Assessing impact submissions for REF 2014: An evaluation

Catriona Manville, Susan Guthrie, Marie-Louise Henham, Bryn Garrod, Sonia Sousa, Anne Kirtley, Sophie Castle-Clarke and Tom Ling

RAND EUROPE

For more information on this publication, visit www.rand.org/t/rr1032

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

RAND® is a registered trademark.

© Copyright HEFCE 2015

RAND Europe is a not-for-profit research institute whose mission is to help improve policy and decisionmaking through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Support RAND
Make a tax-deductible charitable contribution at
www.rand.org/giving/contribute

www.rand.org
www.randeurope.org

Preface

The four UK higher education funding bodies[1] asked RAND Europe to review the assessment process for the impact element of the Research Excellence Framework 2014 in the UK, in order to assess the process and understand how it could be further improved.

This report provides headlines from our study (overarching summary) and is supported by an in-depth analysis of the data gathered relating to each part of the assessment process. It is intended for those responsible for the REF and, more broadly, for those in the higher education sector. It may also be of interest to others working in the evaluation of research impact.

RAND Europe is an independent not-for-profit policy research organisation that aims to improve policy and decisionmaking in the public interest, through research and analysis. RAND Europe's clients include European governments, institutions, non-governmental organisations and firms with a need for rigorous, independent, multidisciplinary analysis. This report has been peer-reviewed in accordance with RAND's quality assurance standards.

For more information about RAND Europe or this document, please contact:

Catriona Manville

RAND Europe
Westbrook Centre
Milton Road
Cambridge CB4 1YG
United Kingdom

Tel. +44 (1223) 353 329
manville@rand.org

[1] The Higher Education Funding Council for England (HEFCE), the Higher Education Funding Council for Wales (HEFCW), the Scottish Funding Council (SFC), and the Department for Employment and Learning, Northern Ireland.

Abstract

The Research Excellence Framework (REF) is a new system for assessing the quality of research in UK higher education institutions (HEIs). For the first time, part of the assessment included the wider impact of research. RAND Europe was commissioned to evaluate the assessment process of the impact element of REF submissions, and to explore the strengths and weaknesses of the assessment process in delivering reliable, robust and fair outcomes, identify any unforeseen issues emerging during the assessment process and outline broad implications for the assessment process of impact in future REF exercises. This report provides a summary of our key findings, complemented by a detailed analysis of the process using the following methodologies: focus groups with panel members, users and impact assessors; interviews with panellists; and a survey of all panellists.

This evaluation complements another piece of research that RAND Europe undertook for the UK higher education funding councils to evaluate the submission process HEIs underwent to prepare for the impact element of REF 2014 (Manville et al. 2015a, Manville et al. 2015b).

Table of contents

Overarching summary xi

S.1. Overview of the assessment process xi
S.2. Key messages xii
S.3. Conclusions xvi

Acknowledgements xvii
Abbreviations xix

Chapter 1: Introduction 1

1.1. Background to the REF 2014 assessment process 1
1.2. Evaluation methodology 4
1.3. Caveats and limitations 8
1.4. Overview of this report 8

Chapter 2: Panel structure and membership 11

2.1. Structure of the panels 11
2.2. Recruitment of panellists 13

Chapter 3: Allocation of impact case studies and impact templates 15

3.1. The process of allocation 15
3.2. Expertise to assess impact submissions 21

Chapter 4: Calibration exercise 23

4.1. The process of calibration 23
4.2. Opinions varied regarding the selection of case studies for the calibration sample 23
4.3. Panellists' perspectives on the calibration process 24
4.4. Calibration raised awareness of different perspectives 25
4.5. Challenges with the calibration process 25

Chapter 5: Review of impact case studies and individual scoring 27

5.1. Score levels 27
5.2. Rules and guidance 27

Chapter 6: Review of impact templates 37

6.1. Importance of the quality of writing in the assessment of the impact template 39

6.2. Other issues reported with assessing the impact template 39
6.3. Options for the future 39
6.4. Relationship between impact case studies and template scores 40

Chapter 7: Scoring agreement and moderation 43

7.1. Use of clusters to assess impact submissions 43
7.2. Resolution of disagreement through discussion 44
7.3. Scores allocated from the process 45
7.4. The effect of small submissions 46

Chapter 8: Evidence checking and audit 47

8.1. Types of audit queries raised 49
8.2. Areas where secretariat provided advice 49

Chapter 9: Overview of the perceived benefits and challenges associated with the assessment process 53

9.1. Benefits 53
9.2. Challenges 55

Chapter 10: Burden associated with the assessment of the impact element of REF 2014 59

List of references 61

Appendix A: List of panels and units of assessment for REF 2014 63

Appendix B: Semi-structured protocol for focus group interviews 65

Appendix C: NVivo code book for coding interviewee responses from the focus groups 67

Appendix D: Survey protocols for those involved in the assessment of the impact element of REF 2014 69

D.1. Academic sub-panel survey 69
D.2. Impact assessor sub-panel survey 74
D.3. Adviser/secretariat sub-panel survey 79
D.4. Main panel member survey 83
D.5. Additional main panel questions (for survey respondents who are already answering another section of the survey) 87

Appendix E: Protocol for individual interviews with panellists involved in the assessment of impact 89

Table of figures and tables

Figure S.1: Overview of the impact assessment process xi
Figure 1.1: Overview of the impact assessment process 1
Figure 1.2: Outline of impact case study and impact template documents 3
Figure 1.3: Project schema 5
Figure 2.1: Level of knowledge of impact assessment prior to involvement with REF 2014 14
Figure 3.1: Number of panellists reviewing a given case study 16
Figure 3.2: Percentage of case studies by number of panel members assessing each case study. 17
Figure 3.3: Number of panel members assessing each impact template 18
Figure 3.4: Percentage of impact templates by number of panellists assessing each impact template 19
Figure 3.5: The distribution of impact case studies reviewed by role and main panel. 20
Figure 3.6: The distribution of impact documents (case studies and templates) reviewed by role and main panel 21
Figure 3.7: Sub-panel members' and impact assessors' assessment of whether they had the expertise to assess case studies allocated 22
Figure 4.1: The extent to which sub-panel members and impact assessors felt that the impact calibration exercise allowed impact case studies to be assessed fairly and reliably (n=596) 24
Figure 5.1: Extent to which survey respondents felt the rules and guidance enabled them to determine the eligibility requirements for impact case studies. 28
Figure 5.2: Extent to which survey respondents felt that the rules and guidance enabled them to assess impact case studies fairly and reliably 28
Figure 5.3: Extent to which survey respondents felt the format of the case study template enabled them to assess impact case studies fairly and reliably. 29
Figure 5.4: Extent to which survey respondents felt the criteria of reach (above) and significance (below) enabled them to assess impact case studies fairly and reliably 31
Figure 5.5: Extent to which survey respondents felt the rules and guidance enabled them to determine the quality of underpinning research for impact case studies 32
Figure 5.6: Extent to which survey respondents felt the rules and guidance enabled them to determine the contribution of research to impact for impact case studies 34
Figure 6.1: Views of different respondent groups on the format of the impact template..... 37
Figure 6.2: Percentage of 4-star case studies and templates, by main panel 38
Figure 6.3: Correlation between the scores for a submission's impact case studies and template, by main panel and submission size 41
Figure 7.1: Perceptions on the process of agreeing scoring and the management of disagreements 44
Figure 7.2: Views by type of respondent on satisfaction with the efficiency of the process of reaching agreement and management of disagreement 44
Figure 7.3: Percentage of impact submissions with each level by main panel 46

Figure 7.4: Percentage of submissions with each level in overall sub-profiles 46
Figure 8.1: Distribution of audit queries across the 36 sub-panels 48
Figure 8.2: Distribution of reasons for audit queries by main panel 49
Figure 8.3: Distribution of audit queries by level awarded 50
Figure 8.4: Areas secretariat and advisers thought sub-panels found most challenging 51

Table 1.1: Types of panellists involved in the impact element of assessment 2
Table 1.2: Level definitions for assessing impact as part of REF 2014 3
Table 1.3: Response rates to the surveys 7
Table 1.4: Overview of caveats and limitations of the study 9
Table 2.1: Types of panellists by role 13
Table 6.1: Classification used for size of submissions 40
Table 7.1: Size of submission by main panel 46
Table 8.1: Breakdown of audit queries across panels 47
Table 8.2: Type of audit queries 49
Table 9.1: Top five responses received when asked to provide up to three benefits of engagement with the impact element of the REF process 53
Table 9.2: Top five responses received when asked to provide up to three challenges of engagement with the impact element of the REF process 55
Table 10.1 Time spent on the REF process, in hours, by sub-group 59

Overarching summary

In 2014 the impact of academic research outside of academia was for the first time assessed as part of the Research Excellence Framework (REF) in the UK, determining future funding for institutional research. Impact was assessed and graded by academics and research users. Guidance and support was provided by the higher education funding bodies. In total 6,975 impact case studies and 1,911 impact templates were assessed and graded. This report explores the strengths and weaknesses of the assessment process and identifies unforeseen issues and their management. It is based on focus groups, interviews, survey and documentary analysis. It details the process that took place, examines the perceptions of academics and research users who were involved in the process, arrives at overall judgements about the conduct of the impact element of REF 2014, and discusses the implications for future assessments.

S.1. Overview of the assessment process

The impact assessment process undertaken in REF 2014 comprised six main stages, as shown below in Figure S.1.

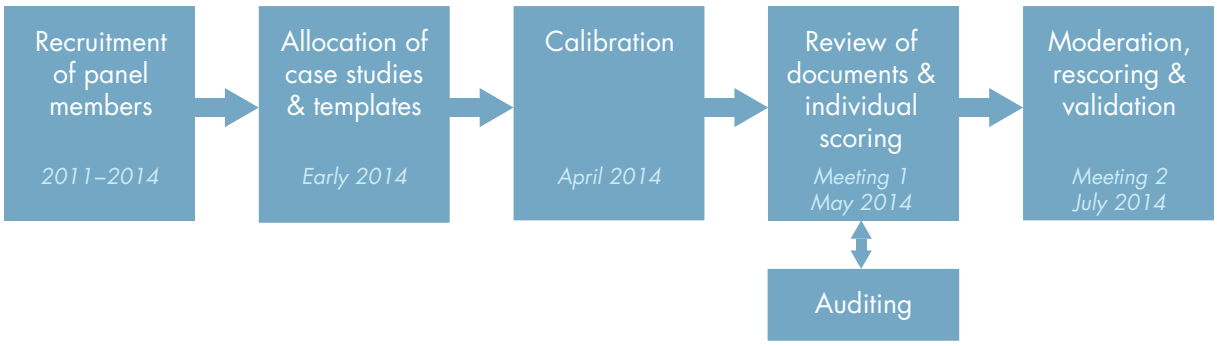
The assessment of impact was made on the basis of two different types of document submitted by higher education institutions (HEIs): impact case studies (REF3b) and impact templates (REF3a). Case studies provide examples where research has led to impact, and templates set out the wider strategy around facilitating the translation of research into impact.

Submissions are made at the level of the sub-panel (i.e. each HEI will make multiple submissions, one per unit of assessment (UOA) that is relevant to their research). Each submission consists of one impact template, plus a number of case studies (where the number depends on how many staff are submitted by the institution in that UOA, approximately one case study per 10 staff).

In REF 2014, the case studies and templates were allocated for assessment to sub-panel members and impact assessors by the sub-panels, who were asked to review and score them each individually. They were provided with some guidance to do this, and there was a calibration session early in the process to allow them to calibrate their scoring to the rest of the sub-panel. Each case study and template was reviewed by more than one individual and exact practice on the allocation and scoring processes differed between main panels and sub-panels.

Once the case studies had been scored individually, there was a process of moderation of those scores. Initially there was some moderation at the sub-panel level (though exact processes for this varied), and then at the main panel level. There was also some limited validation between main panels.

Figure S.1: Overview of the impact assessment process



An additional step alongside moderation was audit, a process by which panel members could request checks on the case studies and templates to check their eligibility against various criteria, or to access the sources of evidence referred to in the case studies. Audit was not conducted systematically across all case studies, but rather on the basis of sub-panel requests. Some 74 per cent of institutions had at least 5 per cent of their case studies audited.

Finally, scores for case studies and templates in each submission were combined to produce an overall impact 'profile'. The profile indicates the proportion of the submission that is at each level from unclassified, to 1 star, 2 star, 3 star or 4 star (the highest rating). Ultimately these profiles are combined with the output and environment profiles to produce an overall profile for each HEI submitting to each UOA. Funding will ultimately be allocated by the funding bodies on the basis of the REF outcomes, with 3 star or above being the requirement for funding allocation.[2]

S.2. Key messages

S.2.1. By a large majority, panellists felt the process enabled them to assess impact in a fair, reliable and robust way

> *I've been struck all the way through by the emphasis on fairness in the efforts being made on the structure of the exercise to ensure that there was really a fair and proper assessment.*[3]

> *Above all, there was extremely intensive and carefully considered discussion of the application of the criteria as a whole in practice, reviewing and revisiting until we were clear collectively that the process was absolutely fair and reliable.*

Panellists felt there was reasonable consistency in the scores given by individuals assessing case studies. This agreement was within main panels and between the different panellist roles. It was recognised that this element of the assessment was new, and although there were areas for improvement for future exercises there was much that had gone well. As the first attempt at assessing impact on a large scale it was seen to be successful, though panellists recognised there were areas for improvement. One panellist stated: 'There is much to commend [it] although there are improvements to be made and much to be learnt'. Consistency was supported by various steps in the process described in the following paragraphs.

Many panellists commented positively on **calibration** as 'creating a shared understanding of impact', and allowing the case studies to be assessed fairly and reliably. Where a minority of impact assessors were not involved in calibration, there was concern that their understanding was not aligned with those who had been through the process. Following calibration, there was **continued discussion** that was viewed by participants to be inclusive and fair. If a consensus was not reached it was escalated to the whole sub-panel, or from the sub to the main panel for resolution. In sub-panels with fewer submissions (such as Agriculture, veterinary and food science (UOA6)) all case studies were discussed by the full sub-panel and a score agreed. Larger sub-panels only discussed cases where a consensus could not be reached among the initial reviewers. The scores were then **moderated** within both the sub-panel and the main panel to ensure case studies were graded equivalently. On occasions where scoring was thought to be out of line, sub-panels were asked to review their decisions by the sub-panel chair or a member of the relevant main panel. The sub-panel chairs sat on the main panel, further facilitating communication and reinforcing comparability across sub-panels. In addition, the main panels had a number of research user members, as well as international advisors who floated between sub-panels to support sub-panel members and impact assessors in their deliberations. In particular panellists from areas of Main Panel A noted how useful their international advisors were in supporting them to be fair to the impacts they were reviewing.

There was less confidence from panellists about the consistency between main panels. This was most often seen by sub-panel members who were not involved in the process across the main panels; those engaged at the main panel level were considerably more satisfied with the process. This could imply that the issue is one of transparency and the information the sub-panel level received. On the other hand, confidence in the process was enhanced by how case study **allocation** was managed. In the majority of sub-panels each panellist reviewed case studies with a variety of other panellists, allowing learning to be shared. Overall, panellists felt

---

2 For further information on the process, see the following chapters, and overview reports published by each main panel (REF 2015).

3 Quotations used throughout this report are taken from the focus groups and survey responses.

that the **rules and guidance** helped them to fairly and reliably assess case studies.

S.2.2. There was variation in the way the process was conducted

> *Plenary discussions of scoring disparities between individuals were possible for small sub-panels but impossible for the largest ones.*

Due to the devolved governance of the process and the understanding that one model for assessment would not suit all disciplines there was variation in the process across sub-panels. This occurred at multiple points in the process – from number and combination of reviewers, to the application of eligibility criteria and the use of corroborating evidence. The flexibility this provided reflected disciplinary differences but may have led to unwarranted variations.

For example, individuals in our focus groups reported differing interpretation of guidelines when identifying the thresholds for the eligibility of a case study, such as the 2-star quality of underpinning research, and when assessing the link between research and impact. We note that this problem of identifying thresholds is not unique to assessing research impact.

Variation in interpretation was also seen with the use of corroborating evidence, and confirming the statements made in impact case studies. It was unclear to panellists the extent to which they should use prior knowledge and the point at which an audit query needed to be raised to check details. This led to differing practices between panellists.

Panellists used their judgement to select the appropriate level of validation for a case study depending on content, and carried out differing levels of checking. This is understandable given the different levels of complexity and different kinds of claims being made. However, participants also reported differing interpretations of how checking of evidence should be conducted, and what was allowed. Consequently, different practices occurred, within and across sub-panels. In general there was a desire from many panellists to have unmediated access to evidence, although the additional burden this would entail was acknowledged. Overall, there was a view from panellists that the process of corroboration was fair and added rigour but at the same time there was an anxiety that there may have been an unknown number of case studies where submissions were over-claiming which may have 'slipped through the net'.

Panellists in Main Panel D stressed the need for a level of expertise in order to fully assess impact and evidence claims. However it was also recognised that this might introduce unfairness if individuals on the panel had varying levels of knowledge. Examples were given in areas of Main Panels A and C where those who knew about specific case studies 'were told [by sub-panel chairs or other panellists] they weren't allowed to use that information'. This was frustrating, as panellists believed it was this very knowledge that partly justified why they had been selected for the panel.

As described in the following section, all main panels used ½ stars to create more granularity. In addition, areas of Main Panel A developed a framework of up to 8 stars when structuring their discussions about scores. The rationale behind this was that there were different levels within each star, and it ensured that 4-star impacts were not downgraded due to comparisons with exceptional examples (a problem discussed below in Section 1.2.3). In Main Panel A, when awarding the scores, case studies scoring 4-8 were all graded as 4-star.

Our study did not review the case studies to examine the scores awarded, so we cannot assess whether variation in interpretation produced variation in scores. However, we can report that the process aimed to limit the impact of variation through calibration at all levels; through creating a shared understanding at a sub-panel and main panel level; and moderating scores at a main panel level and between them. For example, it is important to note that when a sample of Main Panel A impacts were assessed by the other three main panels they also scored highly.

S.2.3. Panellists felt they were able to differentiate between submissions in more detail than the scoring process allowed them to express

> *One challenge we had, which makes us appear generous but we weren't, was that sometimes we found an exceptional 4 that was really off the scale, which made it hard to see that others could still be a 4.*

Several panellists said that there were two different parts to calibration: consistency and level. It was felt that the former was easier than the latter. The difficulty in awarding levels, based on a four-point scale, is reinforced by the fact that all panels used a more granular scale for awarding levels. Sub-panels across all four main panels used ½ stars in their scoring. The difference described however, is not reflected in the scores.

The scores for the impact element are very high, with nearly 40 per cent of all impact sub-profiles rated as 4 star and over 80 per cent considered 3 or 4 star.[4] This apparent 'bunching' of scores suggests consistently high-quality case studies and impact templates when graded against the level criteria.

For the highest score there is a quality range which goes from just achieving the threshold up to a theoretically infinitely high quality. To address this, some areas of Main Panel A, as noted above, framed their discussions using a framework of up to '8-stars', awarding case studies thought to be 4-star or above as 4 star, i.e. so that the very best should not determine the threshold between a 3 and 4-star score. The use of this conceptual framework was described by one panellist as helping them 'to get the calibration right, rather than setting the very best ones as a 4, which would have made 3 a severe judgement. This encouraged us to think that there were lots of 4s'.

Panellists from Main Panels B and D stressed an awareness of the funding implications of the boundary between 2 and 3-star.

Panellists stated that they found it difficult to verify the link between the underpinning research and the impact and sometimes the links were more tenuous than they would have liked. One suggestion from Main Panel A was that attribution could be used as an additional criterion to differentiate where there were multiple submissions claiming the same impact.

S.2.4. There were particular challenges in assessing the impact templates

> *It felt more aspirational than anything based on reality.*

Panellists felt that the impact templates were more difficult to assess than the case studies. In particular, there was a concern that the impact template was not an effective way to assess how institutions support the translation of research to impact, and many questioned its value in the submission.

The lack of requirement to evidence the claims made meant that the quality of writing had a large effect, and this made it difficult to discriminate between templates. In addition, it was felt that it was too easy to retrofit a strategy that may not have been in place throughout the period of assessment. On occasion panellists were uncomfortable with their judgement as they felt that the template encouraged plans for the future that could be 'a figment of the [template author's] imagination'. However, there were some who felt the impact template was a 'vital document' to HEIs and to assessment, as it helped panellists to better understand what institutions were doing to support their staff in encouraging and facilitating impact.

There was a range of suggestions for how to improve this element in future assessments. Some suggested it should be completely removed as the environment could be evidenced by the case study examples whether or not the HEI had a strategy. Others suggested combining the impact and the environment template, as the impact environment is a subset of the overall research environment. However, there were those who felt the environment within which impact is fostered should continue to have a separate submission. In particular there were examples of where impact assessors found it helpful to review the impact template to understand the context within which impact had occurred and this might be more difficult if it was subsumed within the environment template (REF5). Alternatively, to reduce the influence of the quality of writing on the submission, some suggested greater use of narrowly factual information. Whatever the future for the impact template, there is a need for guidance to HEIs to ensure greater clarity.

S.2.5. Through involvement, research users built useful networks, but the burden of involvement was a significant challenge

> *[One of the benefits was the] opportunity to make new contacts and strengthen [my] existing network in academia and industry.*

The most cited benefit among impact assessors (mentioned by 44 out of 74 respondents) was networking and collaborative working. This included the enjoyment of engaging with academics (and the academic sector more broadly) as well as networking opportunities with the academic sector for further professional activities. This is closely linked to the second most-cited benefit (listed by 37 out of 74 respondents), which was learning about and gaining an understanding of academic research taking place across the UK. Respondents noted they had gained an awareness of different types of research, had been exposed to new research ideas and

[4] A sub-profile was formed of the impact element for each submission, taking into account the levels awarded to each impact document.

had a better overview of academic disciplines of relevance to their work, among other things.

These benefits, however, came at a cost for those involved. Burden was by far the most frequently mentioned challenge of involvement in the assessment, being mentioned by 45 out of 74 respondents. Based on the quantitative data collected through the survey, the median amount of time spent on the process by impact assessors was 11 days. There was, moreover, variation in the commitment made by individual impact assessors – the interquartile range of the estimates of time spent by impact assessors from the survey spanned from 7 to 15 days.[5] Despite seeing influencing and networking as benefits, the burden was particularly significant for those not employed by an HEI, since time away from work may have been given to cover the time attending meetings but it is likely that the assessment of impact documents was completed outside their normal working hours.

S.2.6. Engagement by academics in the process offered benefits for academic careers and institutions

> *It will be useful for me in preparing future impact case studies for the next REF.*

Academic sub-panel members reported direct benefits for careers and institutions resulting from participation. By far the most frequently cited benefit (mentioned by 174 out of 396 academic sub-panel respondents) related to their improved understanding of the assessment process. Specific examples included: learning how the process of assessing impact worked in practice; understanding how other people judge impact; developing a deeper understanding of the rules and how they should be applied; gaining reassurance of the robustness of the process; and learning how to approach the exercise for future assessments – including how to present good case studies and templates. This reflects the increasing importance of impact assessment in academia more widely.

Other frequently mentioned benefits included learning about the range of academic research taking place across HEIs (mentioned by 99 of 396 respondents) and learning about the diverse range of impacts emanating from research (mentioned by 77 of 396 respondents). The benefits are both for academics' own careers and for their institutions more widely, in terms of building an understanding of impact and its assessment, and an inside perspective on what other academics and institutions are doing in this area – effectively giving participants a 'competitive advantage'.

In terms of challenges, the burden of engagement is the most frequently listed (mentioned by 86 of 396 respondents), but several other response categories are mentioned almost as frequently, notably the challenges of using evidence (mentioned by 79 of 396 respondents) and of comparing and fairly assessing different types of impact (76 of 196 respondents). Unlike impact assessors, the issue of burden, though still important, is not so dominant for this group, which also reflects on the challenges of the process itself rather than challenges for themselves in engaging with the process. This re-emphasises the more direct engagement that this group necessarily has with the process and its outcomes, which directly affect their sector. While understanding the process and gaining an 'inside' perspective is beneficial for this group, the challenges and limitations of the process have direct impacts for this group and as such are of significant personal concern to them.

S.2.7. Bringing together different perspectives of academics and research users was seen to be successful and valuable

> *It was interesting to hear remarks from the panel about how the REF had tempered the behaviour of the academics; it also tempered our behaviour as well […] and it was a stroke of genius to get people together to get that consensus generated.*

A significant feature of the impact element of REF 2014 was that it included and synthesised the views of both academics and research users in the assessment process to a greater extent than other elements of the assessment process. There was a possibility that since these two groups operate in related but different worlds that they might struggle to form a shared understanding and build sufficiently trusting relationships to work well together in the assessment process. We raised these issues with both academics and research users in separate focus groups and also conducted one-to-one interviews with representatives of each group. Our survey data also helped inform our understanding of how well this worked.

Both academics and research users in the focus groups and the interviewees highlighted the value of including

[5] The interquartile range is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles.

both perspectives.[6,7] It was widely agreed that the two perspectives moderated each other and added to panellists' confidence in the process. There was a widespread sense that the exercise was inclusive and that views were listened to equally. This is echoed in our survey which shows that there was a high percentage of satisfied and very satisfied respondents, with the highest levels of satisfaction regarding all members being listened to equally.

S.3. Conclusions

Confidence in outcomes, and in the processes used in the research impact assessment, were relatively high following completion of REF 2014 among those most involved. Based on the evidence presented in this report, this confidence was due in large part to: the formal guidance and processes used; an academic-led and inclusive approach to leadership; and the high level of commitment shown by academics and research users. The introduction of an impact element in REF 2014 might have been expected to generate concerns because of the relative novelty of the approach and because of obvious difficulties in measurement, but in general it has succeeded.

There are a number of incremental improvements which could be made to the process. These include: access for panellists to underpinning research and corroborating evidence; potential use of closed questions detailing eligibility; the stringency with which to apply the rules around the format; and clearer guidance to HEIs and panellists (on how to present eligibility information within the case studies, and the requirement, or not, for impacts to demonstrate both reach and significance).

