefore the risks of a low IRR are increased. In particular, this article explores how individual approaches to assessing Impact were normalized during group deliberation around Impact and how these relate to evaluators' perceptions of the advice given during the pre-evaluation training. This article is published as part of a collection on the future of research assessment.

[1] Centre for Higher Education Research and Evaluation, Educational Research, Lancaster University, Lancaster, UK
Correspondence: (e-mail: g.derrick@lancaster.ac.uk)

## Introduction

As public expectations from research evolve to include societal outcomes as well as academic achievements, so too does the framework used to include the evaluation of these notions of research excellence. One method used to promote the legitimacy and authority of new and untested criteria such as societal impact (Derrick and Samuel, 2016b), is by employing traditional academic peer review for such assessments. As stated by the British Academy, "the essential principle of peer review is simple to state: it is that judgements about the worth or value of a piece of research should be made by those with demonstrated competence to make such a judgement" (Academy 2007). However, the employment of peer review comes at a risk that not all "experts" will be able to reach a consensus on their assessment of impact, as differing research traditions, worldviews and even methodologies alter an expert's notion of excellence. Difficulty in reaching a group consensus results in low inter-reviewer reliability (IRR) and a longer, less efficient peer review process as assessors spend more assessment time negotiating a consensus and building a committee culture (Olbrecht and Bornmann, 2010) which is necessary for guiding group evaluations. However, for evaluation processes that determine the allocation of public funds, a high degree of IRR gives the public reassurance that the evaluation process was sufficiently objective and accurately reflected agreed-upon priorities (Tan *et al.*, 2015). Reaching a desirable level of IRR through group consensus is therefore both a political and public necessity for public research funding agencies keen to promote the legitimacy of their assessment process and criteria.

The evaluation of "impact" under the UK's Research Excellence Framework 2014 (REF2014) represented the world's first formal, ex-post (after the event) assessment of how research has had an impact on society beyond academia. Debate before the REF2014 focused on the variety of definitions of impact, and the consequent issues associated with its formal assessment, where there was a lack of consensus about what constituted excellence in impact. Furthermore, the inclusion of assessments of the societal outcomes from research has also necessitated that research extend its definition of who represents a research "peer" to include non academic actors. These new peers aim to contribute differing expertise and experience, and to provide a non-academic perspective about the assessment of how research influences society (Derrick and Samuel, 2016a). For many evaluation processes, the REF2014 included, there are a variety of new, non-academic actors incorporated into the panels, especially for societal impact assessment.

However, extending the range of expertise available to panels to guide assessments also increases the amount of conflicting values and definitions of excellence, and therefore makes a group consensus more difficult to achieve—especially if the criterion is new, such as with impact (Langfeldt, 2001; Samuel and Derrick, 2015; Derrick and Samuel, 2016b). Indeed, previous research has highlighted the difficulty associated with peer review panels resolving other, unfamiliar and potentially ambiguous criteria such as "interdisciplinary research" in light of different interpretations of the concept within the panel (Lamont, 2009; Huutoniemi, 2012). This increased level of "noise" (Danziger *et al.*, 2011) during the evaluation process results in a low level of IRR that results in decisions taking a lot of time to achieve, inappropriate proxies being used in order to base evaluations, or assessments not being made (Cicchetti, 1991). Reviewer training before the assessment is one technique that reduces panel "noise" in session (Cicchetti, 1991) and ultimately improves IRR (Sattler *et al.*, 2015) and the efficiency of the evaluation process. Very little research has focused on achieving improved IRR in peer review panels, and even less research has done so using qualitative methods since previous research of IRR within peer review panels has focused solely on producing quantitative indicators of evaluator concord.

In contrast to previous studies of IRR, this article explores the consequences of a structured peer review process that aims to increase IRR and ensure an efficient evaluation. It does this by utilizing a series of interviews conducted with evaluators from the REF2014 prior to (per-evaluation interviews) and then again after the evaluation process had taken place (post-evaluation interviews). It explores evaluators' perceptions about how a tool of a structured evaluation process, the calibration exercise, influenced their approaches to achieving a consensus within their peer review panel. In light of the lack of publically available information about the content of the calibration exercises, this study's qualitatively analysed accounts from the participants in the calibration exercise provide useful insight into its influence on individual evaluation approaches. Moreover, this "in vitro" approach to research—combining two sets of interviews before, and post assessment (Derrick, forthcoming), allows for the measurement of changing attitudes as a result of panel discussions in session, as well as provides a proxy for investigating the interplay of values, tensions and power that occur as peer review committees conduct evaluations. For the purposes of this article, the pre-evaluation training exercises are referred to as the "calibration exercises", to reflect the terminology used by study participants.

With governments and funding agencies around the world interested in incorporating formal societal impact criteria into their research assessment protocols, this article provides insights into how societal impact criteria can be introduced into a traditional peer review model of evaluation. These insights are also essential to consider when the criterion (Impact) is new or unfamiliar to evaluators and there is a risk of low IRR alongside unstructured peer review processes. Likewise, the insights are important for future REF2014 assessment exercises that will involve more seasoned impact assessors (for example, REF2021), when providing a structured peer review process involving pre-evaluation training may not be as necessary as it was in REF2014. Therefore, a high IRR could be achievable within panels through an unstructured approach that would allow for informed discussions between panel members to guide the assessment process.

The next section will briefly provide a review of the literature and outline the main arguments behind the inclusion of training exercises in peer review panel-based exercises, as well as their influence on the ability of peer review panels to reach a consensus (high IRR). In particular, it will concentrate on the influence of such training exercises on the validity, efficiency and fairness of review processes, as well as discuss the potential benefit committee discussions have on the assessment process and outcome. The methods section will discuss the methods used in this study. The results section will explore how reviewers anticipated their role of assessing the "impact" criterion before the assessment process, and reviewers' perceptions about how the training influenced their formal assessment of Impact. Specifically, it will debate how the inclusion of the calibration exercise, on the one hand, worked towards forming a consensus around an unfamiliar criterion in a relatively short amount of time. However, this was arguably, at the expense of sufficient time and intellectual space in which to resolve the myriad of actor (peer) values concerning impact independently from the advice received during the calibration exercises. For future frameworks that incorporate an impact criterion, allowing sufficient time for group debate in lieu of pre-evaluation review, is important for the future of societal impact assessment in peer review settings, regardless of the risk of a low IRR.

## Literature review

**IRR in peer review: Validity and reliability**. The literature surrounding IRR has primarily focused on the review of journal manuscripts, rather than within peer review panels. For these studies, IRR is the extent to which two or more, independent reviews of the same scientific document agree (Cicchetti, 1991). In peer review panels, IRR relates to the concept of consensus building and the extent that evaluators agree on the decision to award research funding or grant a measure of esteem (award, promotion and so on). Relatively little research has specifically focused on the IRR between evaluators within peer review panels, and even less research has investigated the formation of consensus and achieving a high IRR using qualitative methods. To utilize these methods allows one to extend common questions about *if* IRR was reached and to *what extent,* to instead focus on *how* it was achieved and the processes underlying how evaluators reached a consensus.