In addition, we have flagged some areas for further consultation. These include: revisiting how to manage the risk of random variations having excessive effects in small units of assessment; how to avoid the risk of unsubstantiated and false claims being made; how to clarify the processes for assessing different kinds of impact; and how best to capture the information pertaining to the wider HEI environment for nurturing and developing impact.

6 Although a handful of academics worried that users were not suitably equipped to assess impact.

7 In addition, main and sub-panel members from areas of Main Panel A highlighted the value of their international advisers, and the perspective they brought to the process.

Acknowledgements

RAND Europe would like to thank Kim Hackett and Steven Hill at HEFCE for their guidance throughout this project, as well as the REF Impact Steering Group (Michelle Coupland, Chris Darby, Alice Frost, Roger Kain, Ian Leslie, John Rogers, Sue Smart, Malcolm Skingle, Robert Slater and Jeremy Watson). We are also very grateful to all the individuals who provided us with their impressions of the process, through engagement in focus groups, one-to-one interviews and in responses to our survey.

The authors are grateful for the interest and assistance of all those who contributed to this project, in particular Michael Frearson, Molly Morgan Jones, Jessica Plumridge, Claire O'Brien, and Calum MacLure. In particular, we would like to acknowledge the valuable input provided by our Quality Assurance reviewers, Steven Wooding and Linda Butler.

Abbreviations

DEL	Department for Employment and Learning, Northern Ireland
FTE	Full-Time Equivalent
HE	Higher Education
HEFCE	Higher Education Funding Council for England
HEFCW	Higher Education Funding Council for Wales
HEI	Higher Education Institution
HESA	Higher Education Statistical Agency
ITT	Invitation To Tender
QR	Quality-Related
QALY	Quality Adjusted Life Years
RAE	Research Assessment Exercise
REF	Research Excellence Framework
REF3a	REF 2014 section for the impact template document
REF3b	REF 2014 section for the impact case study document
R&D	Research and Development
SFC	Scottish Funding Council
UOA	Unit of Assessment

Chapter 1 Introduction

> *There is much to commend [it] although there are improvements to be made and much to be learnt.*

This report, commissioned by the UK higher education funding bodies in August 2014, presents our evaluation of the assessment process for the impact element of REF 2014.[8] It is aimed at those who wish to better understand the REF 2014 process and those thinking about subsequent assessments.

1.1. Background to the REF 2014 assessment process

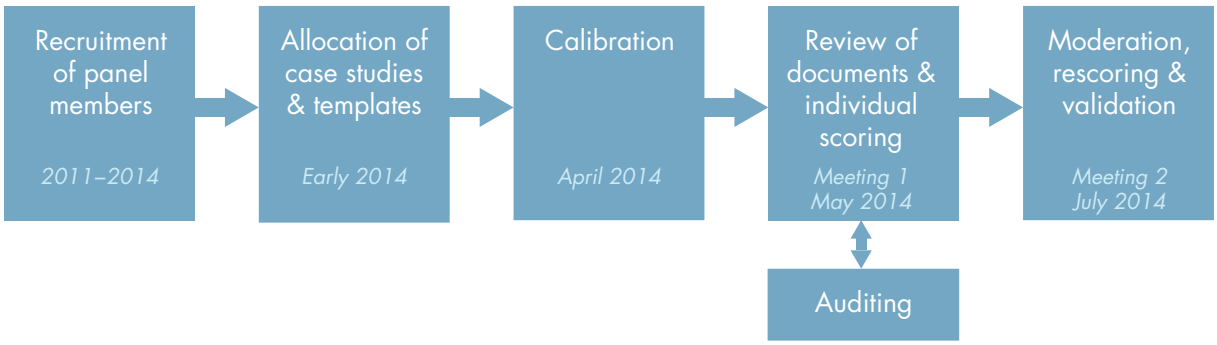
The Research Excellence Framework (REF) is a new system for assessing the quality of research in UK higher education institutions (HEIs).[9] It replaces the Research Assessment Exercise (RAE), which occurred on a (near) quinquennial basis from 1986 to 2008. The RAE assessed research excellence in HEIs by the quality of research outputs and other measures of the research environment, including research students, income and evidence of esteem (RAE 2005). The REF also assesses research excellence, but on the basis of three main criteria: the *quality* of research outputs, the *wider impact* of research and the *sustainability* and *vitality* of the research environment (REF 2011).

The outcomes of REF 2014 were published in December 2014. It was undertaken by the four UK higher education (HE) funding bodies, but is being managed by the REF team based at the Higher Education Funding Council for England (HEFCE) and overseen by the REF Steering Group, consisting of representatives of the four funding bodies.

The REF has three main purposes:

- The HE funding bodies intend to use the assessment outcomes to inform the selective allocation of their research funding to HEIs, with effect from the academic year 2015–16.
- The assessment provides accountability for public investment in research and produces evidence of the benefits of this investment.
- The assessment outcomes provide benchmarking information and establish reputational yardsticks.

Figure 1.1: Overview of the impact assessment process



[8] These are the Higher Education Funding Council for England (HEFCE), the Higher Education Funding Council for Wales (HEFCW), the Scottish Funding Council (SFC), and the Department for Employment and Learning, Northern Ireland.

[9] A higher education institution is a university or higher education college. HEIs across the UK can choose whether to submit to the REF, which leads to funding allocation. Submissions are organised by subject areas, defined as Units of Assessment (UOAs).

The impact assessment process consisted of six main stages as shown above in Figure 1.1. It should be noted that in parallel there was an assessment of research output (looking at academic publications) and the research environment at institutions. The process as set out here and as evaluated in our work refers only to the impact element of the assessment.[10] However, many of the individuals involved in the assessment of outputs and environment were also involved in the assessment of impact, and the processes will have overlapped to some extent (this is clarified as required throughout).

The first step in the assessment was the recruitment of panel members. A breakdown of panel membership is set out below in Table 1.1. The assessment was based on submissions at a disciplinary level, corresponding to 36 units of assessment or UOAs. These UOAs each have their own sub-panel, each of which sits under one of four main panels, A to D.[11] Each sub-panel is made up of impact assessors, sub-panel members, a secretary, an advisor, a deputy chair, and a chair. Observers may also attend some sub-panel meetings.

The assessment of impact was made on the basis of

Table 1.1: Types of panellists involved in the impact element of assessment[12]

Role	Description of the group	Responsibilities	Number across all panels
Main panel members	Senior academic staff with expertise relevant to the main panel	The main panel consists primarily of the chairs (and some deputy chairs) for each sub-panel, plus the chair of the main panel (who does not sit on any individual sub-panel). They are responsible for developing guidance and moderating results across each main panel.	42
User members	Research users representing relevant organisations in the UK	Invited to join main panel meetings to provide a research user's perspective	15
Secretariat	Typically administrative staff from HEIs with some experience of the evaluation of research and research impact	Seconded to support the administration of the REF process, including impact assessment. Each secretary is responsible for several (2-3) sub-panels within one main panel	15
Advisers	Typically administrative staff from HEIs with a comprehensive understanding of the assessment of research impact and REF processes	Seconded to support the administration of the REF process including impact assessment. Each advisor sits on one main panel and several (2-3) sub-panels falling within its remit. They are responsible for providing advice to sub-panels on the REF rules and guidance and have some responsibility for audit	13
International advisers	Academics and research users from outside the UK	Invited to join main panel meetings to provide an international perspective	21
Observers	Other interested parties. Typically representatives of funding bodies and learned societies in the UK	Invited to join main panel and/or sub-panel meetings to observe the process	10
Sub-panel members	Largely academics with expertise relevant to the sub-panel. However, the group also includes a limited number of research users	Developing approach at the sub-panel level to assessment. Assessment of case studies and templates. Also involved in output and environment assessments	735
Impact assessors	Research users with expertise relevant to the sub-panel	Developing approach at the sub-panel level to assessment. Assessment of case studies and template	158

[10] Output assessors were also used in the assessment of outputs which is outside the scope of this study.

[11] The analysis presented here is aggregated to the level of the four main panels due to limited coverage of perceptions at the level of each of the 36 sub-panels. When describing opinions from panellists at a sub-panels and a main panel level within one main panel the phrase 'areas of Main Panel x' is used.

[12] Throughout this report, 'panellist' is used to describe any individual involved with the assessment of the impact element of REF 2014. Types of panellists are referred to as described in this table. Research user is used to describe those involved in the process from outside the HE sector (i.e. both user members and impact assessors).

two different types of document submitted by HEIs: impact case studies, and impact templates. Case studies provide examples where research has led to impact, and templates set out the submitting unit's wider strategy around facilitating the impact of their research. An outline for a case study and an impact template are provided in Figure 1.2 below.

As described above, submissions are made at the individual UOA level (so that each HEI will make multiple submissions, one per UOA that is relevant to their research). Each submission consists of one impact template, plus a number of case studies (where the number depends on how many staff are submitted by the institution in that UOA, approximately one case study per 10 staff).

In REF 2014, the case studies and templates were allocated for assessment to sub-panel members and impact assessors by the sub-panels, who were asked to review and score them each individually. They were provided with some guidance to do this, and there was a calibration session early in the process to allow them to calibrate their scoring to the rest of the sub-panel. Case studies (REF3b) were reviewed on the criteria of 'reach and significance'.[13] The guidance stated these criteria should be 'taken as a whole', rather than assessed separately. In assessing the impact template (REF3a) the panel considered the extent to which the unit's approach described in the template was conducive to achieving impacts of 'reach' and 'significance'.

Scores were allocated based on the level definitions described in Table 1.2.

Table 1.2: Level definitions used for assessing impact as part of REF 2014

Star level	Description
4	Outstanding impacts in terms of their reach and significance
3	Very considerable impacts in terms of their reach and significance
2	Considerable impacts in terms of their reach and significance
1	Recognised but modest impacts in terms of their reach and significance
Unclassified	The impact is of little or no reach and significance; or the impact was not eligible; or the impact was not underpinned by excellent research produced by the submitted unit

Each case study and template was reviewed by more than one individual and exact practice on the allocation and scoring processes differed between main panels and sub-panels.

Once the case studies had been scored individually, there was a process of moderation of those scores.

Figure 1.2: Outline of impact case study and impact template documents

Impact template (REF3a)

Institution:
Unit of Assessment:
a. Context
b. Approach to impact
c. Strategy and plans
d. Relationship to case studies

Impact case study (REF3b)

Institution:
Unit of Assessment:
Title of case study:
1. Summary of the impact *(indicative maximum 100 words)*
2. Underpinning research *(indicative maximum 500 words)*
3. References to the research *(indicative maximum of six references)*
4. Details of the impact *(indicative maximum 750 words)*
5. Sources to corroborate the impact *(indicative maximum of 10 references)*

[13] The criteria used in REF 2014 are defined as follows: 'reach' is the spread or breadth of influence or effect on the relevant constituencies; 'significance' is defined the intensity of the influence or effect (REF 2012a).

Initially there was some moderation at the sub-panel level (though exact processes for this varied), and then at the main panel level. There was also some limited validation between main panels.

An additional step alongside moderation was audit, a process by which panel members could request checks on the case studies and templates to verify their eligibility against various criteria, or to access the sources of evidence referred to in the case studies. Audit was not conducted systematically across all case studies, but rather on the basis of sub-panel requests. 74% of institutions had at least 5% of their case studies audited.

Finally, scores for case studies and templates in each submission were combined to produce an overall impact 'profile'. The profile indicates the proportion of the submission that is at each level from unclassified, to 1 star, 2 star, 3 star or 4 star (the highest rating). Ultimately these profiles are combined with the output and environment profiles to produce an overall profile for each HEI submitting to each UOA. Funding will ultimately be allocated by the funding bodies on the basis of the REF outcomes, with 3 star or above being the requirement for funding allocation.[14]

1.2. Evaluation methodology

The aims of this evaluation are set out in Box 1.1. This study looks at the process of the assessment of the impact element of the REF. It does not attempt to assess the scores that were allocated to case studies and templates. It follows on from work undertaken in 2014 by RAND Europe to understand the submission process (Manville et al. 2015a, 2015b).

The overall approach is summarised in Figure 1.3. It comprises seven tasks organised in three phases. The inception phase involved familiarising ourselves with the available documentary evidence covering the overall processes from initial planning through to the implementation of the impact element of REF 2014. We used the research questions identified in the ITT to orientate this work. We also drew on previous relevant evaluative evidence. Our data collection phase included a detailed analysis of the panel reports and associated materials (for example, guidance) as well as an assessment of panel scoring data, a survey of panel members, and focus groups and interviews with panel members. All data collection from panellists was conducted before the REF results were published on 18th December 2014. This was to ensure that perceptions of the process were not influenced by reactions to the results, and to collect data in the same circumstances as the evaluation of the submission process. The third and final phase involved synthesising the evidence produced and arriving at an assessment of the burden. We approached the data in the synthesis phase both 'bottom-up' (what are the overall messages emerging from the evidence?) and 'top down' (what does the data tell us about the overall research questions identified in the Invitation To Tenders (ITT)?). Each of the main components of the methodology are described in further detail below. This approach has generated a rich and robust understanding of the processes and the experiences of those most closely involved. However, it will not – and is not intended to – deliver an assessment of the actual scale and consequence of the impacts themselves. For example, it allows us to assess whether participants regarded the processes as fair and robust, and whether this is aligned with other evidence about the process, but it does not allow us to judge whether the impacts were as claimed by HEIs.

1.2.1. Document review

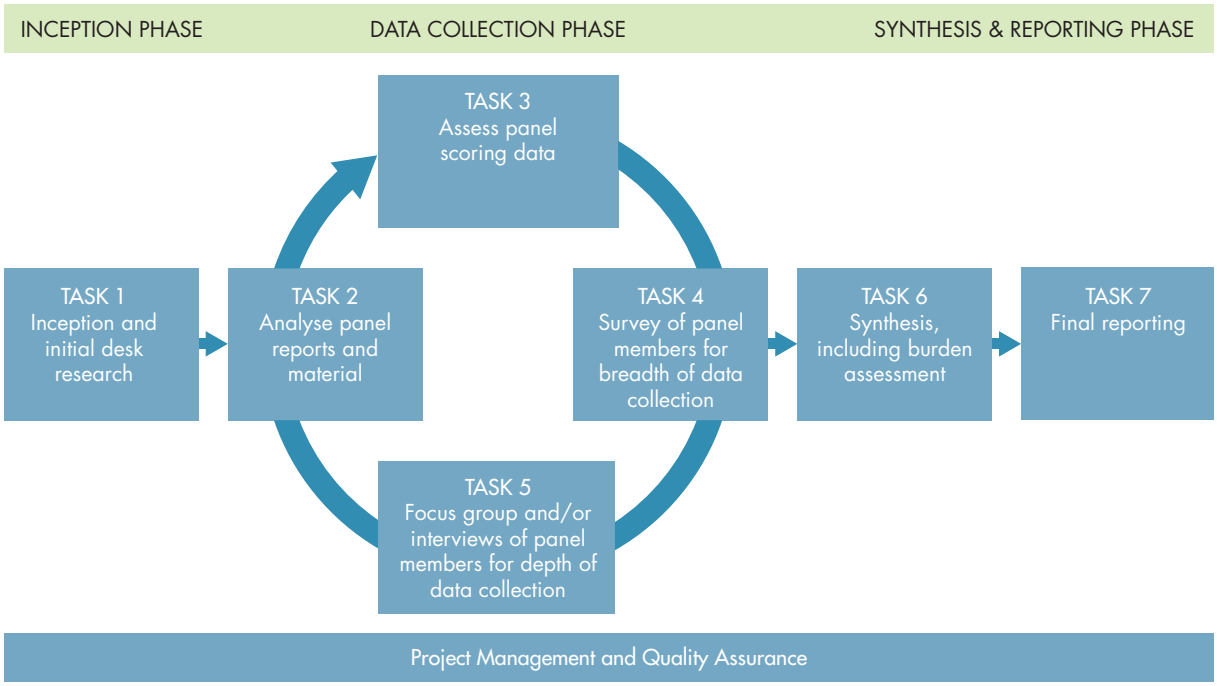
The research team conducted a document review of publically available material that is linked to the REF 2014 process, paying particular attention to details of the assessment process. All documents were reviewed prior to undertaking fieldwork to develop an understanding of the process, and inform the protocol development.

Box 1.1: Aims of the evaluation

- Explore the strengths and weaknesses of the assessment process in delivering reliable, robust and fair outcomes.
- Identify any unforeseen issues emerging during the assessment process, and evaluate the management of these.
- Outline broad implications for the impact assessment process in future REF exercises.
- Understand the study's findings within the context of the evaluation of the submission process.

[14] For further information on the process, see the following chapters, and overview reports published by each main panel (REF 2015).

Figure 1.3: Project schema



Documents included the results of the impact pilot (Technopolis 2010), guidance on submission (such as REF 2011, 2012), and main panels' reflections on the REF process (REF 2015). In addition, various documents were made available to the research team via HEFCE. These included:

- Panel membership
- Secretariat guidance – Impact meeting 1 preparation, Jan, 2014
- Panel briefing – Impact, Feb, 2014
- Main Panel B – example impact case studies
- REF Impact briefing for user briefing events – Main Panels A-D
- Secretariat guidance – selection of impact audit sample

1.2.2. Focus groups with academics and research users

Two focus group meetings were convened by HEFCE in late November. The first (21st November 2014) was for research users (either sub-panel members or impact assessors), and the second (28st November 2014) was for academic panel members.[15] One research user panel member or impact assessor per sub-panel, and two academic panel members per sub-panel, attended their respective meetings. The participants in these groups were nominated by their sub-panel chairs. Groups were held at a main panel level.[16] The session on the impact element lasted one hour and discussion was facilitated by two members of the team from RAND Europe. During the meeting panellists also attended other sessions, focusing on metrics, research environment for the academics and research users' experiences for research users. The focus groups varied in size, from 9 to 20 panellists. Due to this size and limited time, we conducted a survey in advance of sessions asking a range of questions on perceptions of the process, in order to determine the topics to focus on in the meetings. The pre-focus group survey data was analysed by panel and by type of panellist (research user and academics).[17] The themes for discussion were selected based on areas of disagreement or consensus over where aspects of the process did not work optimally. There were three main themes that were explored in the focus group discussions:

- Scoring, moderating and calibrating impact case study assessments;

[15] Academic panel members also attended a second focus group in January, although it did not have a session on impact and RAND Europe did not attend.

[16] In the academic focus groups there was also a cross main panel group.

[17] The volume of qualitative data collected through the focus group meetings was unexpectedly high and very granular. It was therefore decided to add this to the NVivo analysis.

- Working with the REF's definitions, rules and templates;
- How corroborating evidence informed impact case study assessment.

The full protocol used is given in Appendix B. Data were collected by taking detailed written and typed notes, as well as audio recordings of each focus group. These were then coded using QSR NVivo 10 International software. The research team developed an NVivo code book for the analysis of unstructured qualitative data based on the parts of the process, areas of focus listed in the ITT, and generic nodes, such as type of panellist and affiliated sub-panel (Appendix C).

Following the sessions, notes and audio recordings were used to write up memos from each focus group. Verbatim transcripts were not made, but detailed notes were developed. Memos were reviewed, and where appropriate revised, by the second researcher on the visit and uploaded into QRS NVivo 10 software, where statements were assigned to descriptive and analytical categories. It is important to note that statements could be coded to as many nodes as applicable. Once coding was complete coding matrices were run across the NVivo data, pulling out comments coded to two nodes to provide a subset of data around thematic codes to review.

1.2.3. Interviews with panellists

The aim of the interviews was to understand the process of assessment and panellists' perceptions of the process. Twenty interviews were conducted, and these were divided between panel advisors, sub-panel members and impact assessors.

The panel advisors were able to provide an overview of the process and the way that it had been carried out. As a panel advisor spanned three or more sub-panels within a main panel they were selected using a randomly generated number approach, starting with the individuals with the lowest number across each main panel. To maximise perspective across sub-panels, in light of the small sample size of interviews, two academic sub-panel members and two impact assessors from within each main panel were selected from the remaining sub-panels using the same approach.[18]

The protocol was framed around the parts of the process providing context and detailed understanding on the effectiveness and suitability of the rules and guidance, the training process and the assessment process. Interviews also explored unforeseen issues and resolution processes.

The full protocol used is presented in Appendix E. Interviews were conducted on the phone and were recorded for note taking purposes only.

1.2.4. Survey of all panellists involved in the assessment of impact in REF 2014

A survey was sent to all panellists involved in the impact element of the assessment. The purpose was to ensure that the views of all of those involved in the process were captured. Panel members were classified as sub-panel impact assessors, main panel users, main panel members, advisors, secretaries and academic sub-panel members.

Seven types of surveys were sent in total to:

- Main panel members
- Main panel users
- Academic sub-panel members
- Academic sub-panel and main panel members (for academic members who had both a role on a sub-panel and a main panel. i.e. sub-panel chairs)
- Impact assessors
- Advisors and secretariat on sub-panels
- Advisors and secretariat with roles on both a sub-panel and main panel.

Questions were asked on each of the following topics:

- The profile of the respondent (including their previous knowledge of research impact assessment)
- The rules and guidance for assessing impact
- The process of assessment (including whether panellists felt they had appropriate expertise to assess allocated case studies, communication within and across sub-panels and comparing different types of impact)
- The burden of participating in the exercise
- Reflections on the exercise for the next REF.

The full protocols used are available in Appendix E.

The surveys were tested at the end of October on members of the steering group and staff from HEFCE, and modified based on their feedback. They were open for four and a half weeks in November and December

[18] Attendees from the focus group were excluded from sample lists as their views had already been captured in this study.

2014. Respondents were sent a personal link and then two reminders, one half way through the survey window and one on the closing date. The response rates are shown in Table 1.3. The data was analysed in Microsoft Excel.

1.2.5. Analysis of the scores awarded

We were provided with two datasets by HEFCE – the scores for each impact case study and impact template submitted to a specific UOA – as well as the overall submission results by institution, at the UOA level. These datasets were then anonymised and interrogated to further understand the process which took place – for example the allocation of case studies, the consequences of the levels awarded and the relationship between impact case study and impact template scores.

Initially the data was cleaned and coded to anonymise individual scores. New grouping variables were developed to conduct within- and cross-group analysis at both the institution and case study/template level. These were:

- *Type of institution,*[19] based on the number of full-time equivalent (FTE) staff submitted:
  - Group 1 – institutions making up the first 50 per cent of FTEs
  - Group 2 - institutions making up the next 30 per cent of FTEs
  - Group 3 - institutions making up the final 20 per cent of FTEs.
- *Submission size*, determined by the number of case studies submitted by one institution to a given UOA:
  - Large: six or more case studies submitted
  - Medium: three, four or five case studies submitted
  - Small: two case studies submitted.
- *Types of audit queries*, distinguishing between:
  - Sources, when the query referred to accessing a corroborating source in Section 5 of the case study including queries marked as managed by the secretariat
  - Reference, when the query referred to accessing research references in Section 3 of the case study
  - Staff check, when the query was about the need for the institution to provide dates of employment of any of the staff members indicated in the case study
  - Secretariat, when the query was managed by the secretariat
  - Institution check, when the query referred to the need for the institution to provide justification that the research was actually conducted in the claiming institution
  - Other.

We used descriptive techniques to analyse the distribution of scores within the impact element, and other elements of REF 2014, at a main panel and sub-panel level. In addition, we assessed the association between impact scores and other elements of REF 2014. In

Table 1.3: Response rates to the surveys

Type of survey	Number of individuals invited to complete the survey	Number of respondents	Response rate (%)
Main panel member survey	27	17	63
Main panel user survey	24	13	54
Academic sub-panel survey	874	424	49
Academic sub-panel and main panel survey	39	25	64
Advisor and secretariat sub-panel survey	16	11	69
Advisor and secretariat sub-panel and main panel survey	16	8	50
Impact assessor sub-panel survey	165	77	47

[19] This grouping was also used in Phase 1 evaluating the submission process of the impact element of REF 2014 (Manville et al. 2015a) and the analysis of case studies (King's College London and Digital Science 2015).

the absence of more detailed data, this was accomplished by calculating the pairwise linear correlations, for example between the percentage of submissions with 4 stars in 'impact' and in 'environment', 'output' sub-profiles and 'overall' profiles.

1.2.6. Synthesis across methodologies

The 'top-down' approach

Due to the volume of data collected, and the diverse methods employed, we developed a 'top-down' approach in order to begin the process of understanding and synthesising across the data sources. We started by asking each member of the evaluation team to independently reflect on the key messages from the methodology they were mainly involved with. These messages were clustered through an internal workshop to generate 'top down' themes, supported by different data streams.

The 'bottom-up' approach

Each of the methodologies generated a rich volume of data and evidence of both a qualitative and quantitative nature. This was analysed by the members of the project team most closely involved in each particular element of the project.

Triangulation across the evidence streams

Once the analysis of each evidence stream was completed, we triangulated between the different sources of data to ensure that the 'top-down' messages were supported by data and merited inclusion, and looked for any themes emerging from the 'bottom-up' analysis. Messages and observations were refined according to the data coming from the relevant evidence streams. This process proceeded in an iterative fashion as the evaluation team worked through each set of analyses and continued testing different hypotheses to ensure that all information and data were captured and synthesised appropriately. This iterative process resulted in the continued refinement and assessment of the key findings to ensure that each one was fully supported by a robust evidence base that drew upon all appropriate aspects of the evaluation. The result was a series of eight key findings and observations, outlined in the Overarching Summary.

1.3. Caveats and limitations

Table 1.4 summarises the limitations of each methodology. The methodologies were chosen to complement each other; although each methodology had some limitations, none of these limitations was common to all methodologies and we were able to base our conclusions on multiple sources rather than relying on any one individual source.

1.4. Overview of this report

The Overarching Summary at the start of this report presents our key findings. The rest of the report is organised around the different parts of the assessment process and includes data from the different methodologies employed. Chapter 2 details the panel structure and members, and is followed by a review of the allocation of case studies in Chapter 3. Chapter 4 described the calibration process. Chapters 5 and 6 detail the process individuals undertook to apply the criteria and score impact case studies and templates respectively. Chapter 7 examines how scores were agreed and moderated and Chapter 8 describes the audit process. Finally Chapter 9 presents an overview of the benefits and challenges experienced by individuals engaging in the process, and Chapter 10 an assessment of the burden of being involved.

Table 1.4: Overview of caveats and limitations of the study

	Document review	Focus groups	Interviews	Survey	Scoring data analysis	Cost estimation	Synthesis
There was limited opportunity to interrogate data further	✓			✓			
We had limited control over available data	✓				✓		
There may have been reluctance to air unpopular or minority views		✓					
Sampling bias		✓	✓				
Semi-structured protocol meant not all questions were asked on all occasions		✓	✓				
Time limitations meant that not raising a view was not the same as not holding a view		✓	✓				
Not all individuals interpret all questions in the same way		✓	✓	✓		✓	
Recall bias		✓	✓	✓		✓	
Contradictory points could be raised from within one grouping and it was not always possible to give a representative view or even a sense of scale		✓	✓	✓			
Different members of the evaluation team had different coding styles		✓	✓	✓		✓	
The sample size was small relative to the entire pool			✓				
The accuracy of the time estimates may vary				✓			
Views are restricted to those of the project team							✓
Wide variation in time estimates provided						✓	
Limitations of external data sources required for estimation[20]						✓	

[20] This includes: no direct information on salary levels available for individuals, draws on external estimates for costs of other components of the process, assumptions about future QR funding allocation required.

Chapter 2 Panel structure and membership

> *The inclusion of research users brought a real 'credibility' to the exercise.*

2.1. Structure of the panels

2.1.1. Role and composition of the sub-panels

REF 2014 divided the academic disciplines across four main panels, comprising one sub-panel for each of the 36 units of assessment (UOAs) (see Appendix A). Focused around subjects of research, these were expert sub-panels that assessed submissions of outputs, impact and research environment. Each was composed of:

- A chair and deputy chair
- Approximately 10 to 30 members,[21] the majority academics and the remainder research users
- Assessors specifically involved in the assessment of either outputs (in the case of academics) or impact (in the case of research users)
- A secretariat.[22]

Research users (user members and impact assessors)

To complement the academic panellists, some full members of the sub-panels were from research user organisations, such as the British Library, the Overseas Development Institute, the BBC, Royal Museums Greenwich, Oxfam, BT, BAE systems, and the Bank of England. In addition, a number of research users were brought into the process to assess the impact submissions alongside panel members. REF guidance stressed that 'Assessors will play a full and equal role to sub-panel members in developing the sub-profiles for either the impact or outputs element of the assessment. They will be fully briefed, take part in calibration exercises and attend panel meetings at which the relevant aspects of submissions (outputs or impact) are discussed'.[23]

Panellists' perceptions of the inclusion of research users in the assessment of impact

They majority of academic panellists in our focus groups felt that research users were a very worthwhile inclusion in relation to the assessment of impact. The diversity of panellists provided both benefits and challenges to the assessment process, which were discussed in the focus groups. Academics and research users felt that the latter brought a different perspective to the process, being external to the academic sector, and that their experience of carrying out, using and commissioning research was valuable. On the other hand, some individuals noted that it took time for some user members and impact assessors to become familiar with the quality levels used throughout the REF.

Research users were employed very differently in different panels. On occasions impact assessors also assessed some outputs, in which case there was not a clear distinction between panel members and impact assessors, whereas the majority were used solely to assess the impact submissions. Many had links with the academic world which helped them understand the process and to feel involved; however, it was questioned whether this knowledge was a good or a bad thing in an assessment that was aiming to obtain a perspective from outside academia.

---

[21] Number depending on the scale and breadth of the panel's remit.

[22] Further details on the roles and responsibilities of the panels and process for appointing them are given in REF 2010.

[23] For panel criteria and working methods see REF 2012.

When asked about potential improvements, academics and research users from areas of Main Panel A thought that the impact assessors would have benefitted from being part of the panel earlier on, since some felt it was more difficult to be included once the sub-panels had been functioning as a group for a while, and that it did not build a cohesive, positive community spirit in the process if individuals came and went without integrating. However, they understood this would increase the time commitment required, and were conscious that the level of burden was already an issue to some (see Section 9.3). Where research users came into the process at a late stage, anecdotal evidence suggested that they applied the guidance very rigidly and consequently their scores were lower than those awarded by the sub-panel in the calibration phase. Owing to the short timeline of engagement and the compressed period for assessing impact, there was little time for this discrepancy to be addressed prior to the assessment. Another panellist commented that impact assessors would have benefitted from listening to the discussions around some of the outputs and environments to inform their understanding of the assessment process.

2.1.2. Role and composition of the main panels

The role of the main panels was to provide guidance to the sub-panels and oversight of the process.[24] They were responsible for ensuring adherence to the published criteria and working methods so that consistent assessment standards were met across sub-panels. The main panels were composed of

- A chair
- Members:
  - Sub-panel chairs
  - Between four and six international members
  - Between three and six user members
- Between two and five observers from UK Research Councils
- Three or four panel advisers.

Focus group attendees stressed the importance of the main panels in maintaining consistency. The main panel international and user members circulated around the sub-panels to make sure there were no significant discrepancies between the processes at the sub-panel level. They also contributed to ensuring that a wider perspective was taken when scoring impact and were able to provide constructive feedback following the calibration exercise.

In particular sub-panellists from across the four main panels stressed the importance of guidance from the main panel chairs. As one user member stated: 'At the beginning of the process there were differences across the sub-panels that the main panel chair intercepted to guide everyone onto the same page'. The chairs also provided assurance and confirmation that their assessment was comparable to other sub-panels within their main panels. Respondents also highlighted the role of the main panel chair in ensuring the criteria were followed across the sub-panels. One panellist said: 'our chair was very consistent in reminding sub-panel members that [we] had to justify the assessment that we were providing in terms of the criteria'.

International members

As stated above, each main panel had a number of international members. They brought an international perspective to the assessment, as well as attending meetings across the sub-panels to support consistency in the process, alongside other main panel members. They engaged in the calibration process (see Chapter 4) and provided international benchmarking between the sub-panels which was key to ensuring that assessment processes were robust and reliable in an international arena. Members of areas of Main Panel A and areas of Main Panel C commented how beneficial this had been.[25]

Panel advisers

Main panel members (including international members) highlighted the great importance of panel advisers in regards to the logistic support and expert understanding of the process and criteria which they provided.

Consistency in the process was also provided via the REF team, and the REF manager, Graeme Rosenberg, who frequently attended main panel and sub-panel meetings in order to provide consistent interpretation of the evaluation criteria across all panels.[26]

---

[24] Further details on the roles of sub-panel and main panel are outlined in REF 2010.

[25] As mentioned in REF 2015 for Main Panels A, B and D for the Research Excellence Framework (REF 2014) exercise.

[26] As mentioned in the REF 2015 for Main Panel A.

Table 2.1: Types of panellists by role

Sub-panel members[27]	Secretariat and advisors	Impact assessors	Output assessors	Main panel members	Observers
762	88	160	139	105	24

2.2. Recruitment of panellists

The recruitment of panellists for REF 2014 commenced in 2010 with the appointment of the four main panel chairs by an application process. In late 2010, sub-panel chairs were appointed by application. In 2011, the recruitment of main and sub-panel members (academics and users) commenced with an invitation to organisations to make nominations. The majority of panel members were appointed in early 2011, with a small number appointed through nomination or co-option in spring 2011. Impact assessors were nominated in late 2012 and early 2013, and appointed in early 2013. Simultaneously, in late 2012, HEIs were asked to provide an indication of the anticipated volume and content of their submissions.[28] Using this data, the sub-panels were able to recruit further in areas where they believed they had insufficient panellists for the volume of submissions, or specific areas which the already recruited panellists could not cover. A small number of further appointments was made following submissions in late 2013. This was an iterative process designed to ensure the appropriate expertise was in place to assess the impact element of the submission. At the time of assessment, there were a total of 1,285 panellists (see Table 2.1).[29]

It is important to note that there were full sub-panel members who were research users and were involved in assessing all elements of the REF.[30] For example, individuals included a deputy director in the R&D directorate of the Department of Health (UOA1), a director of Arup (UOA14), the head of research at the British Museum (UOA17), and the general manager of the BBC Symphony Orchestra (UOA35). To complement this, 160 impact assessors were recruited onto the panels for the assessment of the impact element of REF 2014.

In general the assessment of impact documents was undertaken by sub-panel members and impact assessors. Taking into account the sub-panel members who were research users, 27 per cent of panellists assessing the impact element of REF 2014 were research users. In general, the two perspectives were valued (see Section 4.4). During the focus groups the ratio of academics to research users was discussed. A few individuals (academic and research users) thought that it would have been better to have more equal numbers of academics and research users assessing impact as they felt that research users had been the minority voice. For example, one research user noted that there were times when their scores were different to others and they felt they really had to argue their case. It is important to note that the experience may have differed depending on how accustomed the research users were to engaging with academia.

Expertise and experience

Through the survey we asked panellists about their level of knowledge of impact assessment prior to their involvement with REF 2014 (see Figure 2.1). The majority (over 70 per cent) felt they had a good or very good knowledge of research impact.[31] Some 70 per cent of academic sub-panel members (including sub-panel chairs) were involved in preparing their institution's impact submission.

There was a divergence of opinion regarding the importance of expertise and experience in assessing impact case studies and templates. Panellists in areas of Main Panel D stressed the need for expertise on sub-panels to assess evidence claims, for example appropriate indicators of web traffic, and to prevent over-scoring. However there was recognition that this was possibly unfair if individuals on the panel had personal and detailed knowledge to support a more informed critique of some of the case studies, whereas others would be taken at face value. Examples were given in areas of Main Panels A and C where those who knew about

[27] Including sub-panel chairs and deputy chairs where relevant.

[28] REF 2012b.

[29] Note the total number of panellists is higher than the absolute number of panellists, since 154 panellists had more than one role.

[30] Based on their organisational affiliation as listed in the REF membership.

[31] When broken down by type of panellist, only just over 50 per cent of impact assessors rated their prior knowledge of research impact as good or very good.

Figure 2.1: Level of knowledge of impact assessment prior to involvement with REF 2014

specific case studies 'were told [by sub-panel chairs or fellow panellists that] they weren't allowed to use that information'. This was confusing and frustrating for the experts involved, as panellists believed that this very knowledge and expertise partly justified why they had been selected for the panel.

Chapter 3 Allocation of impact case studies and impact templates

> *We ensured that no groups of three were ever the same; that way you were always discussing with different people.*

> *The chair allocated a set of case studies in an area where he thought the user was particularly strong, so that we were able to think about kinds of claimed impact common to that area and have a way of comparing them to each other.*

3.1. The process of allocation

Impact case studies and templates were allocated to two or three sub-panel members or impact assessors by sub-panel chairs in early 2014. Guidance provided at the training for sub-panel members and impact assessors in advance stated that impact assessors were expected to be allocated approximately 30–50 documents. Where a sub-panel did not have the necessary expertise it was expected that they would cross-refer a case study to another sub-panel so that advice could be sought.

Data was provided by HEFCE detailing which individuals reviewed each case study, enabling us to determine the number of case studies reviewed by each individual panellist and to identify how many reviews each case study received. A comparison was also made between the number of reviews a case study received and its final score, and between scores given to case studies and their corresponding impact template. The following sections look at each of these points of analysis in turn.

3.1.1. Mechanisms by which impact documents were allocated

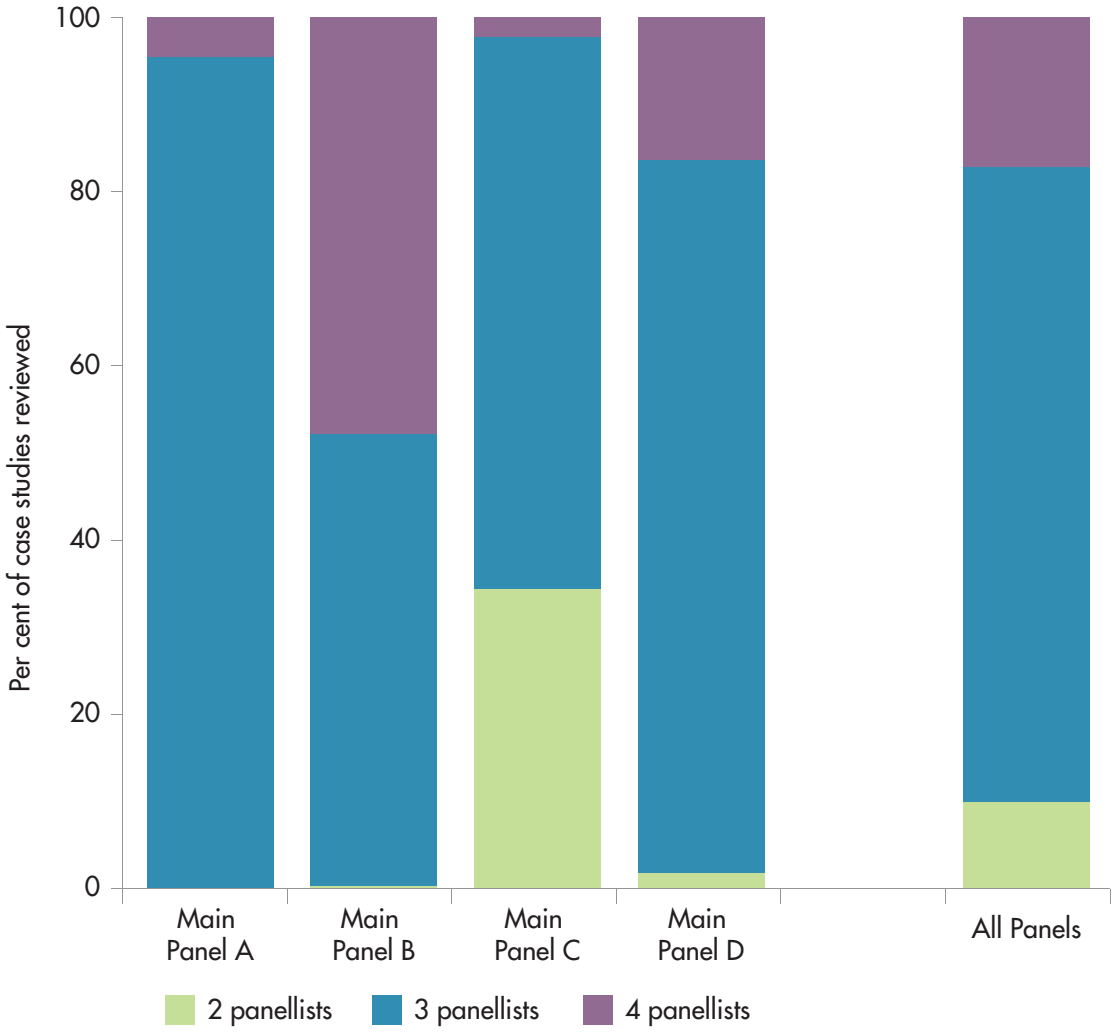
Individuals noted that the way in which impact case studies and templates were allocated differed across the sub-panels. While several individuals noted that allocations of impact case studies and templates were made on the basis of the research user's expertise and experience, others said that case studies were allotted to groups on the basis of the type of impact that they claimed, which 'increased our confidence in our ability to assess claims [of impact]'. It is important to note that individuals were not always aware of the strategy behind allocation in their sub-panel. One commented that it 'appeared to have been done at random'. It could be argued that this element should be more transparent in order to build confidence in the process. On a sample (5.5 per cent) of submissions we analysed whether the impact template was reviewed by a panellist who also reviewed that institution's case studies. This was true in over 98 per cent of the sample.

3.1.2. Each impact case study and template was reviewed by between two and four panellists

All case studies were assessed by between two and four panellists (Figure 3.1).[32] While the number of panellists assessing a case study varied between sub-panels, the majority of case studies (73 per cent) were assessed by three panellists. In Main Panel A, all case studies had at least three reviewers (with 5 per cent of case studies being reviewed by four panel members), in Main Panel D over 82 per cent of case studies were reviewed by three sub-panel members. Main Panel B used almost an equal weighting of three (52 per cent) and four reviewers (47 per cent) per case study. Main Panel C was the

[32] It is important to note that our dataset reflects all impact documents an individual was asked to review, irrespective of the point in the process the request was made. Therefore impact documents where there was uncertainty would have more reviewers if others were brought in to discuss and assess them.

Figure 3.1: Number of panellists reviewing a given case study



only one where a significant number of case studies (one third) were assessed by two reviewers.

The trends seen at the main panel level are reflected in allocation decisions at the sub-panel level (see Figure 3.2).[33] For the majority of sub-panels (26), more than 90 per cent of case studies had a constant number of reviewers, but the number varied within and between main panels leading to the overall patterns seen in Figure 3.1 above. Yet nearly 50 per cent of sub-panels did not use the same number of reviewers across all case studies.

Templates typically had more reviewers on average than the case studies, as illustrated in Figure 3.3, with 30 per cent being reviewed by four panellists (compared to 17 per cent of case studies). This difference is largely due to the approach taken in Main Panel A, where 68 per cent of templates had four reviewers, compared to 5 per cent of case studies. A further 32 per cent were reviewed by three individuals. In the other three main panels a similar pattern to case study allocation was observed. For example, Main Panel C was again the only main panel with a significant proportion reviewed by two reviewers, accounting for nearly one third of impact templates.

Looking at the breakdown of reviewers of impact templates by sub-panel, as for the case studies, most sub-panels had a consistent internal process, with each

[33] Note that data was not available of UOA16, 27 or 32 because the panels did not record data on the spreadsheet, as it did not enable them to record information relevant to their sub-panel.

[34] As above with case studies, data was not available for sub-panels 16, 27 and 32 because the panels did not record data on the spreadsheet, as it did not enable them to record information relevant to their sub-panel.

Figure 3.2: Percentage of case studies by number of panel members assessing each case study

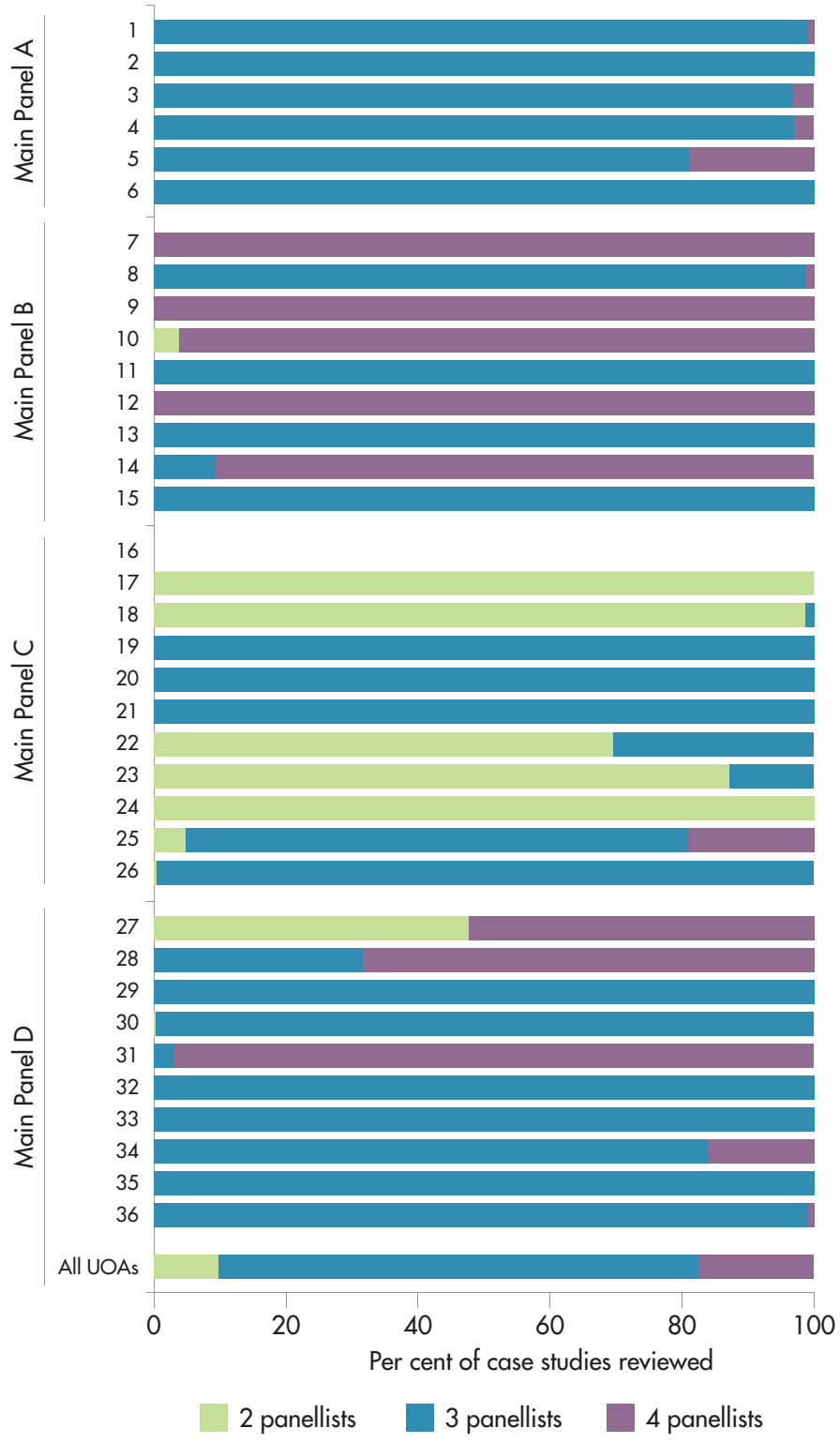
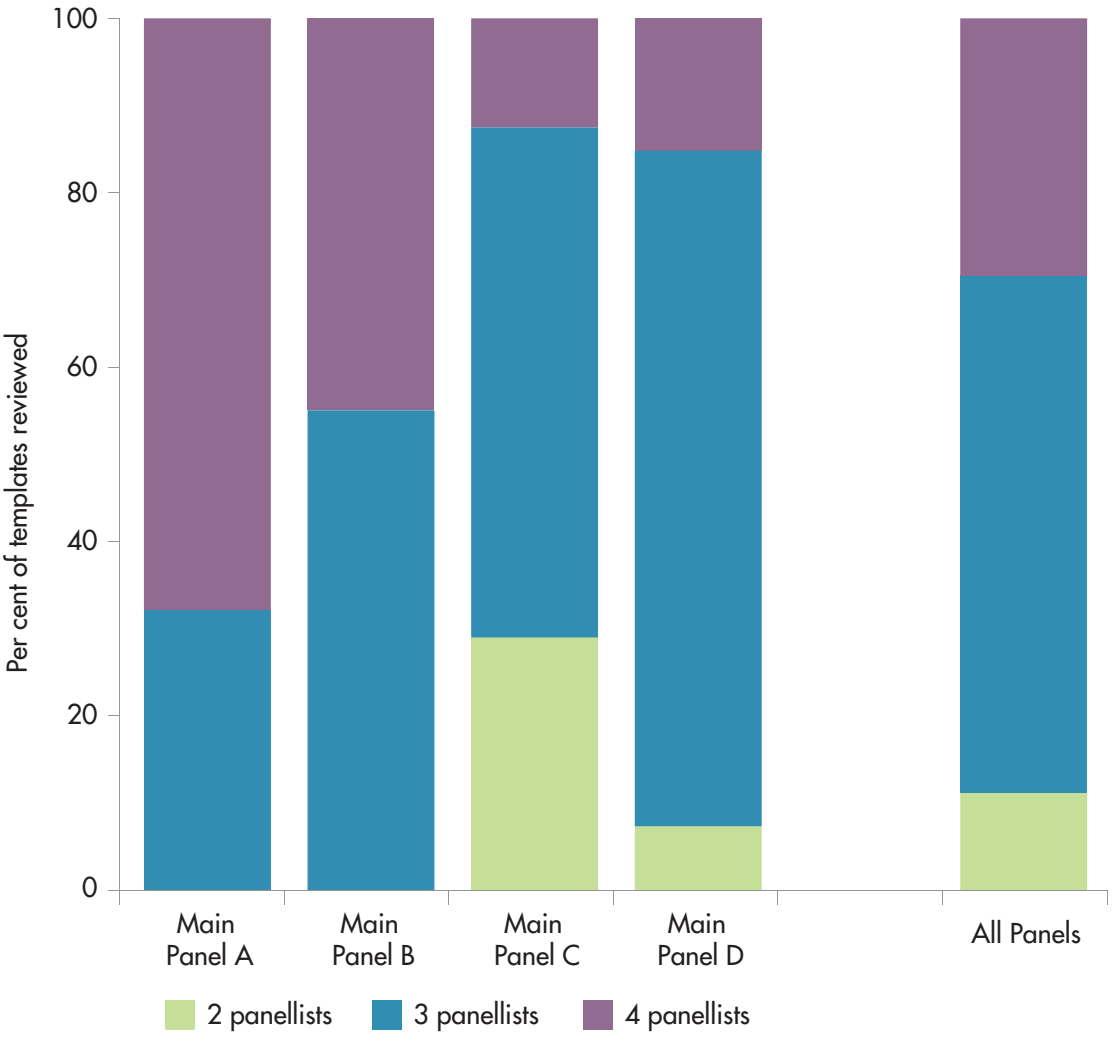


Figure 3.3: Number of panel members assessing each impact template



template receiving the same number of reviewers (see Figure 3.4).[34] This highlights that the variation at the main panel level stems from differences between policies at the sub-panel level. There are, however, some sub-panels which did not have the same number of reviewers for each template (at least one per main panel), and this is particularly notable in Main Panel C, where five sub-panels had a varying number of reviewers per template. Sub-panels which varied the number of case study reviewers corresponded to some extent to those varying the number of template reviewers (e.g. UOAs 5, 22, 23, 25 and 34) but the overlap was not complete with some varying the number of reviewers just for the impact case studies or just for the impact templates.

3.1.3. There was variation in the number of impact documents any individual was asked to review

There was a large variation in the number of case studies any individual was asked to review. On average impact assessors reviewed more impact case studies than sub-panel members (see Figure 3.5). Impact assessors in Main Panels A and B on average reviewed more case studies than those in Main Panels C and D. The median number of case studies reviewed by an impact assessor was 38, compared to 24 for a sub-panel member.[35] The scoring data shows that the range of case studies being reviewed within Main Panel B by sub-panel members was particularly broad as they

[35] This difference may be a result of the panel structure and the number of sub-panel members compared with impact assessors.