Furthermore, whereas journals use the double blinded peer review system as a measure against biases in refereeing (Hemlin and Rasmussen, 2006), guarding against such bias is more challenging for group panel decisions. This is because an important weakness of peer review is the fact that ratings to the same submission given by different reviewers will differ, resulting in a low IRR (Marsh *et al.*, 2008). In addition, in contrast to manuscript submission procedures, panel decisions are made in real time, collectively, and ideally with minimum interference by external sources (in the case of journal submissions, this role would be assigned to the editor). Instead, the process relies on the development of a committee culture (Olbrecht and Bornmann, 2010) to guide the evaluation process. For governments and funding agencies, a high level of IRR is desirable, as it increases the probability of an efficient and therefore less expensive review process that includes the benefits of peer review. However, achieving high IRR in peer review, as discussed below, does not always ensure the best process of peer review for the evaluation.

Typically, as with all group processes, reaching a sufficient level of group consensus can be a highly time consuming exercise, in which reaching decisions are made at the cost of achieving a high level of IRR (Olbrecht *et al.*, 2007). However, for group panel decisions it has been argued that achieving a low IRR is something to be strived for since "too much agreement is in fact a sign that the review process is not working well, that reviewers are not properly selected for diversity, and that some are redundant" (Bailar, 2011). According to (Langfeldt, 2001), the quality of discussion during panel-based assessment is vital to the development of a committee culture that guides the evaluation, and therefore low IRR can actually increase the validity of the review: "low IRR on a panel is not indication of low validity or low legitimacy of the assessments. In fact, it may indicate that the panel is highly competent because it represents a wide sample of the various views on what is good and valuable research" (Langfeldt, 2001). Indeed, the famous Cole *et al.* (1978) study on NSF grant submissions suggested that the low level of reliability in peer review evaluations are not an artefact of the peer-review system or of reviewer bias, but reflect the low levels of cognitive consensus that exists at the research frontier for all scientific disciplines (Cole *et al.*, 1978). This, in turn, reflects the quality of the peer review system where many different viewpoints are actively considered (Eckberg, 1991), and there are usually many, often legitimate, grounds for reviewer disagreements as different reviewers will have different conceptualisations of what represents worth, especially when it comes to research and researchers within a field. In truth, for peer review panel decisions, a certain level of inter-reviewer (un)reliability is to be expected, and for impact this unreliability is to be amplified as evaluators endeavour to deal with new and uncertain criteria (Derrick and Samuel, 2016b), as well as with a greater diversity of panel members (Derrick and Samuel, 2016a). Despite low IRR considered as essential to improve the quality of the discussions for group panel review, high IRR is often a necessary pre-condition for an efficient review process, and there are conflicting opinions on how to achieve this. Below we discuss the role of one method in particular, pre-evaluation reviewer training and its role in a structured peer review process of an uncertain criterion, Impact.

**Structured peer review through providing pre-evaluation reviewer training**. Using a structured review process is not a new concept for peer review, nor is its use unique to the REF2014 impact evaluation process. Funding agencies introduce aspects of structure into the review process to compensate for the perceived unreliability of peer review. This peer review unreliability is typically associated with a low level of IRR (Cicchetti, 1991, Kravitz *et al.*, 2010), and is simply due to the difficulties associated with bringing a large number of (usually) highly educated people with sometimes divergent views together and asking them to reach a consensus in a short space of time. Despite this and infamous unreliability, academic peer review remains the assessment method of choice for government funding agencies, because of its reputation as the "gold standard" for evaluating scientific merit (Chubin and Hackett, 1990; Demicheli and Di Pietrantonj, 2007).

A number of techniques can be employed to increase the IRR within peer review processes. These introduce aspects of structuring the debate to reflect the priorities of the funding agency. These include the ranking method proposed by (Hodgson, 1995); the sandpit method; workshop review; bibliometric data driven decisions; and the use of discretionary grants (Vener *et al.*, 1993, Gordon and Poulin, 2009, Holliday and Robotin, 2010, Wu *et al.*, 2011). For the purposes of this article, one method in particular is discussed, the use of reviewer training (calibration exercises).

The political objective of peer-evaluation reviewer training in peer review is to facilitate an efficient review system that, through the authority associated with academic peer review, increases IRR, and therefore results in a relatively fast and inexpensive evaluation. Higher IRR in peer review panels has been reported to occur when the evaluation priorities of the funding agency are in line with those of the reviewers (Abdoul *et al.*, 2012). However, for criteria where this priority is yet to be determined, and is in constant debate, this alignment of views is unlikely to happen naturally and some level of reviewer education is deemed necessary to increase IRR and also to maintain the legitimacy and validity of the review process itself. In addition, there are obvious advantages associated with this approach when the numbers of applications are high (Abrams, 1991), and/or when dealing with a large number of peers in committee (Thornley *et al.*, 2002).

In his commentary to Cicchetti (1991), Delcomnyn (1991) suggests that the greater refinement of the peer review system should either make the review criteria uniform or more explicit; or "train" the reviewers (Delcomyn, 1991). The first recommendation (to make the review criteria more explicit), based on Cicchetti's (1991) finding on IRR in manuscript revisions was that two reviewers make similar comments about a paper, but have differing recommendations as to its acceptability. The second recommendation (to train the reviewers) was based on a personal reflection that one's academic reaction to other work becomes more sophisticated with time, experience and input from other reviewers. Cicchetti (1991) suggested that for the peer review of manuscripts at least, training of reviewers is pivotal in

increasing both the reliability and validity of peer review outcomes and therefore directly related to IRR. However, there is little evidence to see how pre-evaluation training could influence consensus building and therefore IRR within peer review groups.

As many studies have shown, peer review panel discussions did not improve the reliability (Fogelholm *et al.*, 2012) or fairness (Olbrecht *et al.*, 2007) of evaluation outcomes, instead they point to the group discussion aspect being the major limiting factor of an efficient panel-based peer review system. Inconsistencies in reviews is further exacerbated by differences among committee membership, where despite the existence of funding guidelines and strict criteria for assessment, committees tend to develop different standards and unique cultures. In addition, the consistency with which assessment criteria are applied by the committee is also affected by the complex personal interrelationships that develop as reviewers work together in closed spaces, and under pressure. It is this unique culture, and consideration of the differing standards and opinions of panel members that, it is argued, contributes to robust peer review outcomes (Langfeldt, 2001, Langfeldt, 2004).