Figure 3.4: Percentage of impact templates by number of panellists assessing each impact template

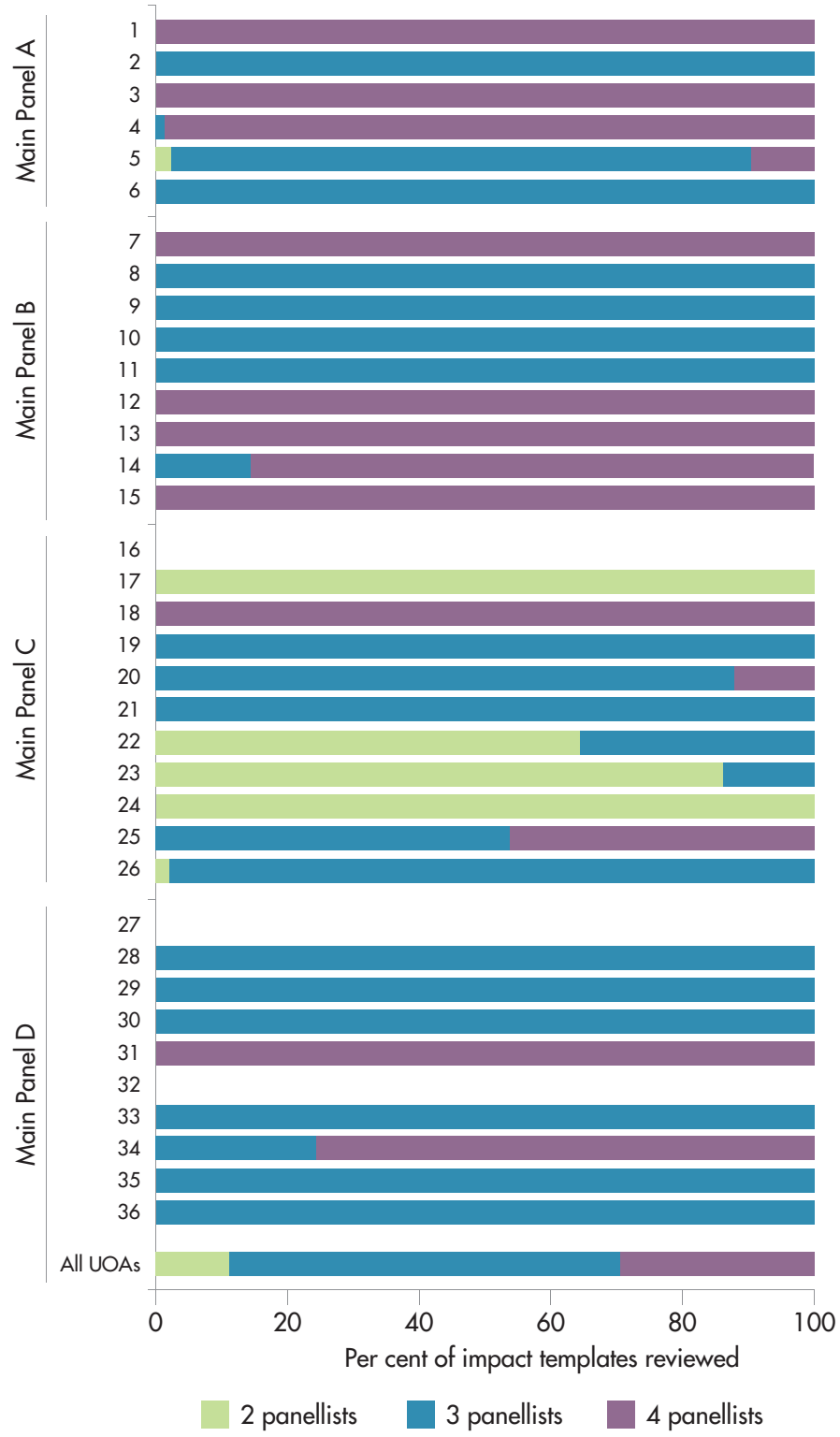
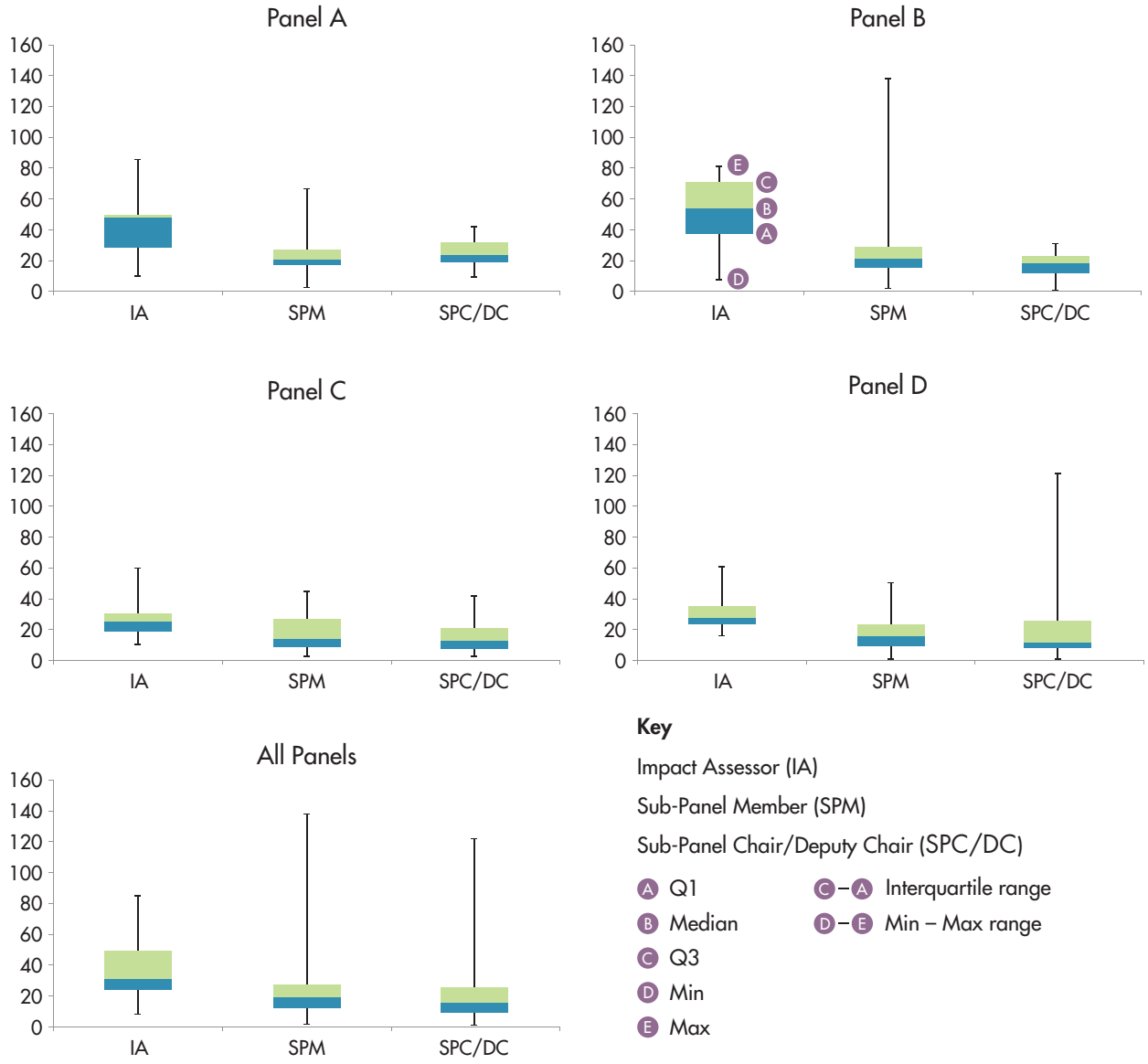


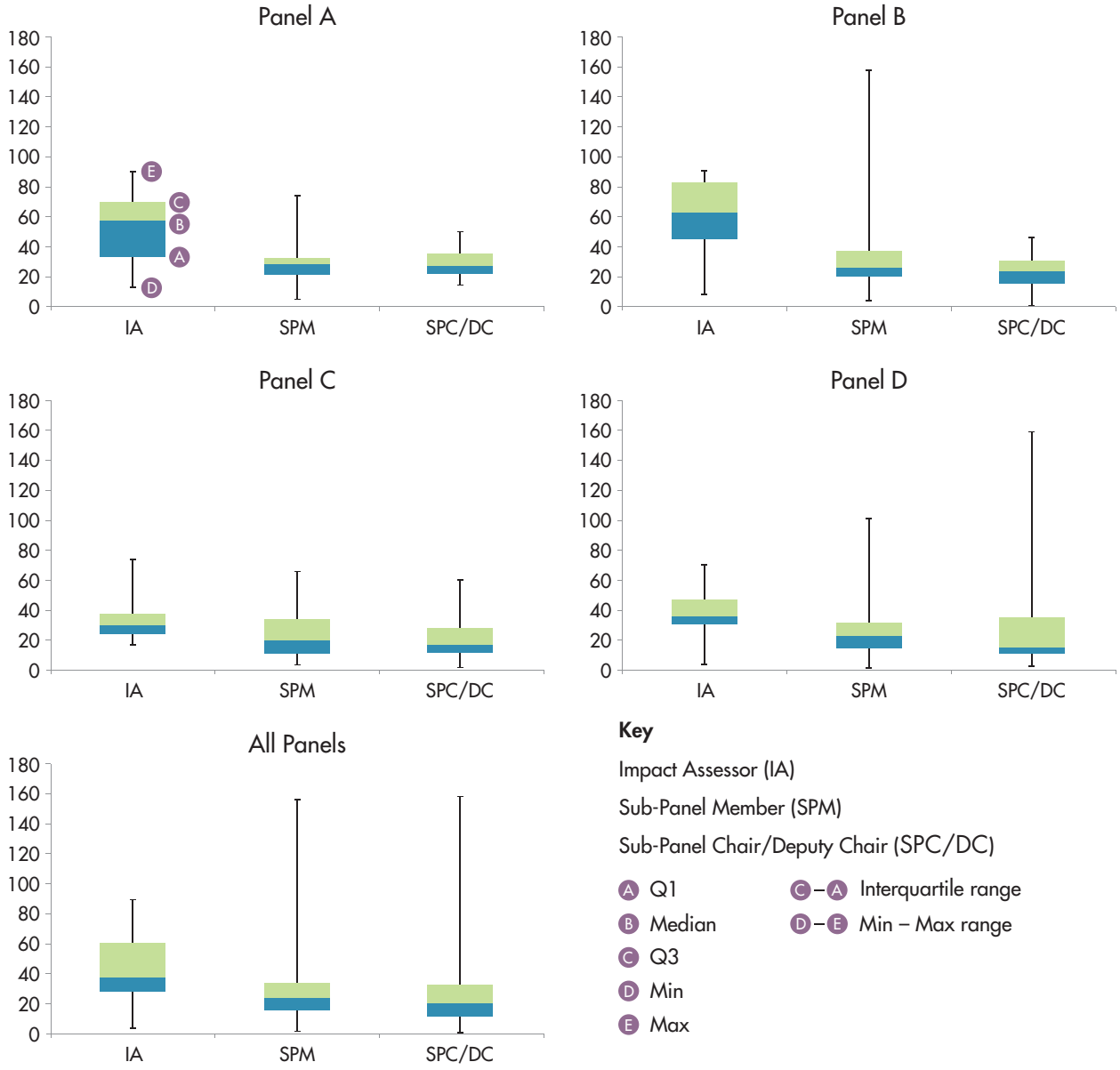
Figure 3.5: The distribution of impact case studies reviewed by role and main panel



were reviewing anywhere between 2 and 138 case studies per person (a median of 21).

Including impact templates in this analysis, we still see a high level of variation between individuals. The median number of impact documents (impact case studies plus impact templates) reviewed by each panellist was 27 (a median of 38 and 24 for impact assessors and sub-panel members respectively) (see Figure 3.6). Looking at the ranges we can see that there is still significant variation between individuals within as well as between main panels. Again, sub-panel members had a particularly wide level of variation, with the total number of impact documents reviewed by any individual ranging from one to 157. Some panellists in focus groups and interviews felt that seeing a body of documents supported their judgements in impact. A handful of research users questioned whether subject specific expertise was required to assess impact. They suggested that research users could be used as a floating resource pool, rather than linked to specific sub-panels. As one panellist stated: 'I found that impact tended to be more generic and non-discipline specific than knowledge generated through research outputs. I think there are strong grounds, on this basis, for impact to be assessed in the future by a specific impact panel that operates across all, or a cluster of, sub-panels within a main panel'.

Figure 3.6: The distribution of impact documents (case studies and templates) reviewed by role and main panel

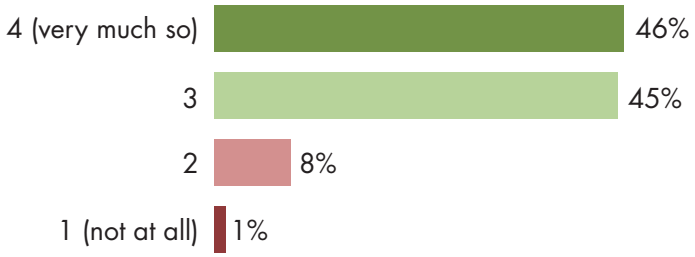


3.2. Expertise to assess impact submissions

Based on the survey data, panellists felt that they had the expertise necessary to comfortably assess the case studies that they were allocated (see Figure 3.7). There were no significant differences between main panels or between the views of sub-panel members and impact assessors. Research users in areas of Main Panel D felt an important role that they played was in interpreting audience behaviour data which was often submitted as evidence of impact in case studies they assessed.

Figure 3.7: Sub-panel members' and impact assessors' assessment of whether they had the expertise to assess case studies allocated

Q. To what extent were you allocated case studies that you felt you had the expertise to assess?



Chapter 4 Calibration exercise

> *"Certainly I think it is really helpful for people to acknowledge whether they are doves or hawks in their scoring both for the individual and for the group… so I know where I sat within my dovishness and hawkishness and I was then able to look at my average scores for the calibration set with the average scores that came out of each of the sub-panels.*

4.1. The process of calibration

Calibration meetings provided the opportunity to reach a shared understanding of the star levels and approaches to assessment. In addition, they were a chance to discuss sections of the guidance and address questions regarding impact assessment in the context of REF.

Initially, calibration occurred at a main panel level. Each main panel calibrated a sample of items from their respective sub-panels. The main panel then shared its views about what impact was and how it was constituted in each of the sub-panels as well as across main panels in order to ensure consistency of understanding and approach. Main panel members then attended calibration meetings at a sub-panel level, which was thought to have helped to achieve 'integrity across all panels in a more consistent way'. At a sub-panel level, in advance of the calibration meeting, panellists were sent a spreadsheet listing calibration sample case studies (which included both main panel and sub-panel items) and guidance on how to assess them, in advance of the assessment.[36] Once sub-panel members and impact assessors had reviewed and scored the sample case studies independently of one another, they attended briefing calibration meeting to discuss them. Each calibration meeting included all sub-panel members and impact assessors, as well as some members of their respective main panels. Inclusion of main panel members in this element allowed advice to be drawn from the main panels' international and user members and it also enabled a shared understanding of and approach to each element of the assessment to be reached. It was thought by main panel members that this iterative process between main panel and sub-panels facilitated an effective cross-disciplinary calibration. Panellists thought that being involved in calibration at both levels was useful as once main panel members had been involved in the calibration process for one sub-panel they could provide comparative feedback to other sub-panels regarding scoring practices. Other individuals noted different ways in which feedback was provided. In some instances it was reported that main panel user members reviewed all calibration case studies and compared 3-star or 4-star borderline cases against case studies from other sub-panels to calibrate and provide feedback to the sub-panels. In Main Panel C, these meetings were led by the user panel members.[37]

4.2. Opinions varied regarding the selection of case studies for the calibration sample

Our document review highlighted that sub-panels had different approaches to selecting sample case studies for calibration. Whilst in Main Panel D[38] case studies

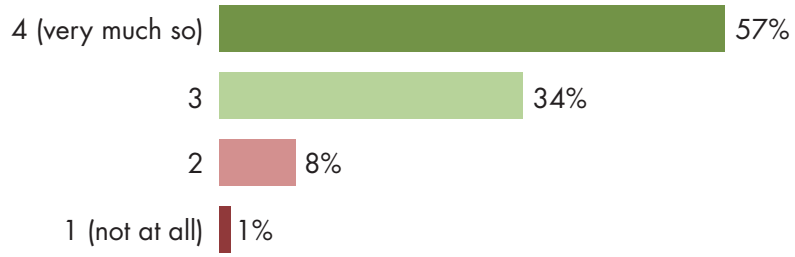
[36] In addition, at this point panellists were allocated the case studies and impact templates that they would be reviewing.

[37] REF 2015, for Main Panel C.

[38] REF 2015, for Main Panel D.

Figure 4.1: The extent to which sub-panel members and impact assessors felt that the impact calibration exercise allowed impact case studies to be assessed fairly and reliably (n=596)

Q. To what extent do you think the impact calibration exercise allowed impact case studies to be assessed fairly and reliably?



were chosen to highlight assessment issues specific to disciplines,[39] other sub-panels chose case studies from another sub-panel or main panel. Some individuals noted that in instances where sample case studies (in particular those from other sub-panels) were not very similar to those which they went on to assess, the calibration process was challenging and of limited help in informing their scoring of impacts. On the other hand, one panellist commented that they felt that for future assessments it would be more helpful to have 'wider calibration exercises using a common set of case studies, rather than each calibration exercise being more or less contained within each [sub-panel]', on the grounds that 'this would help to ensure greater consistency in scoring and provide some statistical assurance of consistency between sub-panels'.

4.3. Panellists' perspectives on the calibration process

Results from the survey indicate that views on the impact calibration exercise were broadly positive, with no significant differences between main panels, impact assessors and sub-panel members (as illustrated in Figure 4.1).

Individuals in the focus group discussions also highlighted the value of the calibration exercise as it was thought to have given people confidence in how to assess the case studies and to have made the assessment process feel more straightforward. One panellist found the process one 'of the most valuable elements of the whole REF'. It was acknowledged that the process was time-consuming but that such in-depth discussions allowed panellists to really 'tease out things that perhaps you overlooked' and encouraged 'extremely intensive and careful consideration… of the application of the criteria as a whole in practice… until we were clear collectively that the process was absolutely fair and reliable'. Several individuals commented that they found the calibration exercise to be much more important in the assessment of impact case studies then the assessment of outputs. Panellists in areas of Main Panel B noted that they felt much more confident in the consistency of their scoring in the impact part of the exercise, compared to outputs, as a result.[40] The Main Panel A overview report highlights that panellists found 'that the calibration element of the exercise was the backbone of this evaluation process' and was an invaluable mechanism for ensuring consistent assessment of impact, as well as enabling real-time interchanges between the sub-panels via the main panel members (REF 2015). It was stated by several panellists that they found the calibration exercise of value and that they would have 'heavily benefitted from a deeper calibration exercise' as whilst it 'was a useful [exercise]… it was [not] as complete and comprehensive' as it could have been if they had had more time to dedicate to it. Another individual noted that having more explicit roles in the calibration process might help 'to make that a stronger more robust process which then could have fed through more easily to the moderation process'.

[39] Such as the impact period, the range of HEI types, the range of impact types and the broad range of evidence types.

[40] Some individuals stated that they would have benefited from additional training prior to the whole REF process 'on calibration, on understanding how to measure the differences [and] how you understand the pathway to impact better'. It was also suggested that a team building exercise earlier on in the process would have supported and encouraged 'trust, team cohesion' and an open calibration process.

4.4. Calibration raised awareness of different perspectives

There was a perception among those involved in the focus group discussions across main panels areas that the calibration exercise was beneficial in helping them appreciate other panellists' perspectives. Individuals noted that, at first, the sub-panels had different interpretations of impact and that these discussions allowed them to agree what was and what was not impact and to 'find a view of impact that both the academic and user members could sign up to'. The calibration exercise allowed discussion of different perspectives. As one panellist notes: 'differences [between academics and research users] resulted from differing expectation of impact for the sector'. In particular some highlighted the critical role of calibration in interpreting the level (1 to 4-star) used in the assessment. It was noted by one user that they found themselves unaware of the

> *…crucial nature of these fours and threes and twos [star levels], and [users were] just trying to interpret the words…'outstanding' [and] 'very significant' on really very arbitrary terms. We had interpreted [them] entirely differently to the academics and that first calibration stage was crucial in trying to bring us together.*

Not only did the calibration process help user members and academics to align their interpretations of the scoring system but also there was a shared perception across areas spanning all four main panels that the exercise helped panellists identify whether, on an individual level, they tended to mark more harshly or more leniently than others. Focus group participants from both academic and research user backgrounds provided anecdotal examples of research users scoring impact higher or lower than academics. Panellists in areas of Main Panel B referred to statistical analysis of the calibration exercises, which they found helped them to identify their scoring position (as individual panellists and as a sub-panel) and helped them to 'moderate our position if necessary'. However, whilst some panellists said that this awareness of their own scoring behaviour led them to moderate their scoring, others commented that they did not let this influence their scoring behaviour.

4.5. Challenges with the calibration process

Through the focus groups, several panellists from areas of Main Panels B and C said that they noticed that sub-panels seemed to be comfortable with ranking the case studies but were less sure of where to draw the scoring boundaries. Several panellists framed it as two elements to calibration: consistency and level. It was felt that the former was easier to achieve than the latter. One panellist said that at one point

> *Some had drawn grade boundaries on slightly different interpretations of the criteria than others which is why we went through a second audit exercise within the whole main panel to try to make sure we were getting consistency of judgment in the drawing of grade boundaries across the main panel.*

In areas of Main Panels B and D in particular it was noted that whilst some panellists found it easy to assess impact when it was clearly 4 star or 1 star, they found it harder to assess the middle bands (2 and 3-star), especially in regards to reach and significance. It was thought that the small sample which was calibrated included high scoring case studies and low scoring case studies but did not allow panel members to 'examine some of the more nuances around…what may be a very good 3-star impact case study'. It was agreed that a broader set of calibration case studies would have been valuable to work through so that panellists did not only hone their scoring on a certain level. In areas of Main Panel A, they used a conceptual framework of up to 8 stars when framing their discussions. When awarding scores, case studies scoring from 4 to 8 were all graded as 4-star. Panellists felt that this helped them 'to get the calibration right, rather than setting the very best ones as a four, which would have made three a severe judgement. This encouraged us to think that there were lots of 4s'. As stated in the panel's overview report, 'the 4-star category contained some truly exceptional impacts in reach and significance, whereas others still worthy of 4-star, were somewhat less exceptional but nevertheless outstanding' (REF 2015 for Main Panel A).

It was also noted by one individual that academic sub-panel members seemed to approach the calibration exercise with expectations of where things would be due to their experience in assessing outputs and 'the sort of percentages that there were going to be in the different categories'. Linked to this, some thought that in future assessments, in order 'to get a result that is going to be satisfactory you have to say we expect 20 per cent 4s

30 per cent 3s and whatever and then you know that when you are marking'. An additional challenge with the calibration process was that there were a few cases where impact assessors did not attend the calibration exercise meeting which in turn meant that 'they generated scores that were significantly different from those who had been involved in the calibration [and] this rendered their involvement more or less useless'.

Chapter 5 Review of impact case studies and individual scoring

> *I found that I could make consistent and coherent assessments of case studies that had a wide range of different types of impact.*

> *Where there was a clear articulation of impact regarding the type of impact made and this was backed up with evidence, the case study was relatively easy to assess.*

Having been allocated, impact case studies and impact templates were then reviewed by sub-panel members and impact assessors. Every impact document was reviewed by at least two reviewers and the aim was to obtain an academic and research user perspective on each assessment. Each panellist read the submission independently before discussing and comparing their views. After calibration had taken place, panellists then reviewed all impact case studies and impact templates allocated to them and suggested a score (giving appropriate commentary if they felt that the document did not meet any threshold criteria). Scores and comments were all logged in a central spreadsheet.

5.1. Score levels

The impact documents were graded on a 4-star scale. In the focus groups the majority of sub-panels across all main panels mentioned that in practice they used more granular scales, by using ½-stars. There was concern raised regarding the challenge of scoring case studies that fell on a scoring boundary, since whether a case study was scored 3 or 4 stars could make a big difference to the outcome for the HEI. Others highlighted the difficulties in assessing case studies in the middle levels (2 and 3-star), the cusp of which was the threshold for funding. We cannot comment on scores allocated by individual panellists as the data was not collected in a standard fashion across sub-panels to allow a comparison of scores before and after discussion with others.

5.2. Rules and guidance

The survey captured a range of information on panellists' ability to apply the rules and guidance provided to support this process. Broadly panellists were happy with what had been provided, both in terms of assessing eligibility (Figure 5.1) and in the wider assessment of case studies (Figure 5.2).[41] The results were broadly consistent between panels, though viewpoints in Main Panel A in particular were consistently positive by comparison to other panels. Of all scores given across the questions summarised in Figures 5.1 and 5.2, panellists from Main Panel A members gave the highest score (4-star) 45 per cent of the time, compared to 35 per cent across all panels. There was agreement across academics and users that the rules allowed for subjectivity but no consensus on whether this was a good or bad thing.

Panellists in the focus groups in general thought that rules were clear and that they were able to apply them. However, on occasion institutions had still failed to follow them. One panellist wished to highlight that whilst changes could be made to the rules, the guidance and the template, what was crucial and key to the assessment process was the discussions that took place. It was this that brought a robustness to the scoring process. To improve the template, one individual suggested that perhaps some 'must do' bullet points could be introduced, so that details such as an academics' dates at an institution were immediately clear. This point was also raised in the panellist interviews with suggestions that a 'lighter touch' template with 'tick boxes' for standard

[41] This data is presented broken down by type of panellist in the following sections.