Pre-evaluation training, rather than limiting the benefits of group discussion as the above argument suggests, ensures that the themes under discussion are kept on topic, and reflect the evaluation priorities and objectives of the funding agency. Despite fears to the contrary, this kind of structured peer review, where the influence of uncontrolled or superfluous group discussions are minimized, was shown to result in no difference in the level of IRR, compared to those reviews that did not limit the direction or topic of discussion (unstructured reviews) within peer review committees.

Offering pre-evaluation training offers a solution for when non-academic reviewers are included to evaluate societal impact, but lack sufficient expertise or capacity with the technical merit of proposals to be fully incorporated into the academic peer review environment (Chubin, 1994, Olbrecht *et al.*, 2007). These results indicate the possibility that structured peer review, through the provision of pre-evaluation training, can offer many advantages as well as the possibility of an efficient peer review system driven by a high IRR, without any loss in academic rigour or reputation. Finally, for funding agencies, pre-evaluation training offers the advantage of allowing them to capture a direct assessment of specific criteria and assign that criterion a specific weight. In this article, we argue that a structured peer review system could foster the evolution of Impact as a recognized component of evaluating research excellence.

## Methods

**The UK research excellence framework 2014.** The United Kingdom has a long history of research evaluation frameworks, from the first Research Assessment Exercise in 1986 involving a small number of "traditional" universities, to the most recent Research Excellence Framework in 2014 (REF2014) which involved all UK universities (Bence and Oppenhein, 2005). Despite a long history of operationalising these Frameworks, it is only since 1992, that the outcomes of such evaluations became linked to a significant funding allocation between the universities based on the ratings of subject panels (Bence and Oppenhein, 2005).

Though the weighting and naming of the criteria used in these exercises has changed over time, the criteria have firmly remained associated with traditional aspects of research excellence, quality and esteem, such as publications, citations and competitive grant funding. Indeed, the most recent REF2014 has mirrored the importance of these traditional criteria by dedicating 65% of the total score dedicated to a traditionally driven peer review of research "Outputs"; and 15% to an assessment of the Higher Education Institute's "Environment" (HEFCE, 2011). The final 20% of the assessment, however, differed in that it dedicated a formal proportion of its overall evaluation criteria to considerations of excellence beyond traditional academia, or of the societal impact of research. The new criteria, which was labelled "impact", was defined as "… an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" (HEFCE, 2011; 26). The addition of this

**Table 1 | The number of interviews conducted with REF2014 main Panel A and its 6 sub-panels**

| *Sub-panel name* | *Number of members on each sub-panel* | *Number of academic evaluators* | *Number of research user evaluators* | *Number of Pre-evaluation participants* | *Number of Post-evaluation participants* |
|---|---|---|---|---|---|
| Main Panel | 19 | 14 (73.7%) | 5 (26.3%) | 8 (42.1%) | 8 (42.1%) |
| Sub-panel 1—Clinical Medicine | 39 | 32 (82.0%) | 7 (18.0%) | 10 (25.6%) | 9 (23.1%) |
| Sub-panel 2—Public Health, Health services and Primary care | 27 | 23 (85.1%) | 4 (14.9%) | 13 (48.1%) | 13 (48.1%) |
| Sub-panel 3—Allied Health Professions, Dentistry, Nursing and Pharmacy | 51 | 42 (82.3%) | 9 (17.7%) | 14 (27.5%) | 13 (25.5%) |
| Sub-panel 4—Psychology, Psychiatry and Neuroscience | 35 | 28 (80.0%) | 7 (20.0%) | 9 (25.7%) | 7 (20%) |
| Sub-panel 5—Biological Sciences | 35 | 30 (85.7%) | 5 (14.3%) | 6 (17.1%) | 5 (14.1%) |
| Sub-panel 6—Agriculture, Veterinary and Food Science | 29 | 16 (55.1%) | 13 (4.9%) | 4 (13.8%) | 4 (13.8%) |
| TOTAL | | 185 (78.7%) | 50 (23.2%) | 62 (28.8%) | 57 (24.2%) |

assessment criterion represented the world's first mandatory, formal, ex post, peer and end user review assessment of societal impact linked to funding allocation. As such, the REF2014 provided a good model for analysis of societal impact in the context of the process of evaluation through peer review.

Assessment of the "Impact" criterion was conducted according to the generic definition given above, by reviewing 4-page case studies submitted by each university, as well as an impact template that described the wider university strategy of facilitating the translation of research into impact. The structure of the 4-page case studies was tightly controlled by a template supplied by HEFCE, where universities must nominate pieces of underpinning research and then proceed to explain how this research has had an impact. This underpinning research must be considered to have reached a threshold of no less than 2 stars in quality ("quality that is recognized internationally in terms of originality, significance and rigour") (HEFCE, 2011)

As with previous assessment exercises, the REF2014 evaluation was performed by peer review. In total, 4 Main Panels (A-D) which included up to 6, smaller sub-panels or Units of Assessment (UoA), oversaw the evaluation process. Each UoA was associated with a research discipline and included up to 30 research experts, or peers. For the assessment of the impact criterion, this peer group included a combination of academic, and user-evaluators (see Table 1). However, for the purposes of the objectives of this article, demarcation between the results obtained from each evaluator types was not performed.

In recognition of the newness of the "impact" criterion as well as to ensure the consistency of the resulting evaluation, a series of calibration exercises were conducted with each panel and sub-panel (www.ref.ac.uk, 2014). The paucity of information provided by HEFCE regarding the structure of these calibration exercises prior to the evaluation taking place, has meant that all information regarding these exercises has had to be solely derived from the individual experiences described in the pre- and post- evaluation interviews analysed.

*Research design*. The study utilized the *in-vitro* approach to assessing academic evaluative cultures. The approach is discussed at length in (Derrick, forthcoming). In essence, the *in-vitro* approach combines insights from interviews conducted both prior to, and after the peer research evaluation process. Through the triangulation of the (1) raw, baseline views expressed prior to the evaluation process (pre-evaluation interviews); with the (2) views expressed after the process (post-evaluation interviews), an understanding is provided of how the evaluation issues were resolved in committee (Derrick, forthcoming). In light of the well documented issues associated with gaining sufficient access to peer review panels (Holbrook and Hrotic, 2013), the *in vitro* model offers a reliable and robust research design for the investigation of evaluation processes (Derrick, forthcoming).

**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 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 covered six Sub-panels: (1) Clinical Medicine; (2) Public Health, Health Services and Primary Care; (3) Allied Health Professions, Dentistry, Nursing and Pharmacy; (4) Psychology, Psychiatry and Neuroscience; (5) Biological Sciences; and (6) Agriculture, Veterinary and Food Sciences. The number of evaluators assigned to each sub-panel 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 Sub-panel.

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). 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), before the interviews taking place. (Holbrook and Hrotic, 2013; Derrick, forthcoming)

*Interviews*. Interviews were conducted via the telephone, skype, or face-to-face, and were recorded, and transcribed for analysis. Pre-evaluation interviews were conducted before the REF2014 evaluation process and group calibration exercises started, whereas post-evaluation interviews were conducted by GS and GD after the completion of the REF2014 evaluation process. All Interviews lasted between 35 min and 2 h and were semi-structured.

To ensure the interviewees' views about the definition and characterization of impact were not influenced by the interview discussion, the pre-evaluation 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.

The post-evaluation interviews followed a similar format to those above, and explored participants' experiences of the evaluation process. In particular, interview themes were based around findings from the pre-evaluation interviews and included issues relating to any changes of opinion towards evaluators' definition of societal impact because of their experience; a description of the evaluation process —including how they assessed societal impact, how they reached decisions about the relative value of different societal impacts, and any issues concerning the assessment impact; interactions between panel members and reaching consensus; and any tools or insights they learnt as a result of the process. Directly relevant to the objectives of this article, was the line of questioning and prompts in the post-evaluation interviews focused on the calibration exercises and their influence on how individual evaluators approached the formal evaluation of Impact.

*Analysis*. In the interests of confidentiality, all participant information was coded and entered into NVivo (qualitative analysis software package) for 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; Charmez, 2006). This method also empirically grounds theorising to data so that abstract conceptualizations can be developed from a close analysis of the data. Duplicate coding by both the first and second author was cross-checked to ensure reliability of data. A more in-depth discussion of our analysis has previously been provided (Derrick and Samuel, 2016b).

*Quantitative results*. Where possible, the calculation of quantitative data involved taking the code analysed qualitatively using grounded theory, and then manually checking individual sources within nVivo to determine numbers. Other quantitative results were determined by a Likert-based survey that was distributed to participants at the conclusion of the post-evaluation interviews. This survey was used to test the validity of the themes that were identified in the above described, qualitative analysis of the interview data.

## Results

**Anticipated feelings towards impact evaluation: A role for calibration**. Interviews conducted before the assessment of societal impact *(-PRE)* highlighted a strand of uncertainty amongst panellists about how to evaluate this criterion ("I have no idea what's going to happen as to how we rank and score impact" *(P1OutImp2-PRE)*), with many evaluators speaking openly about their concerns ($n = 26/62$). Panellists were apprehensive about the newness of the criterion ("this is a completely new exercise...[…] …we actually haven't got a clue what we're going to do; we have never done this before" *(P3OutImp1-PRE)*), and their relative inexperience as evaluators of societal impact ("for a lot of us it's not within our experience directly" *(P5OutImp4-PRE)*). This is in contrast to their experience of evaluating more traditional notions of research excellence, which researchers described as their "bread and butter". Indeed, the assessment of the impact criterion was perceived as something that was outside the expertise of the evaluators and therefore made them "nervous" ("this impact stuff, we just don't know. So I feel a little bit nervous about it" *(P0P2OutImp1-PRE)*). Panellists were especially concerned about how to resolve the variety of values, views and beliefs evaluators expressed about "impact" as a concept (Derrick and Samuel, 2016b): "I think there is going to be quite a lot of difficulties in getting everybody to the same place in terms of what is impact" *(P0OutImp2-PRE)*.

At the same time as evaluators expressed their "nervousness" about the prospect of impact, they were also aware that they would be offered reviewer training, otherwise known as a

calibration exercise, in order to prepare for the impact evaluation process. Indeed, prior to the assessment and calibration exercises, evaluators felt hopeful that the calibration exercise would provide a level of guidance that would compensate for their lack of experience regarding impact described above ("when we do our calibration thing towards the end of the year…this [assessing impact evidence] will be one of the hottest areas really for us to really focus in on. And we may have to sort of generate some rules at the end of the day" *(P0OutImp6-PRE)*). A number of panellists ($n = 21/62$) anticipated that the calibration exercise would demonstrate both the commonalities and discrepancies that existed between different panel members ("I don't know how much commonality there will be because we haven't done the calibration exercise yet" *(P1OutImp2-PRE)*); standardize differing viewpoints ("I do think there will be a difference at least to begin with. But part of our calibration exercise will be to come and make sure we have…hopefully adopted each other's points of view" *(P0P2OutImp1-PRE)*); and resolve any underlying issues; "probably I will have a better feel for what's going on after that [the calibration exercise]" *(P1OutImp1-PRE)*. Specifically, the calibration exercises were expected to act as an arena where evaluators could openly discuss their differing views about impact and, through this discussion, allow sufficient opportunity for the group to reach a consensus about how the evaluation should proceed in light of the myriad of values associated with aspects of impact (Samuel and Derrick, 2015; Derrick and Samuel, 2016b). Through this calibration exercise, the evaluators were aware of its importance in serving as a tool that would facilitate the evaluation, as well as increase the inter-rater reliability of the assessment process;

> "Research excellence, I think it's more straightforward….. for impact its new, and we haven't gone through comparing different assessors and how consistent their assessment is and how we're going to get agreement at the end, it might be very high, it might not be, I don't know" *P1OutImp2-PRE-*.

A similar calibration exercise was performed for the assessment of the "output" criterion of the exercise. During this calibration, participants described how, following their initial assessment of an output example, that spreadsheets were used to "light up" each score so panellists became aware of (if they were in "the red"), and could later modify, their assessments relative to the average score of other assessors; "there was a spreadsheet, with some people lit up in red and other people lit up in green as their variation from the mean" (*P1OutImp3-PRE*). Despite this, discussion between evaluators about their differences in opinion was encouraged ("we had quite diverging scores on the same paper, and then we discussed it" (*P4OutImp4-PRE*)), so that commonality could be established and consensus reached between evaluators with time; "[we] discussed through until everybody was really clear why they thought it was a particular score and reached[ed] consensus" (*P3OutImp9-PRE*). It was commonly hoped that, similar to the output criterion calibration, the impact calibration exercise would be able to secure a similar consensus; "I hope that we have the same opportunity to discuss these potential areas of conflict before we do it for real" (*P2OutImp7-PRE*).

Panellists stressed that, given their uncertainty about the prospect of evaluating the impact compared to their relative confidence with assessing the outputs, having the calibration exercise was particularly important as it allowed evaluators to listen and "digest" other evaluators' views. Indeed, the newness of the impact criterion was considered an advantage as it insisted that everyone remained "open" to discussion; "I think everybody is very nervous, which means I think that most people are very open (*P2OutImp-PRE9*). It was hoped that the calibration exercise would allow evaluators the time, and tools necessary to assess this new and untested criterion, fairly and efficiently; "I assume we'll have calibration exercises and from that will come a process that we are all content with…to make sure that it is fair" (*P3OutImp9-PRE*). Only a minority ($n = 3/62$) of evaluators explicitly expressed reservations about whether evaluator panels would be able to reach the necessary consensus, especially considering the different types of evaluators present on the panels; "Considering that these users provide a different viewpoint about what impact is and it's important, how do you approach resolving any conflicts between academic and non-academic evaluators about what is valuable in impact?" (*P3OutImp9-PRE*).