Figure 5.1: Extent to which survey respondents felt the rules and guidance enabled them to determine the eligibility requirements for impact case studies[42]

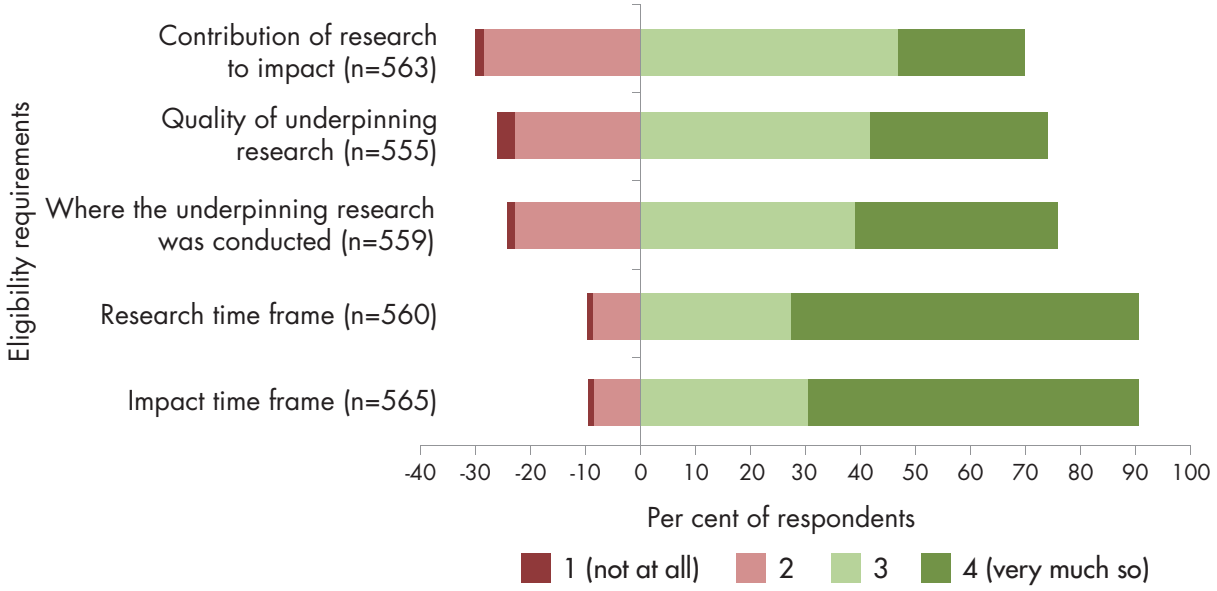
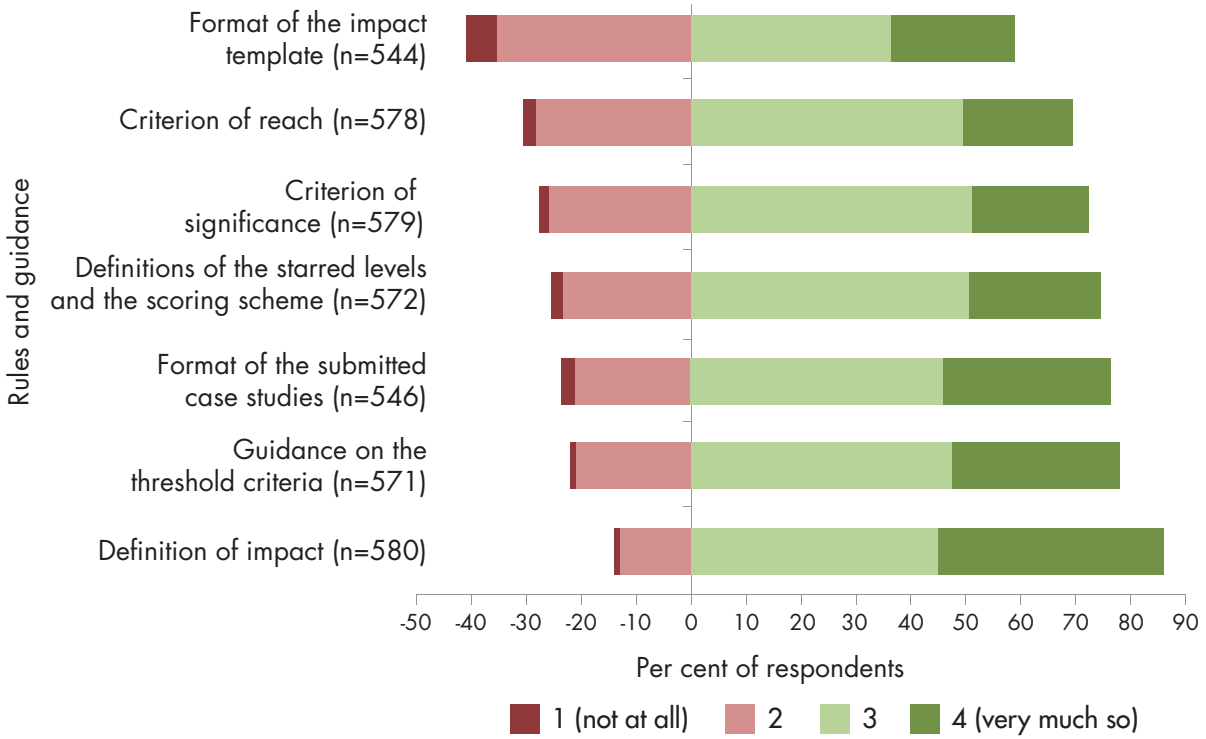


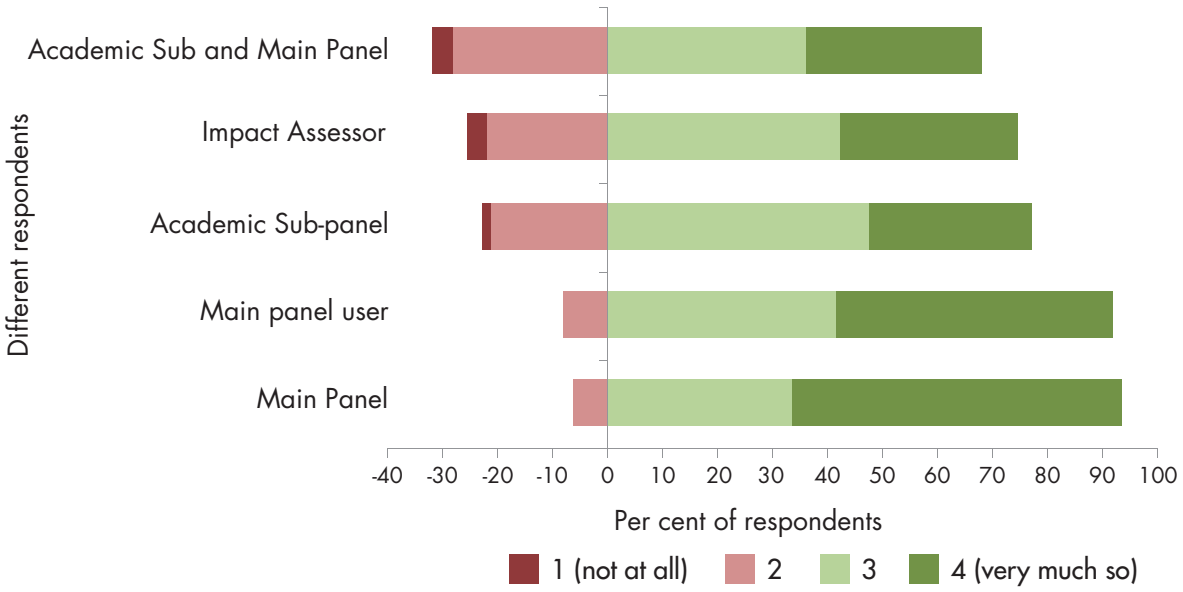
Figure 5.2: Extent to which survey respondents felt that the rules and guidance enabled them to assess impact case studies fairly and reliably[43]



[42] Full question: To what extent did the rules and guidance you received allow you to determine the following eligibility requirements for case studies fairly and reliably?

[43] Full question: To what extent did the following aspects of the rules and guidance allow you to assess impact case studies fairly and reliably?

Figure 5.3: Extent to which survey respondents felt the format of the case study template enabled them to assess impact case studies fairly and reliably[44]



information would reduce the number of audit queries being made. It was also suggested that further guidance on what to include would be helpful.

In the survey data, concerns were typically heightened (i.e. scores were lower) amongst those groups directly involved in conducting the assessment of case studies and templates (sub-panel members and impact assessors). Of concern was the format of the case studies (see Figure 5.3), assessment of the quality of the underpinning research (see Figure 5.5), and, particularly for sub-panel members, assessment of the contribution of research to impact (Figure 5.6), and the criterion of reach (see Figure 5.4). These concerns are discussed in more detail below. It is important to note that these elements were of less concern to panellists than the impact template, discussed in Chapter 6.

5.2.1. Format of the case study (REF3b) template

The format of the case studies was considered to support a fair and reliable assessment of impact by 76 per cent of survey respondents overall. However, there were significant subgroups who felt that the format did not contribute to fair and reliable assessment, giving a score of 1 or 2 in response to this question (see Figure 5.3). In particular, 26 per cent of impact assessors, 23 per cent of academic sub-panel members, and 32 per cent of academic sub-panel and main panel members expressed this concern.

There was a concern from some that the format of the impact case study template channelled 'linear thinking'. One individual noted the value of having a template which allowed so much freedom, as they were concerned that a more prescriptive template could introduce the risk that institutions learn how to 'game play' and score more highly.

There were frustrations from some that rules on submission formats and indicative word and page limits were 'frequently bent or broken but not enforced'. This was felt to disadvantage those submissions that had complied with the guidance, and panellists should be provided with specific guidance on how to deal with this in future.

5.2.2. Definition of impact and types of impact

Although there was agreement that there should be scope for a wide range of impact, there was disagreement between focus group participants over whether that had been adequately achieved. In general panellists felt able to judge different types of impact (backed up by the pre-focus group survey, data not shown). However, panellists agreed that the ease of scoring

[44] Full question: To what extent did the following aspects of the rules and guidance allow you to assess impact case studies fairly and reliably?

varied according to the type of impact, for example impacts on policy, public engagement and 'negative impact'[45] were highlighted by academic and user focus group attendees. However, it is important to note that although these were harder they were still able to make judgements and were conscious that all types of impact should be assessed equally. Panellists stressed that they saw fewer of these types of impact than expected. They hypothesised that this was because they were difficult to quantify and therefore HEIs steered away from submitting case studies in these areas. The examples of types of impact provided in the guidance were thought to be useful. Some panellists requested more examples in the areas mentioned above and examples of case studies scoring different levels. Panellists from areas of Main Panel A stressed the use of proxy measures in assessing impact. For example, what was the impact of an article in a national newspaper, or the impact of a patent or the creation of a spin-out on patient health?

5.2.3. Criteria of reach and significance

There was qualified confidence from the focus groups in being able to assess reach and significance consistently in the majority of cases, provided there was adequate evidence. One user said that Main Panel C had been determined to keep definitions as open as possible, although another questioned whether the generic nature meant they had 'no impact on the assessment'. There was no consensus around the descriptions of the criteria provided, with some stating it was helpful, and others feeling there was too much subjectivity. One individual highlighted that a consequence of this was that during calibration a lot of time was spent establishing what reach and significance meant. For further details on calibration, see Chapter 4.

Relationship between reach and significance

Some 28 per cent of survey respondents felt that the criterion of significance did not contribute to a fair and reliable assessment of impact (scoring 1 or 2 in response to this question); a similar proportion, 30 per cent, expressed this concern for the criterion of reach (Figure 5.4). In both cases, impact assessors and academic sub-panel members were the groups expressing the most concern around these criteria. Some 30 per cent of impact assessors and 32 per cent of sub-panel members felt that the criterion of reach did not contribute to a fair and reliable assessment of impact. For the criterion of significance these percentages were 34 per cent for impact assessors and 29 per cent for sub-panel members.

There was a lack of consensus on whether reach and significance could, or should, be assessed together or independently. Academics across the areas spanning all four main panels thought that reach and significance were not independent or even particularly different from one another, and were keen that they continued not to be scored separately. Therefore, they scored case studies holistically rather than just giving two scores.

On the other hand, several individuals commented that they were not sure if both reach *and* significance were necessary, and that it was hard to compare case studies that had one, the other or both; one academic said that either was acceptable ('apples are as important as oranges') and that they had identified examples of each to use as benchmarks in future cases.

The use of geography in determining reach

There was generally agreement from academics and users in areas across all main panels areas that local impact did not mean limited reach and that international impact did not automatically indicate maximum reach. However, there was some disagreement over whether this was applied in practice, and a number of comments suggested that it was generally harder to assess impact in local cases.

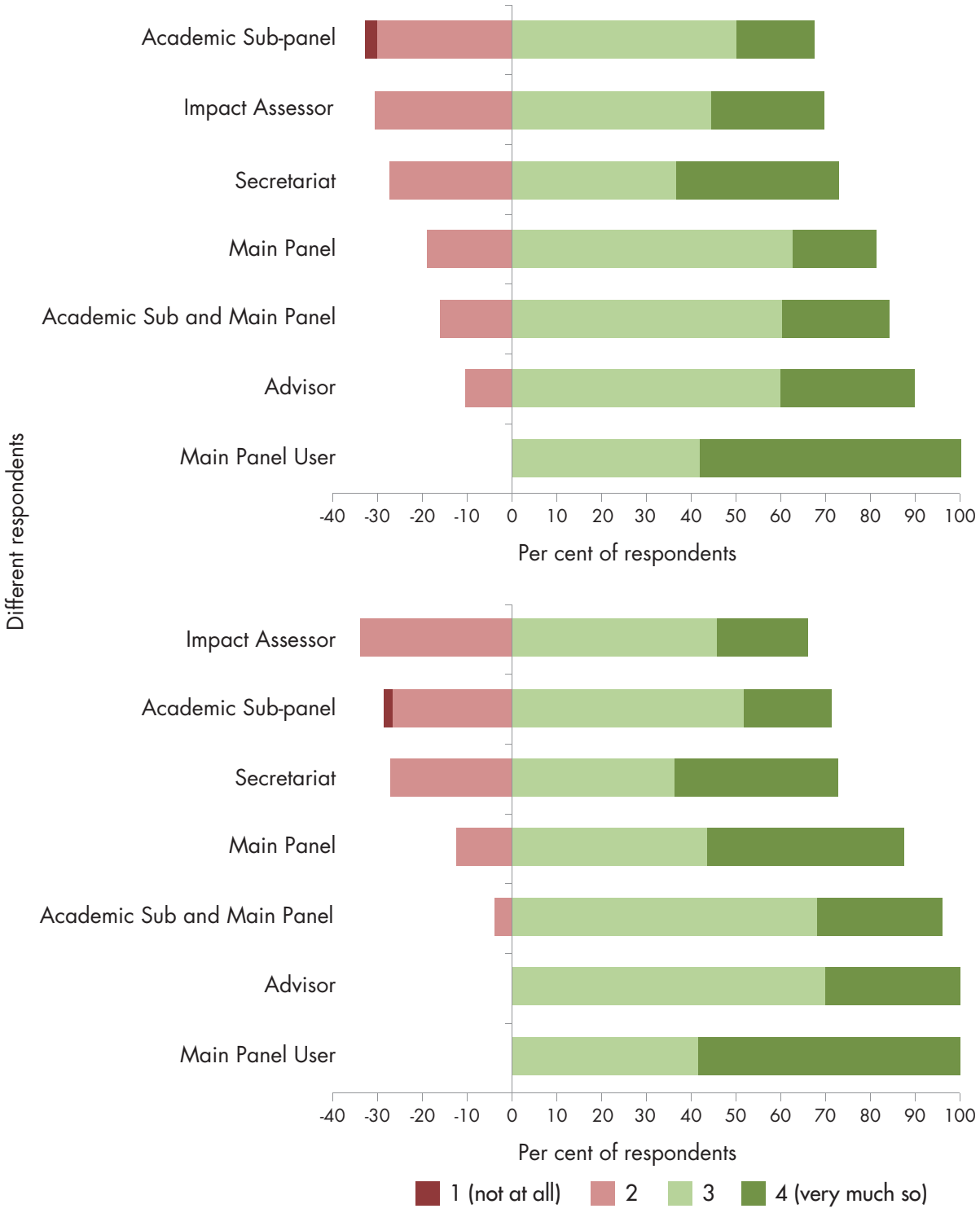
In the focus group there was acknowledgement of the difficulties in defining reach in a way that would apply universally. For example, one person thought that the definition of impact as affecting behaviour was not helpful given that sizes of television audiences were used as examples of reach, and another academic felt that assessing reach was difficult in business and management, as innovation was generally confidential.

Types of impact

Where several impacts of differing significance were claimed within one case study it was unclear how to score it. Some academics thought that there should be more than just reach and significance to judge some types of impact, such as an impact of public discourse. Individuals generally found it difficult to compare reach and significance in case studies with different types of impact, although several people were nevertheless confident that they had done so consistently.

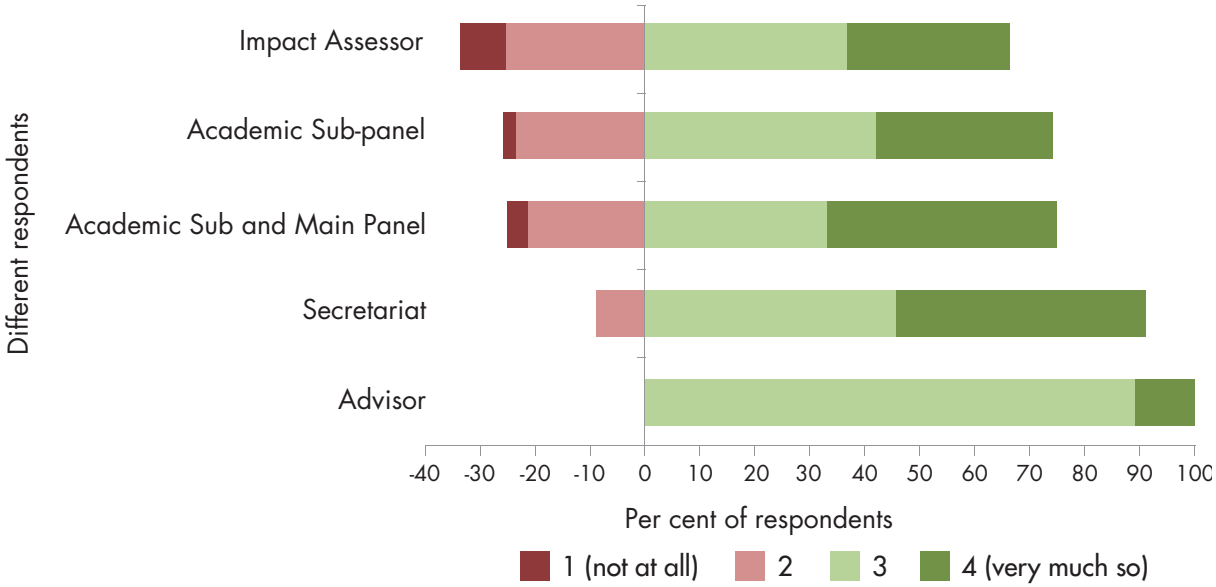
[45] 'Negative impacts' refer to where something was not implemented or was stopped as a result of the research.

Figure 5.4: Extent to which survey respondents felt the criteria of reach (above) and significance (below) enabled them to assess impact case studies fairly and reliably[46]



[46] Full question: To what extent did the criteria of reach/significance allow you to assess impact fairly and reliably?

Figure 5.5: Extent to which survey respondents felt the rules and guidance enabled them to determine the quality of underpinning research for impact case studies[47]



5.2.4. Quality of underpinning research

Approximately 25 per cent of sub-panel members and impact assessors found it difficult to rate the criterion of the quality of underpinning research, in their assessment of impact case studies (Figure 5.5).

Focus group attendees noted that assessing whether work met the 2-star threshold created a lot of challenges, particularly where it spanned disciplines. One research user (from an area within Main Panel B) thought there was a reluctance to judge research as unclassified.

Individuals at the focus groups felt that sub-panels varied on how strict they were on applying the 2-star threshold to underlying research; from limited assessment of the outputs to being 'quite strict'. To assess the quality of the underpinning research panellists employed different methods: several said that they only checked borderline cases; another academic said that they allocated all underlying research to be assessed across sub-panel members; other academics said that they used proxy measures such as impact factors of journals or funding.[48, 49] One panellist raised an example of where research had been assessed both as an output and as underlying research for a case study, and the two groups had reached different conclusions as to whether it was 2 star or not, which was a concern for consistency. Users often left the decision to the academics who were familiar with such assessment through the outputs element of the REF. Although there were anecdotal examples of wasted effort due to the assessment occurring in parallel, where the research user had assessed a case study the academic deemed as ineligible.

Several panellists thought that too much emphasis was put on the importance of the underlying research being 2-star. Most who expressed an opinion disapproved of the existence of the 2-star threshold for a variety of reasons. Several academics across areas covering all Main Panels agreed that 2 star was a surprising threshold given that it was lower than the threshold to receive funding for outputs, and that national recognition – i.e. 1-star - should be sufficient, or the question should just be whether or not it was research. On the other hand, others thought that it was good to have a 2-star threshold. Across areas of Main Panel A there was a discrepancy on whether 'predominantly' defined more than one publication, or whether 2-star quality of the main output was sufficient.

[47] Full question: To what extent did the rules and guidance you received allow you to determine the quality of the underpinning research for case studies fairly and reliably?

[48] It is important to note that this practice is not used for the assessment of outputs (REF 2012a, page 8, paragraph 53).

[49] It is interesting to note that only 11 per cent of audit queries were raised with regard to the underpinning research. However, 25 per cent of these were subsequently unclassified.

In the focus groups with academics from Main Panel A areas, it was recognised that REF includes systematic reviews as underpinning research. However, there was a lack of clarity on whether this synthesis also included review articles. There was also a concern where guidance cited a systematic review rather than the clinical trial it described. In these instances, panellists were concerned the trial may struggle to evidence their part in the impact as the audit trail is not available through citations.

5.2.5. Institutional ownership of impact

Many people in our focus groups, across sub-panels spanning the four main panels, thought that institutions did not make clear where researchers had been in regards to their institutional affiliations. These eligibility issues were time consuming to check, and one panellist suggested there could be a section specifically asking the roles of researchers so that the eligibility of the research was explicit.

The rules stated that for impact to be eligible 'the underpinning research must have been produced by the submitting HEI'.[50] This was different for the rules for outputs, where publications can be claimed by the institution current employing a given individual. Individuals from areas within Main Panels B, C, and D in the focus groups highlighted that they felt that this was inconsistent.

In particular, a number of academic sub-panel members felt that both should move with the individual researcher. However, their difficulties were with the implications of the rule for the sector, rather than the impact of this on their ability to assess the submissions (this is reflected in the fact that when asked research users had much less of a problem with this rule). There was a concern that there was no incentive for HEIs to disclose important relevant pieces of work if they had been done at a different institution, and that there was a risk to institutions in hiring staff who could not 'bring their impact with them'.

On the other hand, it was recognised that linking impact would accentuate the effect of recruitment during the submission period and could have a large effect on a submission if an employee who was central to a case study moved institutions whilst the submission was being put together.

5.2.6. Contribution of research to impact

> *We felt the question was whether the work was necessary – it didn't have to be sufficient.*

The rules stated that an impact case study should 'explain how that research underpinned or contributed to the impact'.[51] There was a lot of discussion in the focus groups about the link between underpinning research and impact, and the need to articulate this link in the impact case studies. Most but not all panellists found it relatively difficult to verify the link and there was agreement that links between research and impact were sometimes more tenuous than they would have liked. For example, one panellist felt that testimonials were 'often vital to prove that a change had been caused by a piece of research'. There was a concern that institutions were not given sufficient guidance on how they were expected to demonstrate the link between research and impact, and that this should be clarified and expanded on for subsequent assessments. Panellists raised examples of where a body of research had led to the reputation of a researcher enabling them to work with and advise public bodies and policymakers.

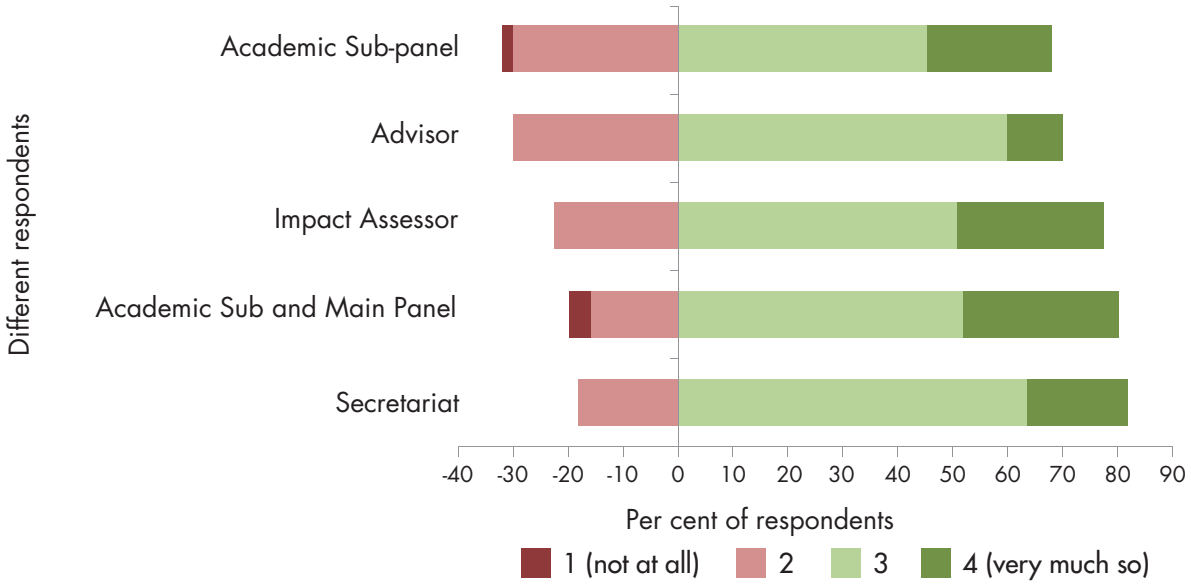
The relationship between research and impact was a particular issue in areas such as public engagement, where the link was to a 'high-impact individual' rather than a particular piece of work. Focus group participants from all main panel areas referred to individuals with high media profiles and questioned the impact that they could claim, as a lot of the work they disseminate is not their primary research.