**Pre-evaluation training (calibration exercises): Resolving different view values and approaches to impact evaluation**. As discussed above, the panellists described a high level of nervousness regarding the evaluation of impact. The calibration exercises, therefore, were seen as an opportunity for panel members, even interdisciplinary panels, to clarify expectations of the assessment process and form a common lens to guide the impact evaluation, especially in relation to the 1–4 star impact scale[1].

> "Having been quite nervous about the potential of it a year ago because the criteria at that point in time weren't very clear. But we had very clear discussions at the calibration which made it—even though we were in a very multi-disciplinary subpanel—very clear what was being expected as impact at each of the star levels, so I think everything became much clearer as you went through the process and did some calibrations." (*P3OutImp9-POST*)

As described by the above quote, much of the pre-evaluation anticipation about the calibration exercise was calmed post-exercise. The calibration exercise served to make panel considerations of impact easier than expected ("it was just made a lot easier for us than we had initially anticipated" (*P4OutImp3-POST*)). Panellists spoke about the calibration exercise being "useful" (*P3OutImp3*) and "necessary" (*P1OutImp7-POST*). Without it the assessment "would have been an enormous mess" (*P0OutImp3-POST*). Only some panellists felt that the calibration offered little benefit ($n = 3/57$, 5.3%) and was no substitute for experience ("the calibration exercise gave me confidence……but it didn't actually give me everything that I needed because you never knew what you were going to read" (*P3OutImp1-POST*)). And by the end of the calibration, there was a sense that panellists had "come together" (*P3OutImp2-POST*) in their approaches to assessing impact, and form a committee culture; "the calibration exercise that the distribution tightened quite considerably and the mean assessment shifted" (*P4OutImp2-POST*). This is not to imply that all issues associated with impact evaluation were addressed, rather, that unresolved issues were flagged "to be taken up later" (*P2OutImp6-POST*).

**Consensus building through discussion**. Panel discussions featured heavily as a method for building consensus during the calibration exercise, "we simply had these ferocious discussions" (*P0OutImp3-POST*). Such discussions aimed to bring evaluators with different views on to the "same planet" by clarifying the REF2014 guidelines on impact evaluation and the scale used to rank each impact;

> "They also helped to get the whole group on the same planet of trying to gauge whether it was four star, three star, two star and one star and non-defined. So you had to do

> this with the whole group and then there were still discussions afterwards, after the scoring, where there were mismatches in scoring. But those were sort of resolved very well. But without the calibration, it would have been impossible to do the job." (*P0OutImp3-POST*)

The calibration exercise also aimed to "get the exact ground rules sorted out" (*P1OutImp4-POST*) in relation to the star ranking profiles for impact assessment. During these discussions, panellists were encouraged to "put all the cards on the table" so that all individual viewpoints and different assessment approaches would be aired and debated. In addition, it was also vital that panellists would listen and be respectful of the views of others, and this openness was essential for the committee to learn and understand how to evaluate impact. Interviewees expressed the need to "re-think how [to] assess things" (*P5OutImp2-POST*) and, on occasion, the need to put aside personal views to reach a consensus; "people were happy to shift or settle" (*P3OutImp5-POST*). The length of the discussion allowed, however, was not finite, with panellists aware that despite the numerous issues associated with impact evaluation (Frank and Nason, 2009, Grant *et al.*, 2010, Martin, 2011; Bornmann, 2013), the purpose of the calibration exercise was not to solve every issue that may arise during the evaluation; "we probably would have agreed to disagree but also agreed that the process was to come to a conclusion" (*P2Imp2-POST*).

To aid these discussions during the evaluation exercise, the role of a "Chair" was essential to ensure that deliberations remained on topic and adhered to the guidelines. In addition, this panel umpire ensured that sticking points were drawn out within the discussions and, where possible, that a conclusion was reached that would guide the future impact evaluation process; "…people had a chance to…put their views forward…he [the Chair] was also very good at pulling them together and say[ing], I've got a sense that this is…the majority of you, this is where we're going" (*P2OutImp2-POST*). The Chair, therefore, both during the calibration exercise and in the impact evaluation, served to remind panellists of the agreements made during the calibration process, as well as to bring confluent viewpoints together to both increase the efficiency of the evaluation, and increase IRR of the panel sessions. This served to reinforce the structured nature of the peer review process and ensure that lessons learnt during the calibration exercise were carried over to, and applied equally to submissions during the formal evaluation of Impact.

**Shifting views through steering exercises**. While discussion facilitation, as described above, was the key to calibration, and for essentially increasing IRR of the assessment, the evaluation also acted as a broader educational and political exercise. Indeed, during the calibration exercise, evaluators recalled that senior figures ("the REF people") who were not formally members of the evaluation panel, educated the panellists about the exact meanings and definitions about impact as stated within the evaluation guidelines, as well as outlining their role as assessors; "the REF people…went through it very, very clear[ly] about what we were to look for and the steps you were to go through in assessing impact" (*P3OutImp4-POST*). The panellists viewed these talks from these senior figures as important to guide them through an assessment process about which they had previously felt nervous and unsure.

> "Despite the fact that we all went in saying, how are we ever going to get this right, how are we going to judge what is a one, two, three or four with these rather vague written cases. But when it came to it, it wasn't nearly as hard….and that's partly due to the fact that we got very clear advice from HEFCE as to how to interpret what we saw and how to grade what we saw". (*P1OutImp4-POST*)

This "clear advice" no doubt helped to provide an efficient assessment of impact, and increase the IRR, but questions remain about the validity of an academic assessment that is not the accumulation of expert opinions, but of anonymous senior figures that would have been possible through a less structured peer review process. This purpose of the calibration exercise is in direct contrast to the purpose of the previously discussed strategy, which was to increase discussion so that panellists reached an agreement about the assessment, thereby organically forming a committee culture around impact. Moreover, panellists commented that the sessions held by these senior figures made the impact evaluation approachable as, rather than having to reach any particular judgements about impact, or needing to come to any conclusions about how to assess the criterion among themselves, panellists were given instructions and were "told what to do"; we were all told what we have to do. So we would just follow what we were told" (*P4OutImp1-POST*).

The result of this instruction was that the consensus built by the panel around impact was the result of a more structured exercise designed to train committee members to adopt a particular form of assessment, rather than to encourage an open, unstructured debate of ideas. As such, discussions and panel conclusions were reportedly steered to conclusions previously decided upon by the overarching main panel ("…had a very strong steer from the whole process" (*P5OutImp1-POST*)). Examples of this included discussions being steered towards evaluating impact in a specific way; "we were told very clearly 'no, it's not about that way of marking' " (*P1Imp1-POST*), or when panellists were "warned" about their evaluation of impact; "we were pretty clearly warned" (*P2OutImp3-POST*).

Although, at times the discussion was key to the exercise, "when push came to shove" these discussions were steered in a particular direction so that they adhered to the impact definition and assessment guidelines provided by HEFCE; "I recall some fairly strong guidance from the senior people from the Main Panel" (*P1OutImp1-POST*); as well as to the political purpose of the impact criterion; "I think we debated that a lot in the pilot phase. I think when push came to shove….it was a little bit of a steer….we were encouraged to be more positive rather than not." (*P4OutImp2-POST*). As a result, panellists felt prepared to evaluate impact by the end of the calibration exercise; "…no matter how diverse we were in terms of our disciplines, we were so well trained that I think if somebody put me into a physics subpanel I probably could have judged their impact." (*P3OutImp9-POST*).

**Putting impact evaluation into practice post calibration**. Following the calibration exercise, the formal impact evaluation process, as to be expected, became an iteration of the messages received during the calibration. In this way, panellists were "encouraged" (*P0P1OutImp1-POST*) to adhere to the HEFCE provided guidelines mechanistically; "it was a fairly mechanistic operationalization [of impact]" (*P2OutImp6-POST*). As such, panellists described "sticking very closely to the guidance" (*P2OutImp1-POST*) during the impact assessment process; "we stuck very much to the letter of what was in the instructions" (*P4OutImp1-POST*). These guidelines, became known as the impact assessment "bible" that were to be followed irrespective of individual assessors' views about impact;

> "…all the REF teams were able to refer back to the REF guidance which after all was a Bible for our assessment, whatever your view of impact or otherwise. The REF

guidance was the thing that you should be referring back to. And so the calibration exercise pulled up a few things in that guidance that people hadn't borne in mind or hadn't remembered or hadn't reached. It was the guidance. So it's actually….so that everybody was aware of exactly what they were supposed to be scoring." (*P2Imp2-POST*)

However, panellists 40/48 (83.4%) interviewed in the post-evaluation interviews did not view these guidelines as restrictive to the assessment process, rather, most interviewees were grateful to have the guidelines available to help the assessment of this new, previously untested criterion (42/48, 87.5%). As one user-evaluator remarked, regardless of what your personal view surrounding the nature and value of impact;

"…it was also helpful in that over and over it was pulling out—all the REF teams were able to refer back to the REF guidance which after all was a Bible for our assessment, whatever your view of impact or otherwise. The REF guidance was the thing that you should be referring back to. (*P3Imp2-POST*)

This reliance on the lessons learnt in the calibration exercises to guide evaluative practice was not restricted to user-evaluators. In fact, academic evaluators also reported how the guidelines and lessons from the calibration exercises were vital to ensure that the group approached valuing the Impact criterion similarly.

Well, no because we were all told what we have to do. So we would just follow what we're told and we just—what we didn't do was go into a room and say what do you think impacts is, what do you think impact is. What we did was come in and say this is the definition, and this is what we're going to stick to, and now we are going to go through a calibration exercise the way that everyone does and that's what we did. (*P4 OutImp1-POST*)

If nothing else, a consequence of the structure approach to peer review through calibration exercises did seem to increase reviewer confidence towards the evaluation of impact, where 47/48 (97.9%) indicated that they now felt more confident in assessing research impact; "I certainly had a much better idea of how I should be scoring these things" (*P1OutImp1-POST*). Upon reflection, panellists spoke about the ease with which they assessed impact comparative to what they originally anticipated; "I think there was a lot of uncertainty … .and then when you started doing it, it was surprisingly easy" (*P4OutImp5-POST*). This confidence is in direct comparison to the previously held "nervousness" that panellists described towards the thought of evaluating impact, in the pre-evaluation interviews. If a reason for introducing a structured approach to peer review is to ensure an efficient evaluation process, then by increasing individual evaluator confidence around the task is key. However, in this way, while the calibration exercise and the constant presence (or reassurance) of the guidelines served to increase evaluator confidence around the criterion, they also arguably steered the assessment away from one that is the outcome of expert peer review and this is an important consequence of such a structured approach.

Finally, it must be mentioned that as a result of the experience of evaluating impact, 35/48 (73%), of panellists noted a change in how they viewed, described and characterized impact. Whereas, for the non-impact assessors or, in other words, the panellists who did not have the privilege of the experience of evaluating impact, only 33% noted a change in their concept of impact. For impact panellists, this change was described as a "broadened" *understanding* (*P5OutImp2-POST*), or a "crystallization of beliefs" (*P2OutImp8-POST*). Indeed, some panellists even declared that their previous understanding of the impact criterion, before the evaluation experience, was incorrect; "in one of the cases we looked at, I was completely wrong" (*P3OutImp4-POST*).

## Discussion

This study explores interviewees' perceptions about how pre-evaluation training (calibration exercises) influenced their approach to the assessment of societal impact. In discussing the results presented in this article, and the benefits of pre-evaluation training, two contrasting aspects to consider emerge. First, the value of an efficient assessment process with high IRR in light of a highly undefined criteria and widespread evaluator inexperience and nervousness surrounding the assessment process; and second, allowing the evaluators the temporal and intellectual space necessary to debate the value of different impacts and approaches to assessment made possible through a less structured peer review process. Below, we debate the benefits and pitfalls of both approaches using current understandings of peer review processes such as consensus forming, and achieving high IRR.

As shown above, the evaluators viewed the calibration exercises as useful, especially considering that they were inexperienced when it came to assessing the societal impact criterion. With neither the experience nor a community definition of societal impact (Grant *et al.*, 2010; Bornmann, 2012, Bornmann, 2013, Samuel and Derrick, 2015) with which to ground the evaluation process, the assessment risked a low level of IRR and a long drawn out evaluation process that risked unexplainable outcomes, and therefore the future legitimacy of the impact criterion as a component of research excellence, if an element of structured peer review was not utilized such as pre-evaluation training. In this situation, with all evaluators acting effectively as societal impact novices, pre evaluation training was not as much useful, as being an essential addition to the evaluation process. Whilst training peer review evaluators can have a number of drawbacks, none less important than the narrowing of evaluators' focus with regards to what constitutes societal impact and achieving a high IRR, for the purposes of the world's first formal, ex-post societal impact assessment process, the calibration exercise was an appropriate, and even vital, aspect for the future of societal impact assessment. Indeed, assessments needed to be completed not only within a particular time frame, but also within a certain monetary budget. With no guidance or training about how to assess the, as yet, undefined societal impact criterion, evaluator's perspectives would have been too disparate based on previous studies investigating the variety of views held by evaluators prior to the assessment process taking place (Samuel and Derrick, 2015; Derrick and Samuel, 2016b), resulting in low IRR. Whilst having a low IRR has been argued to improve the quality of discussion and actually increase the validity of the review (Langfeldt, 2001; Bailar, 2011), this argument does not stand in instances when a consensus needs to be reached quickly and within a closed time frame. In these situations, some form of direction, or at least a starting point, is required to "speed debates along", and to bring diverse opinions, perspectives and values onto the same page. For this evaluation exercise, therefore, pragmatic implications ruled out the possibility of the ideal, unstructured peer review process.