There was much debate over the importance of the size of a particular researcher's contribution to impact, particularly for multi-institutional and multi-disciplinary research. Individuals from Main Panel A areas felt that they had been strict about ignoring the size of the contribution, but were not sure if this had been uniformly adopted across other main panels. There was some concern that this might have led to minor contributions receiving too much credit for their part in an impact. For example, if you contributed to stopping people smoking this was an impact. However, the rules did not allow for differentiation on the level of contribution to achieving that impact. To address this, academics in areas of Main Panel A suggested that as well as reach and significance, case studies should be scored

[50] REF 2012a.

[51] REF 2012a.

Figure 5.6: Extent to which survey respondents felt the rules and guidance enabled them to determine the contribution of research to impact for impact case studies[52]



on the dimension of contribution (scoring 4A, 4B, 4C, 4D). However, elsewhere there was concern that this could disadvantage cross-discipline and institution submissions and therefore discourage collaboration

This links to the discussion of the role of researchers in achieving impact. Various contexts within which impact occurred were presented through the case studies, for example where researchers took steps to ensure impact, where research was sponsored by or designed with industry, and where the researcher had no onwards engagement in the uptake of research. This spectrum of engagement led panellists to questioning whether intended and unintended impact should be rewarded equally. For example, a company or the government might choose not to implement something due to financial constraints or wider political factors and contextual considerations rather than as a result of the quality of the research. As one panellist stated: '[the research] might be lucky and fall on fertile ground but [it] may fall on stony ground and I'm not sure that the researchers themselves are responsible for the terrain they find externally'. As one panellist said: 'we are rewarding [HEIs] for discoveries that led to impact whether or not there was an effort to create that impact'.

5.2.7. Use of corroborating evidence

The role of supporting evidence was important in confirming the impact presented to panellists. On occasions where claims were made without supporting evidence panellists told us that these were marked down.

The limited access to corroborating evidence was a general concern to those attending the focus groups. There was support for the idea of requiring institutions to submit all corroborating evidence to a central repository. In some instances, panellists felt supporting evidence had been poorly presented and was difficult to obtain.

There were suggestions from panellists from areas of Main Panel A and D that more standardised numerical measures could be used (e.g. quality adjusted life years (QALYs), patents, industrial income, audience appreciation index etc.). These were included in some case studies but panellists felt they would have been helpful to have in all case studies describing a comparable type of impact. However, there was an understanding that that the same metrics may not be applicable across main panels and might not work for all disciplines. For example, a subset of research users were concerned about measures when claiming an impact involving interaction with the public. One panellist questioned: 'What is the right number of website hits to become 4-star? How many people do you have to get to your lectures to be 4-star?

[52] Full question: To what extent did the rules and guidance you received allow you to determine the contribution of research to impact for case studies fairly and reliably?

There was confusion amongst focus group attendees as to the level of corroborating evidence that could be accessed without raising an audit query. Some used the web links contained within the submitted documents or even carried out Google searches, whereas others based their assessment only on the information provided in the case study document. One panellist said: 'We were steered back to focusing on case studies on the face of what was difficult because we weren't sure if we were getting the full picture'.

There were a variety of opinions on the importance of testimonials. For example, research users stressed it was essential to have the beneficiaries' voice and feedback on how it had impacted on them. In particular, Area studies (UOA27) highlighted the role of country-specific experts in distinguishing the value of testimonials, depending on the context and the location. The inclusion of a quote from the testimonial in the case study was seen to be helpful in supporting statements made. Although another panellist stressed that you had to be careful in interpretation as the quote did not always reflect the rest of the testimonial.

Chapter 6 Review of impact templates

> *It felt more aspirational than anything based on reality.*

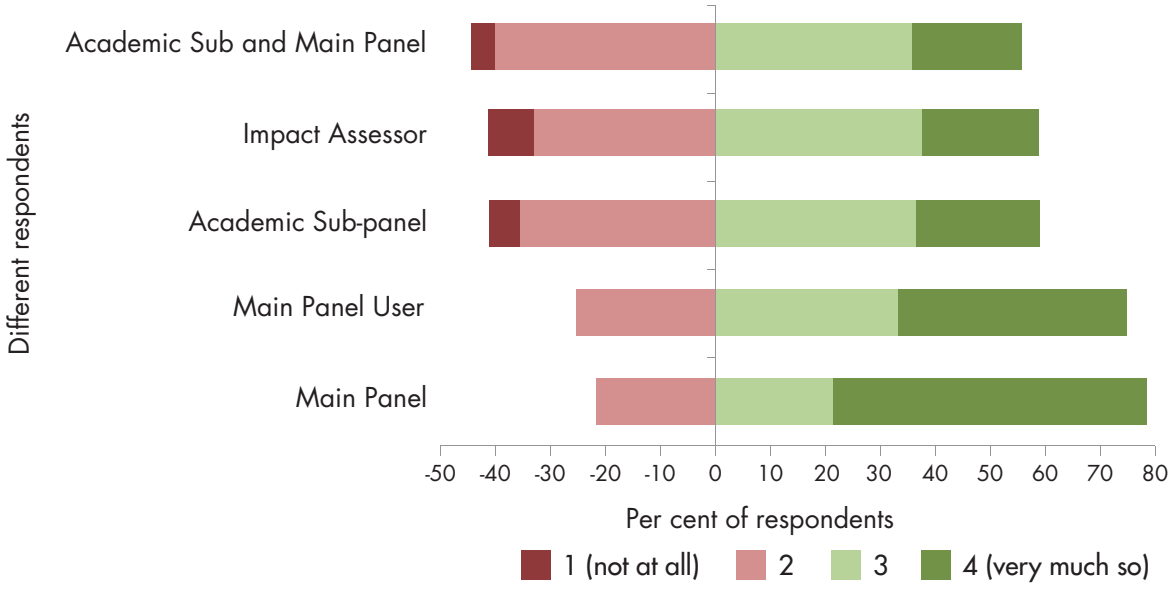
As described above, panellists were broadly positive about their ability to conduct a fair and reliable assessment of impact, particularly with respect to impact case studies. This is reflected in both the survey data (shown previously in Figure 5.2) and discussions in the focus groups and interviews. However, the assessment of the impact templates was of greater concern in focus groups and interviews, perhaps reflecting the concerns about the limitations of the format of the impact template described in the survey. For example, one focus group respondent highlighted that the submissions 'reflected the wooliness in the construction of the template' and that clearer guidance could help HEIs' focus.

When explored by type of respondent, as with other rules and guidance, those assessing the impact submissions (sub-panel members and impact assessors) were more negative than those overseeing the process from the main panels (see Figure 6.1).

In the focus groups, the particular criticisms discussed were that the scoring criteria were not clear, and the template and associated guidance did not help institutions structure their responses appropriately. In particular, they felt that the 'Relationship to case studies' section (Section d) was completed poorly.

Individuals also noted that there was a great deal of repetition from the case studies especially in this final section of the template. It was noted by one panellist that they thought the section in the impact template regarding the relationship to case studies was going

Figure 6.1: Views of different respondent groups on the format of the impact template[53]



[53] Full question: To what extent did the format of the impact template allow you to assess impact fairly and reliably?

to be very useful, however it was 'uniformly the least informative' and it was suggested that there could have been further guidance to prompt people to reflect on the context of the case studies. It was also suggested that the assessment criteria in the case studies could be used to structure the impact templates. One concern raised was that the impact template could discuss institutional support that is being given to new groups without there being a case study to evidence the claims.

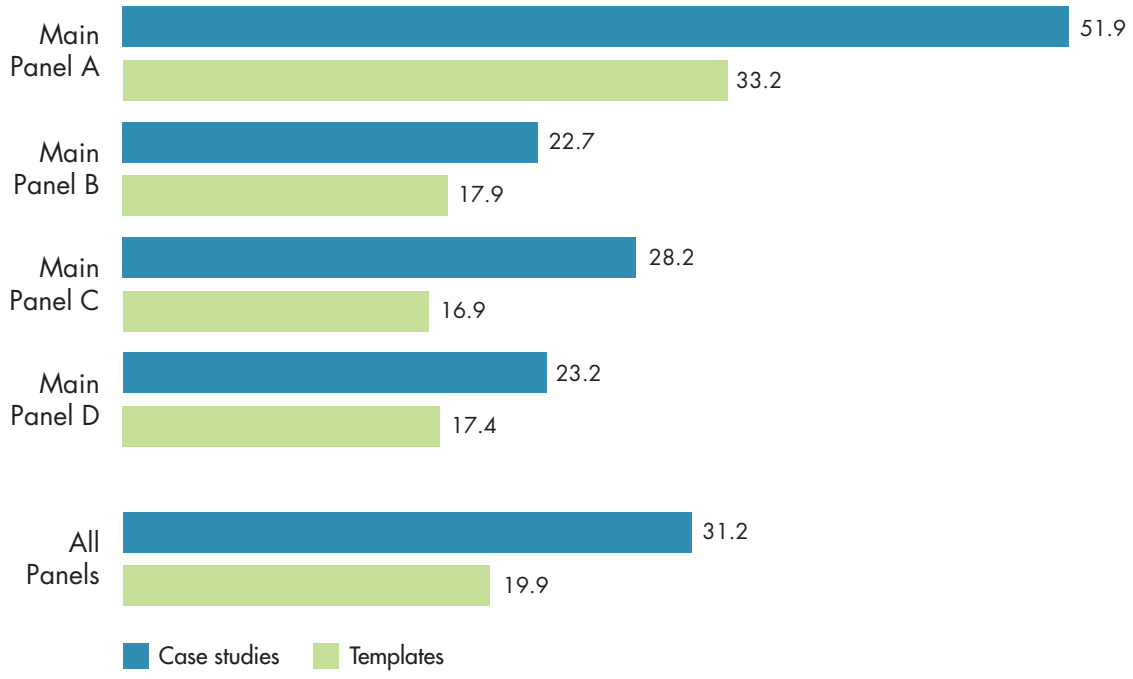
In the focus groups there were diverging views on the value and importance of the impact template within the assessment process. Some panellists felt that the impact template was a 'vital document' and could help institutions to 'craft their whole submission for impact case studies'. In particular, they felt that the impact template really helped to better understand what institutions were doing to support staff and what processes they had in place to encourage and support impact and facilitate engagement with partners, collaborators and industry.

Several spoke of the value that the impact template added as part of the REF submission. Individuals in areas of Main Panels A and B noted that its very presence as part of the process prompted HEIs to think about an impact strategy which positioned them to think about the future and their priorities. Some users and academics stressed it was 'potentially immensely useful to institutions' and had a 'hugely pivotal role' in crafting the whole submission.

On the other hand, many individuals in the focus groups spoke of the 'fairy tale'-like nature of the impact templates, which they felt to be 'a cosmetic exercise'. One panellist referred to it as more 'aspirational than anything based on reality'. It was felt that whilst the information the template provided was good to have, there was no way of verifying claims that were made without having a site visit, and there was no confidence that the impact template reflected reality. One individual said that the template encouraged plans for the future that could be 'a figment of the [template author's] imagination' which made them uneasy when they came to their assessment.

Panellists commented that the impact templates were often of lower quality than the case studies. This is reflected in the scores awarded and there are consistently a lower proportion of 4-star impact templates in each Main Panel than 4-star case studies. This was thought to have sometimes undermined good case studies, lowering the aggregated score and thus impacting the final sub-profile a submission received for impact. It was noted by some individuals in areas of Main Panels B and C that this distinction might not be clear to institutions provided with a combined score for the impact documents, unless explicitly stated in feedback.

Figure 6.2: Percentage of 4-star case studies and templates, by main panel



6.1. Importance of the quality of writing in the assessment of the impact template

As the template's content was not grounded in metrics, many panellists across the areas of all four main panels also commented that they felt that a good template depended on the skill of the writer. Some mentioned discussions within their sub-panel regarding whether panellists were to some extent marking the writer and unless written badly, panellists found it hard to discriminate between scores. As long as texts were well written, panellists believed that most HEIs could write a 4-star impact template. Related to this, there was concern from areas across all main panels that the use of external writers in some submissions led to inequality in the process and unevenness in the system.

The reasons provided for lower scores included a lack of understanding of what to include in the impact template. One panellist said that there was 'too much scope for things to appear in all sorts of different places and it was like trying to grab a bar of wet soap, it really was'. There was a call from areas of all main panels for clearer guidance on what and how much to enter into each section and also how these templates would be assessed.

The variation of the quality and clarity of writing and the wider presentation was also raised with regard to case studies. Some panellists were aware that presentation affected their assessment of the impact. One stated: 'there was a frustration in some cases because in the substance of the piece there was actually something important but it may not have been presented in such a way that you could draw it out very easily'.

6.2. Other issues reported with assessing the impact template

An issue raised by a number of users in our focus groups was that HEIs did 'not seem familiar with writing strategies which referenced overall aims, objectives, trajectories milestones and evaluation'. Another issue was the retrospective nature of the REF 2014 assessment. Users noted that the impact template was hard to assess as many HEIs did not appear to have an impact strategy in place in 2008 (at the beginning of the assessment period) thus introducing confusion in presentation between retrospectively putting a strategy in place or writing about one for the future. Linked to this, a minority of panellists commented there was confusion in submissions and the guidance between whether to present the unit's perspective or an institutional perspective (at the HEI level) on procedures and infrastructure in place.

6.3. Options for the future

There was general agreement across areas spanning all main panels (in our focus groups, interviews and survey, and including users and academics) that it would be beneficial and sensible **to integrate the impact template within the environment template,** with opinion being that the impact strategy cannot be disaggregated from the research strategy. One individual suggested that the template might have been appropriate for the first exercise, since impact was new and the template was about intentions, but in future exercises impact should be assessed on the basis of what had been achieved.

It was thought that this would negate the need to repeat so much information and that one template which addressed how the research and impact environments are inter-related would be useful. Several individuals in areas of Main Panel D thought that there could be a distinct impact element in the environment template that was scored separately. It was suggested that careful labelling of sections 'the environment in which research is undertaken' and 'the environment in which impact is supported' would help to structure the combined template. However, there were concerns that combining the impact and environment templates would disadvantage small submissions that would be solely dependent on their case studies, and one individual stated that they liked the fact that the environment and impact templates were separate from one another.

Several panellists felt that whilst it would be possible to combine the impact template and the environment template together there could be a risk in both elements that this discursive approach could be 'susceptible to spin'. Instead they suggested the need to be more strategic in the approach to impact and providing support for impact – an approach which **captures key empirical information** and could be framed in a discursive context.

There was also discussion that perhaps a series of closed questions is required instead of the discursive nature of the current template. However, it was thought by one individual that devising a set of factual questions answerable across a broad range of impacts could be very challenging.

One individual questioned the need for impact templates, since if an institution has an encouraging and

Table 6.1: Classification used for size of submissions

Size of submission	Number of case studies	Sample size (number of submissions)
Small	2	749 (41%)
Medium	3–5	807 (44%
Large	6 or more	268 (15%)

supportive environment for impact then this will be evident from the submission of good case studies. Others noted that the majority of the impact case studies submitted for REF 2014 were historically based (before institutions would have had impact strategies) and yet they still bore impact. With this in mind they wondered whether the impact template (which demands detail of an impact strategy) was simply another requirement that might detract away from letting impact happen in a more organic environment.

6.4. Relationship between impact case studies and template scores

In assessing the relationship between case studies and impact template scores, we categorised institutional submissions by size (see Table 6.1).

Figure 6.3 illustrates that there is a higher correlation between impact case studies and template scores in smaller submissions, (i.e. there is a smaller difference between case studies and template scores in smaller submissions). It is not clear from our analysis what has caused this correlation.

There was some discussion in the focus groups regarding when it was best to assess the impact case studies and templates. One individual noted that the timing of the impact templates needs careful thought; they argued that the template assessment should take place after all the case studies had been evaluated, so that one informs the other. This individual then also queried whether it may be better to assess impact templates at the same time as environment templates. However, one individual disagreed and could not see what was to be gained by assessing the case studies with the template. They thought that they would be able to make a fairer assessment of the case studies without the impact template.

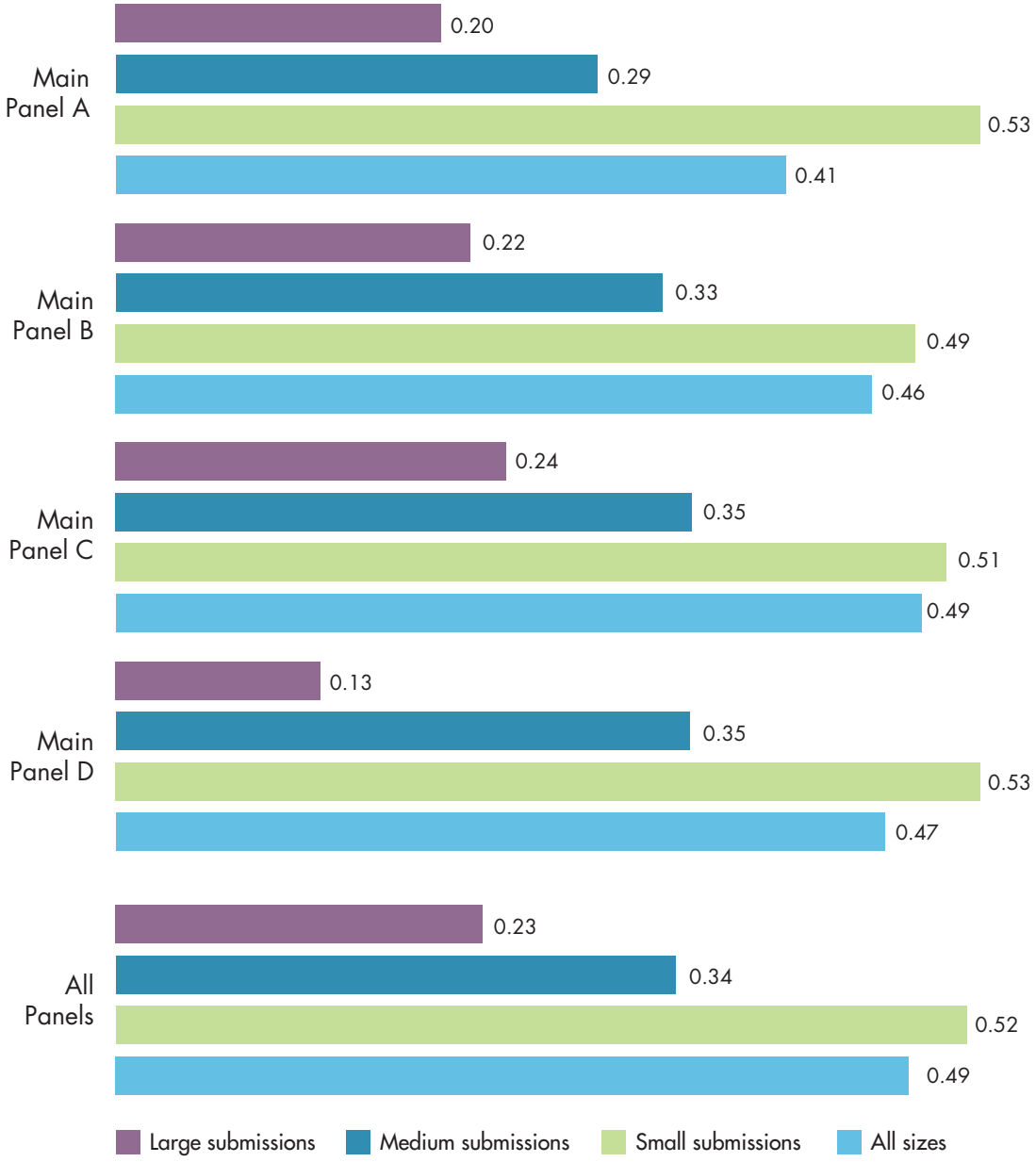
Two individuals thought that it was unrealistic to expect a link between the impact templates and the case studies as the time frame between the two could be decades apart from one another.

It was noted that some panellists did not see the case studies before awarding their final template scores and that even those writing the unit feedback did not see them until the assessment was complete.[54] Individuals stated that they often read the impact templates separately from the case studies, which meant that panellists were unaware of repetition and links between the two documents. Others who did assess (or read) both documents felt that the link between the two wasn't clear and at times was just a single sentence. One individual suggested that there could be clearer coherence between the two documents if the structure of the impact templates reflected the case study assessment criteria.

One panellist commented on the 'strange divergence' between the impact template (which details the procedures supporting research impact) and the case studies (where there is no value accredited to detailing the process supporting impact). Another panellist noted that there was very often nothing to link the two documents other than a sentence. It was noted that there was no requirement for the case studies and the impact templates to 'speak creatively to one another' and that panellists would have liked to see a link between the mechanisms described in the template and how they supported the generation of material evidenced in the case study.

[54] From reviewing a sample of 5 per cent of case studies, we know that over 98 per cent of templates were reviewed by someone who reviewed the case studies, but we do not know what point in the process the template review was conducted, with respect to the case study review.

Figure 6.3: Correlation between the scores for a submission's impact case studies and template, by main panel and submission size[55]



[55] All correlations are statistically significant at the 0.01 significance level ($p<0.01$).

Chapter 7 Scoring agreement and moderation

> *It was like a real intellectual threshing floor, where things were just brought and threshed through, and we ended up with a decent crop.*

As described in Chapter 5, prior to carrying out the assessment each main panel and its sub-panels undertook an initial calibration exercise. After this calibration process they continued with an on-going programme of moderation throughout the assessment phase to continuously review scoring and grading patterns. As one focus group attendee put it: this process helped in 'refining the scoring through discussions between readers in advance of and at meetings, and difficult studies were discussed with the full [sub]-panel'. The moderation programme was conducted by sub-panel chairs and their deputies and discussed at main panel meetings regularly.

There was a high level of satisfaction overall with the moderation process within sub-panels (see Figure 7.1). Respondents indicated that in general they felt all panel members and impact assessors were listened to equally, that the process was transparent and clear and that criteria and guidance were equally applied in every case. There was slightly more concern around the efficiency of the moderation process, with just under 20 per cent of respondents expressing some level of dissatisfaction. Levels of dissatisfaction with the efficiency of the process were highest among the secretariat (36 per cent scoring 1 or 2 on this question), though the largest proportion of respondents that was very dissatisfied (score of 1 on this question) was amongst the academic sub and main panel members (9 per cent) (see Figure 7.2). There were no major differences in viewpoints between main panels, though there was a slightly higher level of concern around transparency and clarity of the process, and the equal application of criteria and guidance among respondents from areas of Main Panel C. Some 15 per cent of respondents linked to Main Panel C suggested that they were dissatisfied with the transparency and clarity of the process, compared to 10 per cent of all respondents, and 21 per cent of respondents in Main Panel C areas suggested they were dissatisfied with the equal application of criteria and guidance, compared to 14 per cent of all respondents.

7.1. Use of clusters to assess impact submissions

Across all sub-panels, clusters of sub-panel members and impact assessors were used as the initial point for discussing a case study or template and agreeing on a score. The size of the clusters reported in the focus groups ranged from two to eight individuals in different sub-panels. In addition, the composition of the clusters varied across sub-panels, in line with the variations in allocations (see Section 3.1). Some individuals highlighted the benefit of mixing up the clusters to ensure that discussion happened across the whole sub-panel, but this practice was not standardised even within a main panel area. It was agreed by many that the clustering of groups worked very well as every case study was marked multiple times from different area studies and scores could be reviewed again in the clustering.

In Earth systems and environment science (UOA7), they brought in an additional individual from the sub-panel to assist clusters in coming to an agreement where opinions were strong and wide ranging. Chemistry (UOA8) in particular noted that the breadth of the skillset on the sub-panel was useful when it came to resolving issues through the use of expert input beyond the cluster.

7.1.1. Communication within clusters

In between official meetings, panellists communicated online via email and Skype and by phone. It was suggested by individuals in areas of Main Panel D that

Figure 7.1: Perceptions on the process of agreeing scoring and the management of disagreements[56]

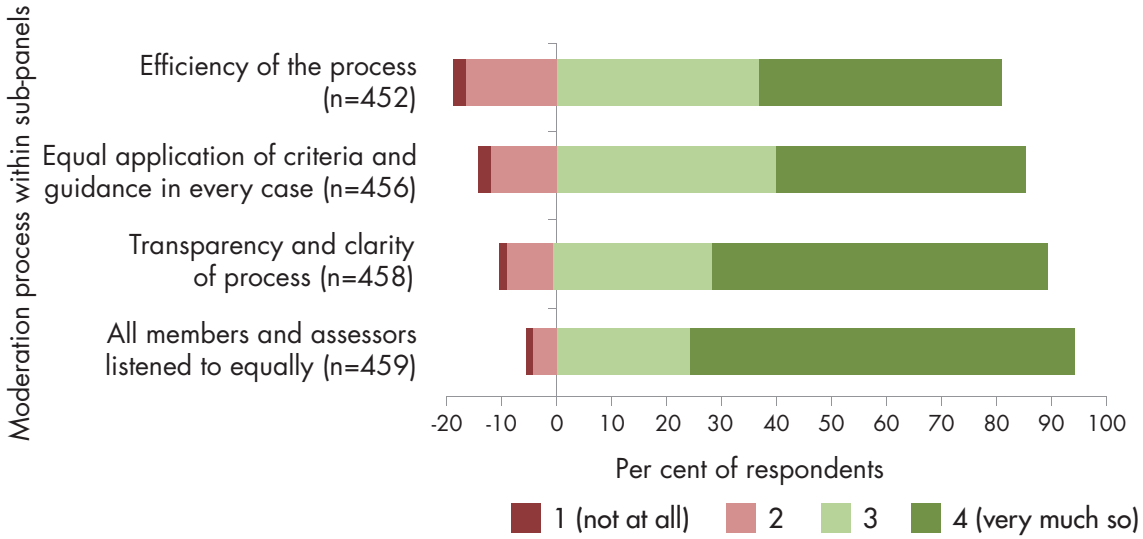
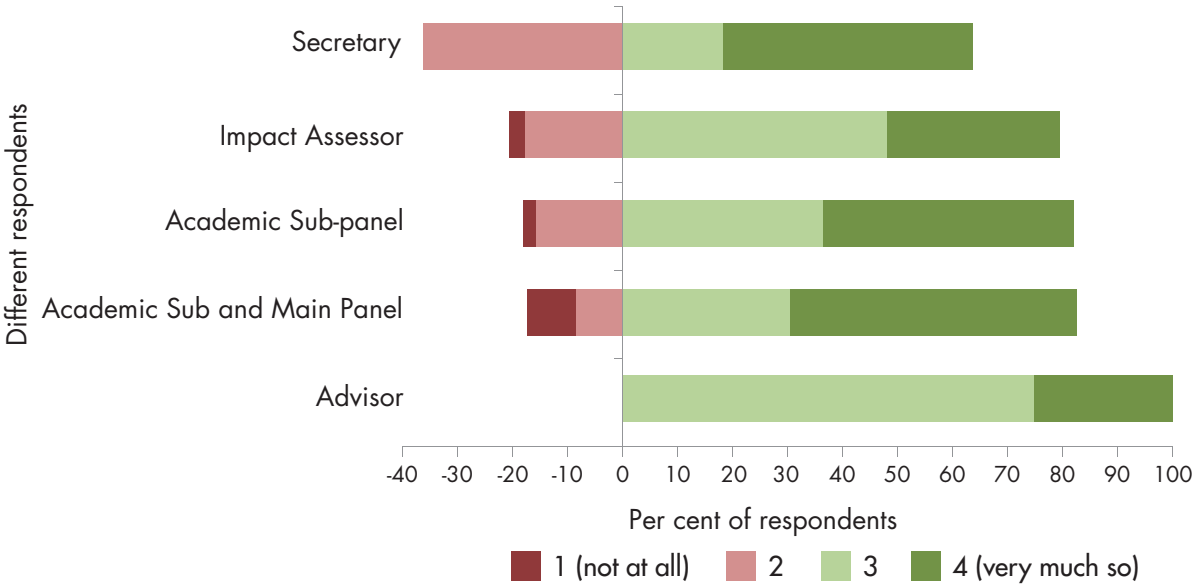


Figure 7.2: Views by type of respondent on satisfaction with the efficiency of the process of reaching agreement and management of disagreement[57]



further funding should be provided for additional cluster meetings as they felt more time was required for discussions at the cluster level and this would have meant they could arrive at a fairly consistent story prior to panel meeting and thereby reduce time spent debating difference in the full sub-panel meeting.