However, by training the evaluators to the point that their evaluation reference point were the HEFCE guidelines, or "…the bible", and that they would "follow what we were told", the calibration exercise inevitably removed the intellectual space expert-evaluators needed to deliberate on the concept of societal impact. The results showed that despite appreciating the role of the

pre-evaluation training, evaluators too felt that their appreciation of societal impact was normalized during the evaluation process, reflecting more the definitions of societal impact dictated within the evaluation guidelines, than a product of intellectual debate. Here it must be noted that the goal of the REF2014 was not to resolve what societal impact is, but rather to define it for the purpose of the particular exercise, that is, to develop an understanding of "REF Impact" previously offered in the REF guidelines and reinforced during the calibration exercises via "the bible".

On the other hand of this debate, is the point that the strength of peer review is in allowing the panel debate component of the process. Free and unstructured debate within peer review leads to a group identity surrounding the criteria, or committee culture (Olbrecht and Bornmann, 2010), and this, once established, drives the evaluation. To allow such unstructured assessment approaches may risk a low IRR in the initial stages of the assessment process and therefore would be more time consuming and expensive. Moreover, politically, it would also risk resulting in evaluation outcomes that are not in line with the initial, government and funding agency (HEFCE in this case) intentions. In theory, however, once a committee culture is established, the evaluation process results in a high IRR, despite being low in the initial stages of the process (Olbrecht *et al.*, 2007).

Allowing such discussions in the early stages of implementing an academic acknowledgement of the impact criterion would, arguably, play a large, longer term role in establishing an academic culture and recognition of impact as a component of research excellence. This is especially the case since individual evaluators brought a myriad of opinions and approaches surrounding Impact to be resolved during the evaluation process. This, in turn would better position the academic community, and therefore its peer reviewers (experienced and novice) to accept and approach future impact assessment processes professionally in peer review settings. However, without sufficient temporal and intellectual space in which to develop this committee culture, the evaluation process was driven more by pragmatic issues than a product of intellectual debate of research "peers". Indeed, informed, unstructured, academic peer debate is a necessary precondition for deliberation about how to define, characterize and value new, untested criteria.

Despite this, structured peer review and pre-evaluation training for this REF2014 was appropriate as all evaluators approached the assessment process as novices, bringing no experience with which to ground the assessment process. Pre-evaluation training will remain important for novice evaluators but future evaluation frameworks that include a formal, ex-post societal impact assessment by peer review will involve panels consisting of both novice and experienced societal impact evaluators. As such, a change in panel dynamics will result, as the more experienced evaluators will bring bias and tendencies based on their experience of the last evaluation exercise, dominate the discussions, and influence novice evaluators. If this last evaluation experience was one that was dictated by predetermined guidelines, this will make it more difficult for funding agencies to manage the review outcomes to reflect changes in funding priorities and therefore evaluation criteria and its weighting.

If in future research assessment frameworks incorporating impact, the guidelines about the concept of impact have changed, in the absence of a pre-established committee culture necessary to fuel a community based understanding of the concept of societal impact excellence, pre-evaluation training may be again necessary to normalize all evaluators (experienced and novices) views. This normalization will be necessary to avoid the prior experience and guidelines influencing the direction of assessment, contrary to the objectives of the funding agency. In this way, the REF2014 was an unrealized opportunity for the future of societal impact assessment in that it presented an academic setting (peer review) with which to trial the formal ex-post societal impact assessment and resolve the ongoing debate about the definition and methods of evaluation.

**Limitations of this study**. It is difficult to compare these results directly or indirectly to other studies for a number of reasons. First, there is a recognized lack of research investigating the evaluation of societal impact (Holbrook and Frodeman, 2011) due mainly to the low availability of evaluation frameworks that include a formal, ex-post evaluation of societal impact such as with the REF2014. Second, the qualitative nature of this study and the analysis of its interviews, means that unlike previous studies that only measure the extent of inter-reviewer agreement quantitatively, this study allows for the study of the mechanism of agreement with peer review panels. The qualitative nature of the paper's enquiry also means that the conclusions drawn about the influence of the calibration exercise are based on the opinions of the participants in these exercises (the evaluators), and not by the content of the calibration exercises. However, in light of the lack of official information about the calibration exercises (apart from the fact that they existed), the individual experiences described here provide a unique insight into how a known tool to increase IRR (pre-evaluation training) influenced individual approaches to assessing the Impact criterion. In addition, this study focused on Main Panel A and cannot make any larger inferenced on the influence of the calibration exercises on individual approaches to assessing Impact within other panels (Main Panels B, C and D).

Finally, the political nature of the REF2014 process, as well as other evaluation processes that determine the distribution of public funds, means that the secret and black box nature of peer review, politically and conveniently desirable. This makes the process, therefore, difficult to investigate empirically, and objectively using other qualitative methods such as observations. The use of observations would have, no doubt, allowed for further questions surrounding the change in group dynamics and its effect on IRR possible.

## Conclusions

Societal impact assessment outcomes will only be accepted by the academic community if the review process is perceived to be fair (Tyler, 2006). This necessitates the use of peer review as an assessment tool for societal impact, as it is long considered the gold standard of academic research assessment, despite its flaws. Therefore, the future of societal impact assessment through peer review requires that peer reviewers, typically practising academics, develop a community understanding of excellence for societal impact.

While, for the purposes of the REF2014 and the world's first formal, ex-post societal impact criterion the implementation of pre-evaluation training was essential, questions remain as to their suitability in future peer review processes of societal impact. The academic legitimacy of societal impact as a component of research excellence depends on all practicing researchers, whether societal impact novice or experienced evaluators, reaching a common agreement surrounding the concept of societal impact. Although the purpose of the REF2014 was not to settle the debates surrounding impact assessment, the missed opportunity of allowing unstructured and unguided intellectual academic debate within a peer review process necessitates that pre-evaluation training and therefore a structure impact review process will be necessary for future assessment exercises.

## Note

1 The Impact criterion was measured by a 5-point star rating where 0 = the impact is of little or no reach and significance; 1 = recognized but modest impacts in terms of their reach and significance; 2 = considerable impacts in terms of their reach and

significance; 3 = very considerable impacts in terms of their reach and significance; and 4 = outstanding impacts in terms of their reach and significance (http://www.ref.ac.uk/panels/assessmentcriteriaandleveldefinitions/)

## References


Abdoul H *et al* (2012) Peer review of grant applications: Criteria used and qualitative study of reviewer practices. *PLoSOne*; **7** (9): e46054.