7.2. Resolution of disagreement through discussion

Through the focus groups, academics and research users across all panels, and the majority of sub-panels (22 out of 36), highlighted the importance of in-depth and

[56] Full question: How satisfied were you with the process of reaching agreement on the scores and the management of disagreements within the sub-panels in relation to the following areas?

[57] Full question: How satisfied were you with the process of reaching agreement on the scores and the management of disagreements within the sub-panels in relation to the following areas? Answers relating to the efficiency of the process broken down by respondent group.

iterative discussion in coming to consistent and reliable scores. This was facilitated by communication within and across clusters as well as at the sub-panel and main panel level. When asked about the robustness of the process the element of discussion was emphasised with one panellist stating: 'The robustness came from the discussions in the room, and time taken over'. Different approaches in different sub-panels were taken to resolve disagreement within clusters. Across all main panels, plenary discussion of the cases of disagreement was the mode employed by most sub-panels. Here, the cluster would discuss their judgement on a case study, and others could comment and have input. Some sub-panels linked justifications directly to the criteria and asked the panellists to 'highlight aspects of case which helped them make the decision'. The extent to which panellists were using the criteria or not was something that was continually challenged and discussed at the wider sub-panel level. This dialogue was critical to ensure 'resolution was achieved by consensus' and not by 'people fighting it out in the corner'. In Agriculture, veterinary and food science (UOA6), where there were only a relatively small number of case studies (128), all case studies were presented and discussed by the whole sub-panel.

Views on the level of discussion varied, with some people highlighting that more discussion would have been useful, and others thinking it was sufficient. For example, one individual felt that a lot of the discussion focused on the credibility of the statements made rather than on agreeing the scores.

7.2.1. Confidence in the agreement process

Panellists confirmed that discussions to achieve agreement were 'vital' and that they were surprised that when they checked their independent scores against those of their fellow panellists in their cluster they were consistent with one another. Panellists in three sub-panels commented anecdotally that in less than 20 per cent of cases there was a discrepancy of more than 1.5 levels and one noted that over time they became more confident and by 'the end of the process they were never more than half a star out in terms of assessment'. This is in line with the findings from the evaluation of the REF pilot in 2010 where 'assessors found that agreeing the scores of the case studies through discussion was mostly a straightforward process and difficult judgements arose from having poor quality information or balancing factors such as having significant impact at an early stage' (Technopolis 2010).

Such moderation discussions were thought to produce 'a realistic and reasonable score, [but] also tempered everyone's views, so that we understood and interpreted the criteria properly, and that went a long way towards negotiating a common consensus amongst the group on where scores lay'. It was also thought that the moderation process highlighted problems with scoring multidisciplinary research and assessing the quality of underpinning research and this provided an opportunity for academic and user members to work well together to discuss the threshold criteria around the quality of underpinning research. One individual pointed out that this element of discussion would likely be essential in subsequent assessments. This would not be negated by the learning which had happened this time, as the unique nature of case studies would mean that next time they would also be different and discussion would be required to achieve consensus.

7.2.2. Perceived challenges with moderation

In Mathematical sciences (UOA10), one individual said that they had to do a lot of moderation after the main panel meeting to try and square off the differences in scores between the users and academic members. This individual suggested that panellists viewed the REF terminology rather differently and these differences were not adequately regulated during the calibration phase. Panellists also noted that they felt they moderated their own scoring behaviour during the process. One user said that they felt themselves being stricter in the assessment earlier on because 'I had higher expectations of what was going to come in terms of sophisticated evidence and arguments around theories of change and complexity and didn't really get any of that at all and so then I calmed down' whilst another panellist noted that they 'moved from a formulaic to a conversational way [of scoring] and by the end it becomes an art rather than a science'.

7.3. Scores allocated from the process

In this section data on scoring performance including differences between the overall, quality, impact and environment scores and the relationship between them, is presented. This highlights some important differences between the main panels in terms of outcomes from the process.

In regards to the impact submission there was a significant difference between Main Panel A and other panels particularly in terms of 4-star rated submissions (Figure

Figure 7.3: Percentage of impact submissions with each level by main panel

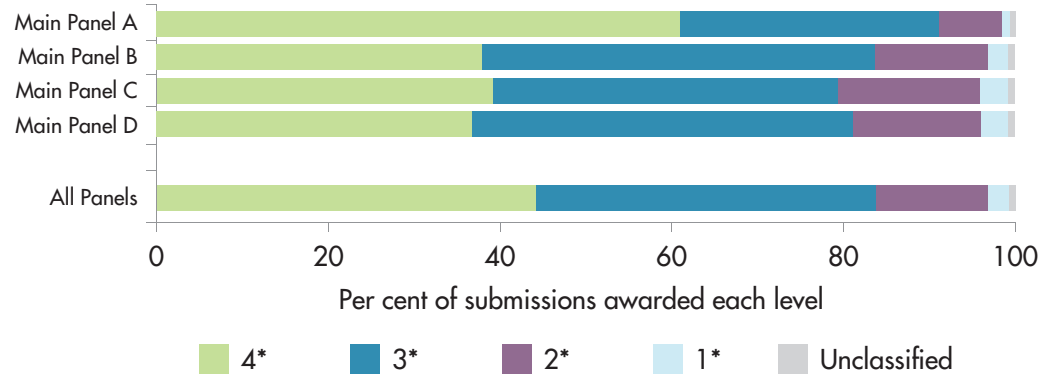
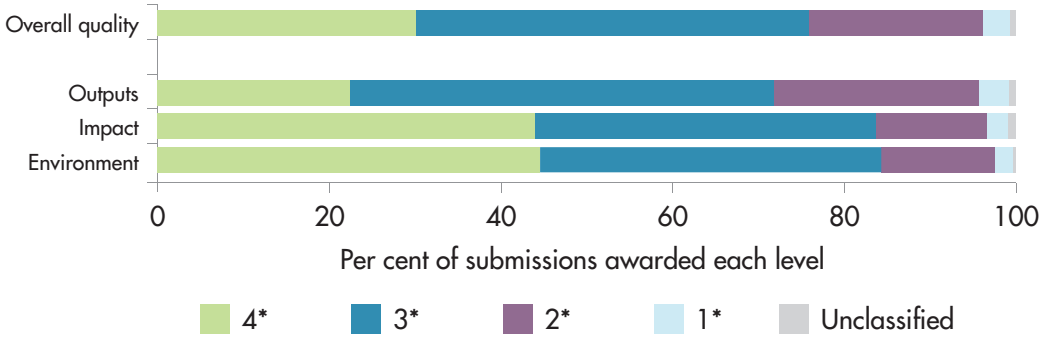


Figure 7.4: Percentage of submissions with each level in overall sub-profiles



7.3). While Main Panels B, C and D had between 36.6 and 39 per cent of submissions with 4-star ratings, Main Panel A achieved 60.9 per cent of the submissions with 4-star rating. There was also a significant difference in the environment submissions in Main Panel A compared with other panels particularly in terms of 4-star rated submissions (data not shown).

Figure 7.4 above gives an overview of scoring results and highlights that the performance in impact and environment elements of the submission was higher than that in outputs, particularly when regarding the number of 4-star submissions. Over 80 per cent of both impact and environment sub-profiles were considered to be either 3 or 4-star.

7.4. The effect of small submissions

Academic panel members from areas of Main Panel D highlighted a particular concern with the reliability of data in small submissions (submitting less than 15 FTEs and therefore requiring two case studies). In these cases it was argued that random variation in the scoring process had more impact on the final profile of a small submission. This problem is unique for the impact element, as the output element requires four outputs per FTE and therefore a submission of 10 FTEs will have 40 publications, but only two impact case studies. This was argued to be a particular problem in Main Panel D, where the average size of submission is smaller (Table 7.1).

Table 7.1: Size of submission by main panel

Panel	A	B	C	D
Per cent of submissions with only 2 case studies	20.1 per cent	17.7 per cent	23.1 per cent	30.5 per cent
Mean number of case studies	13.6	15.7	16.5	12.0
Median number of case studies	7	8.5	13	9
Mode number of case studies	4	4	6	2

Chapter 8 Evidence checking and audit

> *It would be useful to have routine access to corroborating evidence and underpinning research outputs in order to validate claims.*

An audit was conducted on a subset of case studies from each panel as a check on the material submitted. It was intended to be used to verify specific statements where panellists were uncertain and where panellists identified concerns about some of the information within the case studies. For example, further information could be sought from an HEI, corroborating contact or the public domain to inform the assessment, such as details regarding the threshold criteria, access to underpinning outputs and corroboration of impacts. The aim was to sample a minimum of 5 per cent of HEIs' case studies, starting with those identified as high priority by the panellist and complemented by a sample selected by the REF team, where fewer than 5 per cent of case studies were identified.[58] Overall audit queries were raised on over 8 per cent of all case studies (Table 8.1),[59] often exceeding 10 per cent within a given sub-panel (Figure 8.1). Some 74 per cent of institutions had at least 5 per cent of their case studies audited.

It was noted across areas covering the four main panels that the audit process was useful to support the panellists in making their judgements. Interviewees noted they generally raised an audit query when in doubt of the impact claimed, the eligibility of a case study or in instances where none of the impact was publically available. In particular, it was noted that it was hard to assess the significance of an impact where the evidence was 'nuanced' and in the form of corroborating testimonials. Often panellists requested evidence to 'check its validity'. To improve the process panellists would have liked to have easy access to the underpinning research, corroborating evidence and testimonials to check it themselves, instead they had to raise an audit query that took time to respond to. Panellists told us that the time taken to receive an answer to an audit query was approximately three weeks. This caused an additional burden on the panellists, as once the information had been returned they had to re-familiarise themselves with the case study to make their assessment. This extended the time taken to assess an impact case study in these instances.

One individual highlighted the importance of users' contributions in understanding international contexts where it may be difficult to collect evidence.

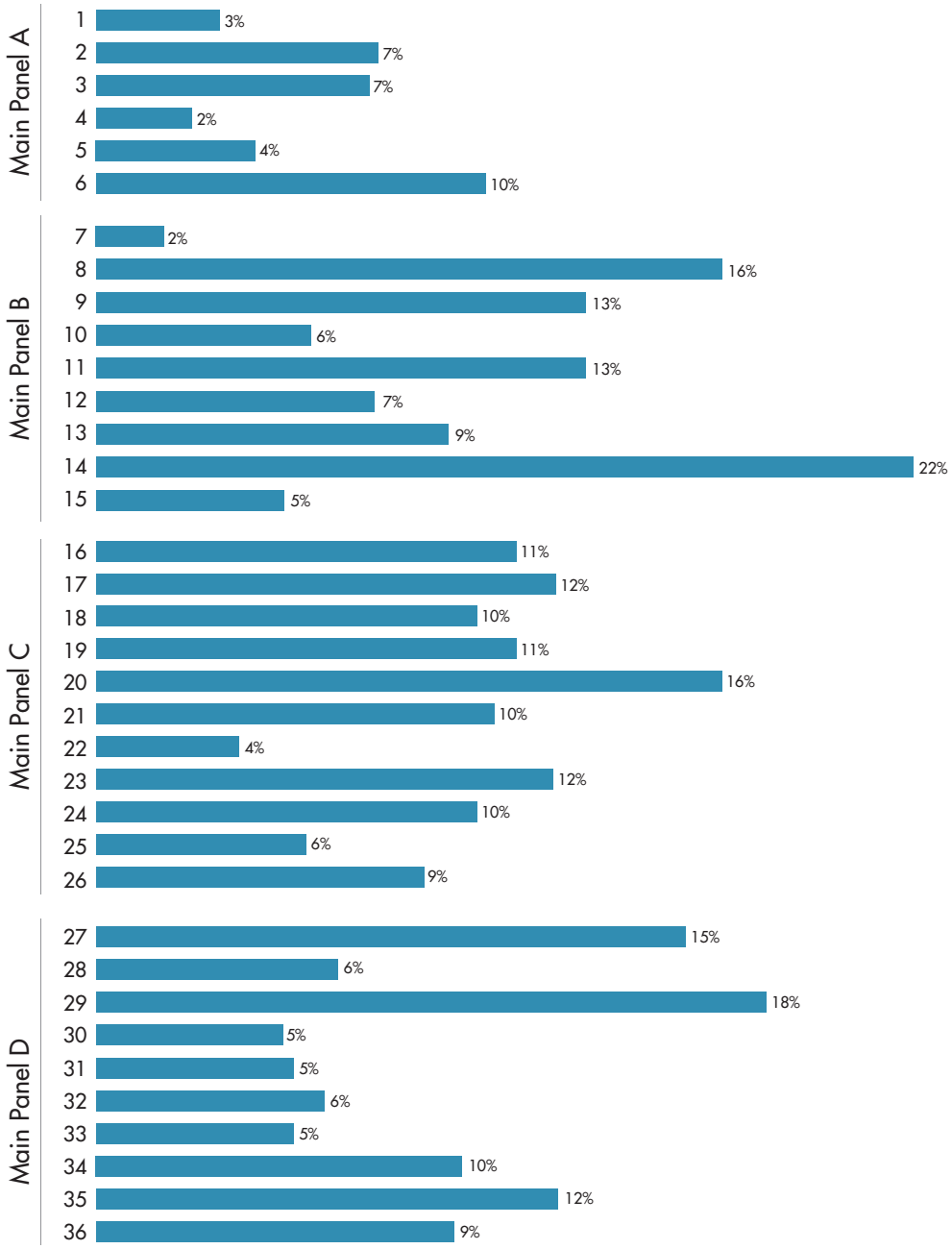
Table 8.1: Breakdown of audit queries across panels

	Main Panel				
	A	B	C	D	Total
Total number of queries	81	153	210	161	605
Per cent of case studies audited	4.9	9.1	10.3	9.8	8.7

[58] The aim was to audit at least 5 per cent of case studies where the HEI's submission met a certain threshold number.

[59] It is important to note however, that some audit queries could include multiple elements.

Figure 8.1: Distribution of audit queries across the 36 sub-panels



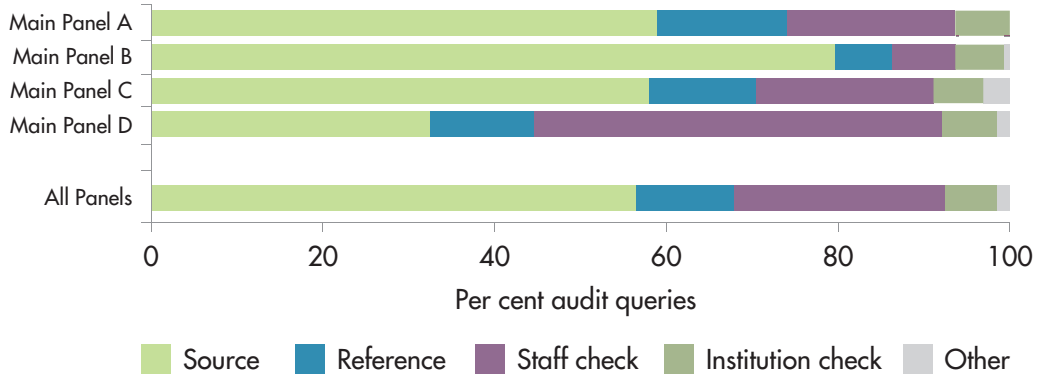
Focus group participants spoke of the challenge of not being able to validate evidence claims due to, for example, not having a reference's contact details. Some panellists said that they did make plenty of requests for corroborating evidence which entailed a significant amount of work for the REF office. There was a mixed response regarding how supportive evidence was of claims made when it was more closely inspected. Whilst some individuals said that they saw a number of cases where the supporting evidence didn't support the claims made, others thought that seeing the evidence didn't change their initial view and they thought the time spent on auditing seemed to be a waste. It was suggested that links to corroborating evidence, within the REF system, would be easier to use.

It was also highlighted by areas of Main Panel C that multiple requests for corroborating evidence were made and that its compilation would have been a huge amount of work. One individual said that the REF team were only auditing 5–10 per cent of case studies and that they did not have the resources to do more

Table 8.2: Type of audit queries

Type of audit query	Description	Number of queries
Source	The query referred to the accessing of corroborating source in Section 5 of the case study including queries initiated by the secretariat	343
Reference	The query referred to the accessing of research references in Section 3 of the case study	67
Staff check	The query related to the need for the institution to provide dates of employment of any of the staff members indicated in the case study	146
Institution check	When the query referred to the need for the institution to provide justification that the research was actually conducted in the claiming institution	36
Other	E.g. pdf of case study not available, missing case study sections	9

Figure 8.2: Distribution of reasons for audit queries by main panel



than that. Another panellist thought that asking for this evidence was not a good use of time as, in their cases, it did not contribute to any changes in scores.

Responses from the survey suggest that obtaining the data for the auditing of case studies was reasonably straightforward for advisors and secretaries to carry out. They indicated that it typically took around one hour per case study for the audit tasks they were asked to conduct.

8.1. Types of audit queries raised

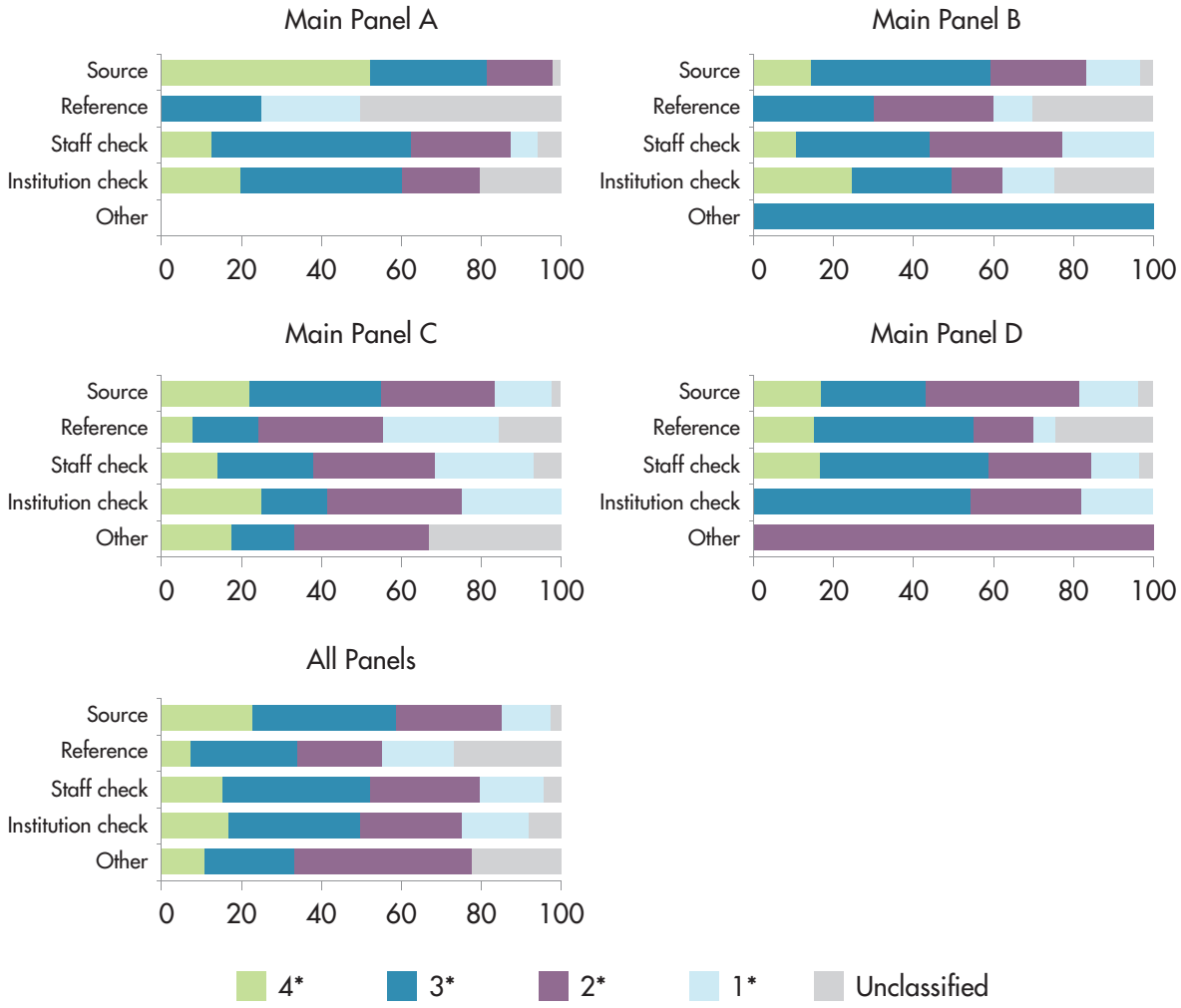
There are a number of reasons a case study could be audited (see Table 8.2). The most common reason was to check the underpinning evidence ('Source'), followed by checks on the location of staff when the research was conducted (see Figure 8.2). The Main Panel C overview report also noted that submissions that did not provide basic information (for example, regarding the dates of employment for researchers who conducted underpinning research) led to a degree of auditing that should not have been required, and could be avoided in future assessments, with clarifications in the template (REF 2015).

When assessing the relationship between the types of audit query and the level awarded, although checking underpinning evidence ('Source') caused the greatest number of audit queries, only a very small number of these were subsequently unclassified. By contrast, over a quarter of those case studies audited on the basis of the underpinning research (described as 'Reference' and accounting for about 11 per cent of all audit queries) were subsequently unclassified (Figure 8.3).

8.2. Areas where secretariat provided advice

In addition to the process of audit, the secretariat also provide advice to panellists. According to survey data, each advisor was asked to provide advice on average regarding 85 case studies and 30 templates, though the exact numbers differed between respondents, and seem to be lower for areas of Main Panel B in particular, where the two responses indicate that each advisor was only asked to provide advice on around 15–20 case studies and 0–2 templates.

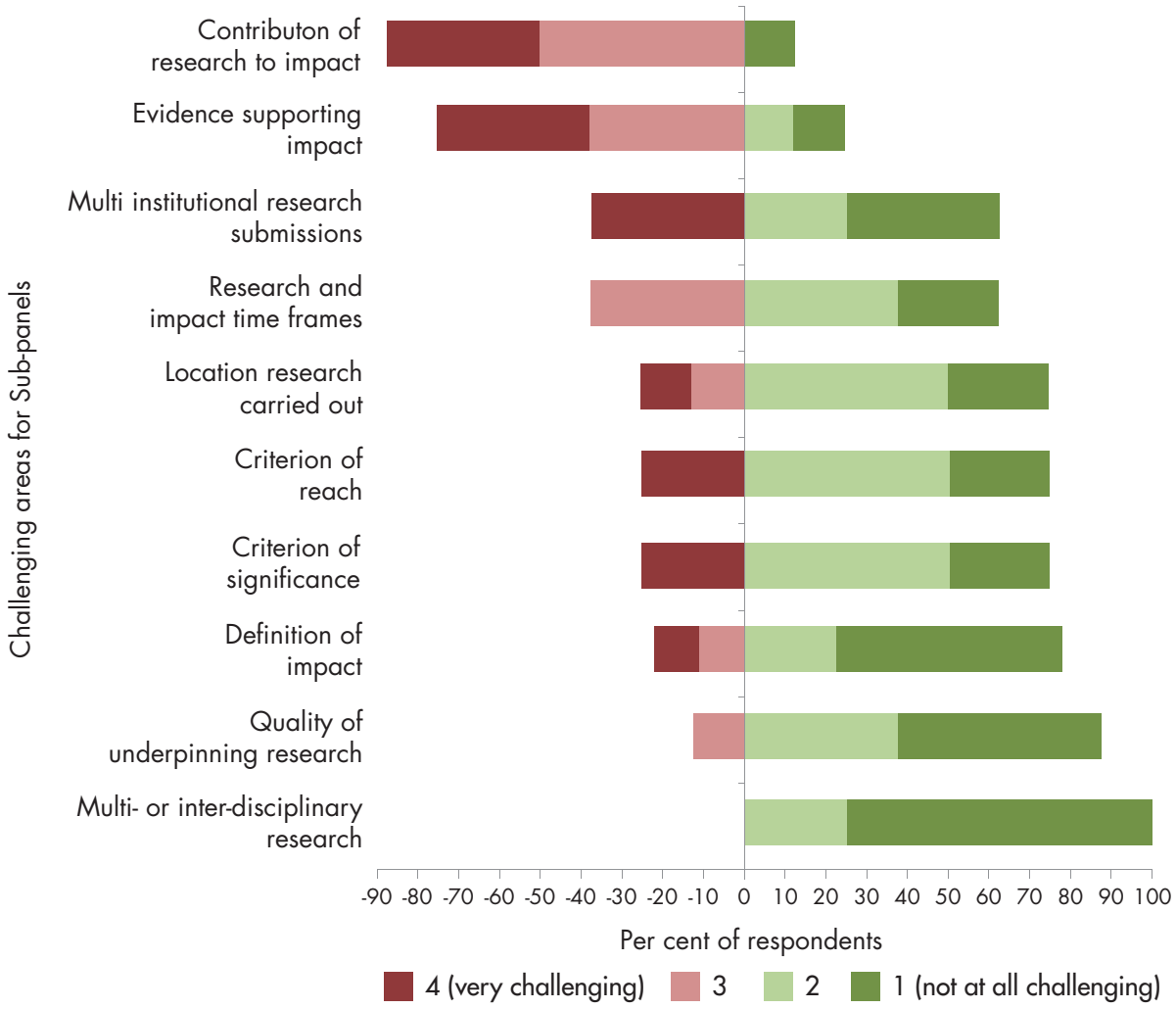
Figure 8.3: Distribution of audit queries by level awarded[60]



When asked about the nature of the queries, advisors reported that the contribution of research to impact, and the evidence supporting impact were the areas of greatest concern to panel members (see Figure 8.4). This is reflected in the data on audit queries presented in Figure 8.2, which shows that evidence supporting impact is the most common reason for audit queries.