Abrams PA (1991) The predictive ability of peer review of grant proposals: The case of ecology and the United States National Science Foundation. *Social Studies of Science*; **21** (1): 111–132.

Academy., T. B. (2007) Peer review: the challenges for the humanities and social sciences, A British Academy Report. The British Academy.

Bailar J (2011) Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences*; **14** (1): 137–138.

Bence V and Oppenhein C (2005) The evolution of the UK's Research Assessment Exercise: Publications, performance and perceptions. *Journal of Educational Administration and History*; **37** (2): 137–155.

Bornmann L (2012) Measuring the societal impact of research. *EMBO Reports*; **13** (8): 673–676.

Bornmann L (2013) What is the societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society of Information Science and Technology*; **64** (2): 217–233.

Charmez K (2006) *Constructing grounded theory*. Sage: London.

Chubin DE (1994) Grants peer review in theory and practice. *Evaluation Review*; **18** (1): 20–30.

Chubin DE and Hackett EJ (1990) *Peerless Science: Peer review and US Science Policy*. State University of New York Press: Albany, NY.

Cicchetti DV (1991) The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*; **14** (1): 119–186.

Cole S, Cole JR and Rubin L (1978) *Peer Review in the National Science Foundation: Phase One of a Study*. The National Academy of Sciences: Washington DC.

Danziger S, Levav J and Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences of the United States of America*; **108** (17): 6889–6892.

Delcomyn F (1991) Peer review: Explicit criteria and training can help. *Behavioral and Brain Sciences*; **14** (1): 144.

Demicheli V and Di Pietrantonj C (2007) Peer review for improving the quality of grant applications. *Cochrane Database Syst Rev*; **2** (2).

Derrick GE (forthcoming) *The Evaluators Eye: Impact assessment and academic peer review*. Palgrave Macmillan: London.

Derrick GE and Samuel GN (2016a) "All this grassroots, real life knowledge": Assessing the value of including non-academic evaluators in societal impact assessment. 21st International Conference on Science and Technology Indicators, 2016a Valencia, Spain.

Derrick GE and Samuel GN (2016b) The evaluation scale: Exploring decisions about societal impact in peer review panels. *Minerva*; **54** (1): 75–97.

Eckberg DL (1991) When nonreliability of reviews indicates solid science. *Behavioural and Brain Sciences*; **14** (1): 145–146.

Fogelholm M, Leppinen S, Auvinen A, Raitanen J, Nuutinen A and Väänänen K (2012) Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of Clinical Epidemiology*; **65** (1): 47–52.

Frank C and Nason E (2009) Health research: Measuring the social, health and economic benefits. *Canadian Medical Association Journal*; **180** (5): 528–534.

Glaser B and Strauss A (1967) *The discovery of grounded theory. Strategies for Qualitative Research*. Weidenfeld and Nicolson: London.

Gordon R and Poulin BJ (2009) Cost of the NSERC science grant peer review system exceeds the cost of giving every qualified researcher a baseline grant. *Accountability in Research*; **16** (1): 13–40.

Grant J, Brutscher P-B, Kirk S, Butler L and Wooding S (2010) Capturing Research Impacts: A Review of International Practice. RAND Europe: Cambridge, UK.

HEFCE (2011) Assessment framework and guidance on submissions. Research Excellence Framework 2014. London, UK.

Hemlin S and Rasmussen SB (2006) The shift in academic quality control. *Science, Technology, & Human Values*; **31** (2): 173–198.

Hodgson CM (1995) Evaluation of cardiovascular grant-in-aid applications by peer review: influence of internal and external reviewers and committees. *Canadian Journal of Cardiology*; **11** (10): 864–868.

Holbrook JB and Frodeman R (2011) Peer review and the ex ante assessment of societal impacts. *Research Evaluation*; **20** (3): 239–246.

Holbrook JB and Hrotic S (2013) Blue skies, impacts, and peer review. *A Journal on Research Policy & Evaluation*; **1** (1).

Holliday C and Robotin M (2010) The Delphi process: A solution for reviewing novel grant applications. *International Journal of General Medicine*; **3**, 225.

Huutoniemi K (2012) Communicating and compromising on disciplinary expertise in the peer review of research proposals. *Social Studies of Science*; **42** (6): 897–921.

Kravitz RL, Franks P, Feldman MD, Gerrity M, Byrne C and Tierney WM (2010) Editorial peer reviewers' recommendations at a general medical journal: Are they reliable and do editors care? *PLoS ONE*; **5** (4): e10072.

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

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** (6): 820–841.

Langfeldt L (2004) Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation*; **13** (1): 51–62.

Marsh HW, Jayasinghe UW and Bond NW (2008) Improving the peer-review process for grant applications: Reliability, validity, bias and generalizability. *American Psychologist*; **63** (3): 160–168.

Martin BR (2011) The research excellence framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation*; **20** (3): 247–254.

Olbrecht M and Bornmann L (2010) Panel peer review of grant applications: What do we know from research in social psychology on judgement and decision making in groups? *Research Evaluation*; **19** (4): 293–304.

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** (2): 79–91.

Samuel GN and Derrick GE (2015) Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*; **24** (3): 229–241.

Sattler DN, Mcknight PE and Mathis R (2015) Grant peer review: Improving inter-rater reliability with training. *PLoSOne*; **10** (6): e0130450.

Tan E, Ghertner R, Stengel PJ, Coles M and Garibaldi VE (2015) Validating grant-making processes: Construct validity of the 2013 senior corps RSVP grant review. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*; **27** (3): 1403–1424.

Thornley R, Spence MW, Taylor M and Magnan J (2002) New decision tool to evaluate award selection process. *Journal of Research Administration*; **33** (2/3): 49–58.

Tyler TR (2006) Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology*; **57**, 375–400.

Vener K, Feuer E and Gorelic L (1993) A statistical model validating triage for the peer review process: Keeping the competitive applications in the review pipeline. *The FASEB Journal*; **7** (14): 1312–1319.

Wu H, Ismail S, Guthrie S and Wooding S (2011) *Alternatives to Peer Review in Research Project Funding*. RAND Europe: Cambridge, UK.

WWW.REF.AC.UK. (2014) *Consistency across UOAs: REF 2014* [Online], accessed 20 September 2016].


## Data availability

The datasets generated during and/or analysed during the current study are not publicly available due to confidentiality reasons, but are available in a codified form from the corresponding author on reasonable request.

## Acknowledgements

This research was funded by the UK Economic and Social Research Council (ESRC) Future Research Leaders Programme (ES/K008897/2).

## Additional information

**Competing interests:** The authors declare that there are no competing interests.



**How to cite this article**: Derrick G and Samuel G (2017) The future of societal impact assessment using peer review: pre-evaluation training, consensus building and inter-reviewer reliability. *Palgrave Communications*. 3:17040 doi: 10.1057/palcomms.2017.40.