[60] No data was recorded for the Main Panel D secretariat.

Figure 8.4: Areas secretariat and advisers thought sub-panels found most challenging[61]



[61] Full question: Based on the queries you received, which of these areas do you think sub-panels found most challenging?

Chapter 9 Overview of the perceived benefits and challenges associated with the assessment process

> *"I enjoyed learning about the wide range of impact taking place in my discipline.*

> *"It was difficult to extract the evidence and understand the arguments in some of the templates submitted.*

In our survey we asked panellists to describe the benefits and challenges they observed from participating in the assessment process. Their free text responses were categorised, and the overall number of responses and respondents naming each category was calculated. The data was broken down by type of panellist and main panel to determine any variation in perspectives. It is important to note that based on the various roles undertaken by panellists there were differing numbers of respondents to each survey. As a result, this data comments in the main on the sub-panel members and impact assessors.

9.1. Benefits

The top five benefits identified for each group of respondents is shown in Table 9.1 and discussed for each respondent group below.

9.1.1. Engaging with the process allowed impact assessors to build networks and relationships, particularly with academia

External to the sector, the reasons for engaging with the process and the benefits enjoyed by impact assessors were necessarily less direct.

The most cited benefit among impact assessors was networking and collaborative working. This included the enjoyment of engaging with academics (and the academic sector more broadly) as well as networking opportunities with the academic sector for further professional activities.

This is closely linked to the second most cited benefit which was learning about and gaining an understanding of academic research taking place across the UK. Respondents noted they had gained an awareness of different types of research, had been exposed to new research ideas and had a better overview of academic disciplines of relevance to their work, among other things.

Table 9.1: Top five responses received when asked to provide up to three benefits of engagement with the impact element of the REF process[62]

Impact assessors	Sub-panel members	Main panel members
Networking/collaboration (44/74)	Process of assessment (174/396)	Process of assessment (20/51)
Academic research (37/74)	Academic research (99/396)	Academic research (13/51)
Process of assessment (27/74)	Breadth of impact (77/396)	Breadth of impact (7/51)
Breadth of impact (10/74)	Impact of research in subject/ discipline (72/396)	Networking (6/51)
HEI impact/engagement strategies (9/74)	Understanding of impact (61/396)	Impact of research in subject/discipline (6/51)

[62] Number of respondents provided in brackets. Main panel members includes user members and those who sat on both the main panels and a sub-panel (i.e. sub-panel chairs).

Also cited by a significant number of respondents was enjoyment and increased understanding of the process of assessment itself. Several noted it was beneficial to understand how the process of impact assessment works. In some cases this was thought to be particularly useful in working with HEIs and understanding their requirements in future.

An understanding of and exposure to the wide range of different types of impact was also perceived as a benefit by ten respondents. In some cases this related to gaining an understanding of various types of impact generally, and in others respondents valued learning about how research had contributed to impacts in their discipline or area of work.

Insight into various HEIs' approaches to impact and engagement was also thought to be valuable by nine respondents. The reasons for this varied and included gaining an understanding of how academia plans to engage with the policy sector, being able to help HEIs engage more effectively with schools and identifying what helps HEIs gain better impact.

Overall, this gives the broad picture that the process for impact assessors was about building and maintaining links and networks which are valuable for their wider activities, supported by an increased understanding of the academic sector.

Other benefits identified by respondents included the exercise highlighting the importance of impact, benchmarking HEIs (i.e. highlighting various strengths and weaknesses across HEIs), the personal prestige of being involved in the process, the opportunity for 'creative', 'useful' and 'informative' discussions with colleagues, gaining a knowledge of how to generate impact from research in future, contributing to society and assessment through the process, gaining new experience, gaining an understanding of the difficulties of impact assessment, increasing knowledge generally, gaining inspiration for future research and impact activities, the sense of doing something new and the opportunity to contribute to improvements in the standard of research impact.

Differences between main panels in terms of these responses were minor. Understanding the breadth of impact observed across different case studies was considered more important by respondents in areas of Main Panels A and B, while understanding the process of assessment and insight into various HEIs approach to impact and engagement were considered more important to areas of Main Panel C and D respondents.

9.1.2. The process and its outcomes have direct implications for academics and their institutions

Academic sub-panel members reported direct benefits for their careers and their institutions resulting from participation in the assessment process. By far the most frequently cited benefit related to their improved understanding of the assessment process. Specific examples stated by individuals included learning how the process of assessing impact worked in practice; understanding how other people judge impact; developing a deeper understanding of the rules and how they should be applied; gaining reassurance of the robustness of the process; and learning how to approach the exercise for future assessments – including how to present good case studies and templates. This reflects the increasing importance of impact assessment in academia not just through the REF but more widely. Academics recognise the importance of engaging with and understanding the concepts around impact and involvement in this process was seen by many participants as an opportunity to develop this understanding in the context of the REF and more widely.

Other frequently mentioned benefits included learning about the range of academic research taking place across UK HEIs and learning about a diverse range of impacts emanating from research. Both of these again reflect benefits of engaging for both academics' own careers, and for their institutions more widely in terms of building an understanding of impact and its assessment, and an inside perspective on what other academics and institutions are doing in this area – effectively giving those participants a 'competitive advantage'.

The next most cited benefit also related to the range of impacts observed, but these comments were specifically in relation to a subject or discipline area. Examples of comments in this category include: 'awareness of the success of UK research effort in the biomedical sciences', 'witnessing the quality of some of the impact in my discipline', 'learning about REF-type impact of mathematics' and 'evidence that civil engineering research contributed to social, economic and environmental impacts, as anticipated'. The final benefit mentioned in the top five was an increased understanding of impact, again reflecting the growing importance of the impact agenda for the academic community.

Differences between main panels were small, with improved understanding of the process of assessment the most important across them all. For areas of Main Panels A and C, knowledge of how to generate impact was a

top five benefit, in place of understanding of impact specific to their discipline (in the case of areas of Main Panel A) or understanding of the breadth of impact (for Main Panel C). For areas of Main Panel D, insight into other HEIs and departments was a panel specific benefit cited, replacing an increased understanding of impact. These minor changes do not change the overall broader picture that for sub-panel members, engaging with the REF was an opportunity to learn more about the process and about impact at other institutions, which is likely to be beneficial for themselves and their institutions in the long run.

This picture was largely repeated for main panel members who are also predominantly academics. The top three benefits listed were as for the academic sub-panel members. Specific responses included gaining a better understanding of how to assess impact, understanding impact within the context of the REF impact assessment, gaining experience for future REF impact exercises, gaining a good overview and understanding of a range of different impacts, and an insight into involving different stakeholders in the assessment process. The next most frequently mentioned benefit was the one listed as most important for the impact assessors, networking and collaborative working, perhaps reflecting the views of the user members that fall within this group. The final benefit listed among the top five was insight into the range of impacts in a particular discipline or research area, mentioned by six respondents. Examples of responses in this category include 'learning about the diversity and quality of impact in my discipline and cognate areas' and gaining 'a more thorough understanding of how the creative and performing arts contribute to the cultural ecology of the UK'. Differences between main panels again were not significant and, as described, the overall picture reflects that seen for the academic sub-panel members.

9.2. Challenges

The top five challenges identified for each group of respondents is shown in Table 9.2 and discussed for each respondent group below.

9.2.1. The burden of involvement was a particular challenge for impact assessors

The burden involved in acting as an impact assessor was a particular challenge. Burden was by far the most frequently mentioned challenge of involvement in the assessment, being mentioned by more than half of respondents. Based on the quantitative data collected through the survey, the median amount of time spent on the process by impact assessors was 11 days. There was, however, some variation in the commitment made by individual impact assessors – the interquartile range of the estimates of time spent by impact assessors from the survey spanned from 8 to 16 days. A more detailed analysis of the burden across all groups is provided in Chapter 10. This burden was particularly significant for impact assessors, since though time out of work may have been given to cover the time attending meetings, it is likely they would have had to largely find the time for the assessment of impact documents in addition to their full time job (in contrast to academic participants who may have been able to assess case studies and impact templates during working hours). Examples illustrating this from the specific responses received include having to find evenings and weekends to read through assessment case studies, all REF preparation work being done outside of working hours, fitting in

Table 9.2: Top five responses received when asked to provide up to three challenges of engagement with the impact element of the REF process[63]

Impact assessors	Sub-panel members	Main panel members
Time/burden (45/74)	Time/burden (86/396)	Time/burden (10/51)
Inexperience/learning new rules (14/74)	Using evidence (79/396)	Calibration (8/51)
Calibration (14/74)	Comparing/fairly assessing different types of impact (76/396)	Comparing different types of impact (8/51)
Managing documentation (10/74)	Calibration (68/396)	Ensuring fair and consistent scoring (7/51)
Integration with academics (9/74)	New process (45/396)	Working with others/Applying rules and guidance/Attribution to research (5/51)

[63] Number of respondents provided in brackets. Main panel members includes user members and those who sat on both the main panels and a sub-panel (i.e. sub-panel chairs).

with heavy workload from their main job, time/burden required on top of their day job, and completing it at weekends and evenings.

Another issue mentioned reasonably frequently was the challenge related to inexperience of the process and the need to learn new rules. This perhaps reflects the fact that this group is external to the sector and as such is likely to be less familiar with the impact agenda, and also perhaps that impact assessors joined the process later and as such would have taken some time to familiarise themselves with the rules and the process compared to sub-panel members who had been involved from an earlier stage. This is reflected in the specific responses received which include: inexperience/learning new rules, getting up to speed; inexperience in assessing impact; absorbing the lengthy guidance and understanding the process; initial concern over unfamiliarity with academic process; trying to understand the process of the scoring; figuring out how the REF exercise works.

Calibration was also a challenge according to impact assessors. Specific responses included resolving differing views on the panel, coordinating discussions on scoring discrepancies, the differing avenues through which research impact can be assessed and how different individuals might put differing emphasis on the impact, too little time taken to calibrate our sub-panel against other sub-panels, reaching agreement and calibrating results, and resolving differing views on the panel.

Managing documentation was the next most frequently cited challenge, which largely reflected challenges that the group had experienced using the REF IT systems. Particular comments referred to issues with the separate email system, the volume of electronic material, challenges with encrypted memory sticks, and more broadly in terms of getting the technology to work effectively.

The final challenge listed in the top five for this group was integration with academics. Here impact assessors mentioned concerns about the voice of users being heard and respected by academics, differences in background and experience, indifference/suspicion among academics, and the feeling that the academics already knew each other.

Differences between main panels are small, with issues around managing documentation less significant for areas of Main Panel B members, and issues around integration with academics not being mentioned in the top five challenges amongst areas of Main Panel D respondents. Overall, these responses largely reflect the practical concerns of the group about their personal involvement in the process, with burden being by far the biggest concern.

9.2.2. Burden was also important for academics, but they also reflect on the challenges of the process itself

In terms of the challenges in engaging in the process, the burden of engagement is listed the most frequently among sub-panel members (though by less than a quarter of respondents in this case), but several other response categories are mentioned almost as frequently, notably the challenges of using evidence and of comparing and fairly assessing different types of impact.

The challenges around the use of evidence here related particularly to issues around corroboration and the level of information available in the case studies. Looking at the specific responses, there was a sense that the statements had to be taken at 'face value' and panel members typically did not have access to evidence around the claims made which was a challenge in assessing the case studies. Sub-panel members were also concerned about the extent to which they were able to fairly assess different types of impact. One respondent referred to this as 'comparing apples and oranges', reflecting the challenge of judging very different types of impact in an equivalent/fair manner.

Calibration was also mentioned as a challenge by several respondents. This refers in particular to reaching a resolution regarding differences of opinion amongst the different case study reviewers. Looking at the specific responses from individuals, resolving differences of opinion and perspective between research users and academics seems to have been a particular concern here.

Finally, the novelty of the process was also one of the five most frequently cited challenges amongst sub-panel members. This refers to the challenge of conducting a process for the first time with no previous exercises to use as 'reference points'. The lack of any 'precedents' was mentioned by several respondents, and this meant that when particular challenges arose, the panellists were required to find solutions themselves, rather than being able to draw on previous experience or fixed practice with regard to this process.

Responses were fairly consistent across the main panels, with some slight variation in the order in which these challenges would be placed. For areas of Main Panels B and C, issues around attribution were also important and would have made the five most frequently cited

challenges. For areas of Main Panels C and D, challenges around applying the rules and guidance were also important.

Overall, unlike impact assessors, the issue of burden, though still important, is not so dominant for this group, which also reflects on the challenges of the process itself rather than challenges for themselves in engaging with the process. This re-emphasises the more direct engagement that this group necessarily has with the process and its outcomes which directly affect their sector. While understanding the process and gaining an 'inside' perspective is beneficial for this group, the challenges and limitations of the process have direct impacts and as such are of significant personal concern to them.

Main panel members describe a wide range of challenges, largely focusing on the process rather than their own experience. Again, the most cited challenge was the time and burden involved, but as with the academic sub-panel it was not as dominant as for the impact assessors. The second most frequently mentioned challenge was that around calibration. In contrast to sub-panel members, here the specific responses were not so focused on disagreements between users and academics and reflected the main panel role, with issues such as calibration between disciplines and sub-panels being mentioned more frequently.

Challenges around comparing different types of impact were also frequently mentioned, as was the challenge of ensuring fair and consistent scoring. For the second of these, the importance reflects the main panel role, with comments again referring to ensuring a wider understanding of the rules and ensuring consistency across sub-panels. Finally, three further challenges were cited equally frequently: working with others, applying the rules and guidance, and attribution to research. Looking at the specific comments, the first of these reflects the fact that many of these respondents were sub-panel chairs and had to take responsibility for the cohesion and discussion among panellists with differing opinions. The issues around the rules and guidance and attribution again reflect concerns about the process and the way the rules are applied in practice.

Overall, main panel members do have a focus on the challenges of the process rather than personal challenges, as with academic sub-panel members, but their focus is a little different reflecting their different role, taking in issues such as the oversight of sub-panels, and ensuring wider consistency across them, which are not as important to sub-panel members.

Chapter 10 Burden associated with the assessment of the impact element of REF 2014

> *I thought that I had a lot to do reviewing case studies, until I talked to panellists who were also assessing outputs. I was astonished by the quantity they had to read and at the dedication of those people, who'd reviewed so much and in such a remarkable manner.*

As described in Chapter 9, across all groups involved in the REF 2014 impact assessment, burden was raised as the most important challenge of their involvement. We have conducted an analysis of the overall burden of the process based on data collected from respondents to the survey. All participants were asked to estimate the number of hours they spent on impact-related activities as part of the REF assessment process (i.e. excluding those activities which formed part of the submission process at their own institution, where relevant). This was broken down into relevant subcategories for the different types of panel members, but it seems from the data received that not all respondents used the sub-categories consistently, so the analysis is on the basis of the aggregated data for each individual on the total time they spent on the assessment of impact.

To ensure the quality of these data we reviewed each response individually, reading the text that was provided alongside the numerical responses. Some respondents had not completed this question or had simply put '0' in each box. These responses were excluded from the analysis. Other respondents had commented in the accompanying free text box that they were not able to respond to this question and that their responses should not be included in the analysis, or indicated that they did not understand the question. Their responses were also excluded.

The number of hours spent on the impact assessment process, broken down by type of panellist, is set out in Table 10.1 below. The first thing to note is that there was significant variation in the amount of time spent on the process within each group. We attempted to run a simple regression to see whether this could be explained by the noted differences in number of impact documents allocated to each individual. However, the number of case studies (or impact documents as a whole) reviewed did not account for a significant proportion of the variation in the time spent on impact assessment. Differences between research users and academics were also not helpful in explaining this variation. Based on these investigations, we have conducted a burden analysis using the median values for each group of panellists. However, the significant differences within each group should be borne in mind when interpreting these data.

Note also that there is some overlap between two of those groups. Chairs (and some deputy chairs) of sub-panels made up the majority of the main panel membership. As such, these individuals were classified as main panel members for the purposes of this analysis. This is because it is likely that they had additional duties through their main panel membership that increased their burden comparative to most sub-panel members. There were a few main panel members

Table 10.1: Time spent on the REF process, in hours, by sub-group

Type of panellist		Impact assessor	Sub-panel member	Main panel member	Advisor	Secretary	International advisor	Observer
Time spent on impact assessment for the REF (in hours)	Median	85	70	167	241	103	84	34
	Inter-quartile range	58-123	50-105	70-224	181-507	88-129	72-229	31-52

(notably the chair of each main panel) who did not have duties on a sub-panel. However, these people have not been categorised separately. This is partly because their numbers were so few among survey respondents, but also because it is likely that main panel chairs had additional duties in their role such that although they were no longer required to sit on sub-panels, it is unlikely their workload was decreased relative to other main panel members. Given the wide variation within groupings seen, it was not considered that this was likely to make a significant impact on the overall analysis.

In order to calculate a monetary burden, the number of panellists in each grouping was determined through an analysis of the full lists of panellists. We converted the number of hours into a number of days assuming a 7.5-hour working day (as used in the analysis of burden in the submission process, Manville et al. 2015a). Salary bands were then allocated to each grouping. This was done using the academic pay scale data from the Higher Education Statistics Agency (HESA), which was also used in the analysis of the burden associated with the submission process (Manville et al. 2015a). For sub-panel members, the professorial band was used. For the main panel members, the 'Senior Management Team' band was used. For impact assessors and observers, it was very difficult to determine what would be appropriate to use as the individuals within this grouping were very diverse, and from a wide range of industries with vastly differing levels of remuneration. For simplicity, it was decided that the professorial band should also be used for these groups; this may be a considerable under- or over-estimate for many individuals. For advisors and secretaries, the 'Senior Administrator' band was used. The pay bands were converted into a daily rate, based on an estimate of the number of working days in a calendar year (allowing for 25 days leave, plus UK bank holidays). No additional allowance for other employment costs (such as employer NI contributions, or overhead costs) was made.

Summing this over all the groupings gives an estimate for the burden of the impact element of the process of £2.1m. This does not include travel costs and expenses for travel to meetings, costs of meeting venues, equipment costs, or any overhead costs within HEFCE for running the process. As described above, the salary bands used are very crude estimates. The result is also sensitive to the variation seen in the survey data on which the analysis is based. Taking the lower and higher values of the inter-quartile range for the time spent for each grouping produces a range of values from £1.5m to £3.3m for this estimate. As such, this estimate should only be treated as an order of magnitude estimation of the likely scale of the burden of the process.

In our analysis of the submission process (Manville et al. 2015b), we estimated that the total costs of the submission of the impact element were £55m, with a lower estimate of £51m and an upper estimate of £63m. Combining this with our estimate of the burden of the impact element of the assessment process, we get an estimate for the total burden of the impact element of the assessment process of £53m to £66m (based on the interquartile ranges as described above and the assumptions set out in Manville et al. 2015a), with a best estimate of £57m. These estimates are to the nearest million (to avoid implying unrealistic levels of accuracy) and are subject to all the caveats set out above and in the evaluation of the submission process.

We can use these data to estimate 'transaction costs', defined as the costs for preparing and assessing impact submissions to REF 2014 divided by the total expected QR funding that HEIs may receive over six years (i.e. between 2015 and an assumed REF in 2020).[64] We estimate QR funding to be £8.2bn for this period (we feel this is conservative as we focused on QR funding for 2013/14, which was £1.4bn and did not inflate this figure for the subsequent six years). The transaction costs for the impact element (preparation and assessment) – that is comparing the costs (£57m) with 20 per cent of QR funding determined by impact (£1.6bn) – were 3.5 per cent. The assessment process increases the calculated transaction cost above the submission costs by 0.1 per cent. As stated in Manville et al., 2015b, by means of comparison, the proportion of funds awarded by UK Research Councils to HEIs spent on administration costs was historically estimated to be around 10 per cent, but this figure may have declined in recent years (DTZ Consulting & Research 2006; Morgan Jones & Grant 2013).

[64] We also estimate the transaction costs associated with the impact element of REF (i.e. the costs of preparing and assessing impact submissions divided by the financial benefit as the 20 per cent weight of QR funding). It is important to note that some other aspects of recurrent research funding are informed by the outcome of research assessment.

List of references

Bazeley, P., & K. Jackson. 2013. *Qualitative Data Analysis with NVivo*. London: Sage Publications.

DTZ Consulting & Research. 2006. *Research Councils UK. Analysis of the External Costs of Peer Review: A Final Report.* Research Councils UK. Ref: 05125137. As of 23 March 2015: http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/prdtz.pdf

King's College London and Digital Science. 2015. *The nature, scale and beneficiaries of research impact: An initial analysis of REF 2014 impact case studies*. Bristol: HEFCE

Manville, C., M. Morgan Jones, M.-L. Henham, S. Castle-Clarke, M. Frearson, S. Gunashekar & J. Grant. 2015a. *Evaluation of Submission Preparation for Impact Assessment REF 2014: Approach and evidence* Santa Monica, Calif.: RAND Corporation. RR-7276-HEFCE.

Manville, C., M. Morgan Jones, M. Frearson, S. Castle-Clarke, M.-L. Henham, S. Gunashekar & J. Grant. 2015b. *Evaluation of Submission Preparation for Impact Assessment REF 2014: Findings and Observations.* Santa Monica, Calif.: RAND Corporation. RR-727-HEFCE.

Morgan Jones, M., & J. Grant. 2013. 'Making the Grade: Methodologies for Assessing and Evidencing Research Impact.' In *7 Essays on Impact*, edited by Dean et al., 25–43. DESCRIBE Project Report for Jisc. Exeter: University of Exeter.

Research Excellence Framework (REF). 2010. *Units of assessment and recruitment of expert panels*. REF 01.2010 As of 13 March 2015: http://www.ref.ac.uk/pubs/2010-01/

——. 2011. *Assessment Framework and Guidance on Submissions*. REF 02.2011. As of 13 March 2015: http://www.ref.ac.uk/pubs/2011-02/

——. 2012a. *Panel Criteria and Working Methods*. REF 01.2012. As of 13 March 2015: http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf

——. 2012b. *Invitation to complete the Research Excellence Framework (REF) survey of submission intentions.* As of 13 March 2015: http://www.ref.ac.uk/pubs/surveysubmit/

——. 2015. REF 2014 Panel overview reports. As of 13 March 2015: http://www.ref.ac.uk/panels/paneloverviewreports/

Technopolis. 2010. REF *Research Impact Pilot Exercise Lessons-Learned Project: Feedback on Pilot Submissions – Final Report*. Brighton: Technopolis. As of 13 March 2015: http://www.ref.ac.uk/media/ref/content/pub/refresearchimpactpilotexerciselessons-learnedprojectfeedbackonpilotsubmissions/re02_10.pdf

Appendix A List of panels and units of assessment for REF 2014

Main panel	Unit of assessment / sub-panel	Full title
A	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
	6	Agriculture, Veterinary and Food Science
B	7	Earth Systems and Environmental Sciences
	8	Chemistry
	9	Physics
	10	Mathematical Sciences
	11	Computer Science and Informatics
	12	Aeronautical, Mechanical, Chemical and Manufacturing Engineering
	13	Electrical and Electronic Engineering, Metallurgy and Materials
	14	Civil and Construction Engineering
	15	General Engineering
C	16	Architecture, Built Environment and Planning
	17	Geography, Environmental Studies and Archaeology
	18	Economics and Econometrics
	19	Business and Management Studies
	20	Law
	21	Politics and International Studies
	22	Social Work and Social Policy
	23	Sociology
	24	Anthropology and Development Studies
	25	Education
	26	Sport and Exercise Sciences, Leisure and Tourism

Main panel	Unit of assessment / sub-panel	Full title
D	27	Area Studies
	28	Modern Languages and Linguistics
	29	English Language and Literature
	30	History
	31	Classics
	32	Philosophy
	33	Theology and Religious Studies
	34	Art and Design: History, Practice and Theory
	35	Music, Drama, Dance and Performing Arts
	36	Communication, Cultural and Media Studies, Library and Information Management

Appendix B Semi-structured protocol for focus group interviews

Scoring, moderating and calibrating

1. How did the panel ensure consistency of treatment of the case studies, such that the score given by one reviewer could reasonably be expected to be the same if the study had been reviewed by a different member of the panel?
2. What was the most difficult thing in judging impact scores for impact case studies?
3. Did the panel agree in advance the parameters in which the various scores for impact could be allocated (e.g., 4 star as opposed to 3 star), and the types of things that would, and would not, be 'counted' as impact – for example, the place of 'public engagement' in the assessment? How formulaically was this applied?
4. What similarity of approach did, or did not, exist across the sub-panels – and indeed the main panels – with respect to the review and scoring of the impact template and case studies?

Working with REF definitions, rules and templates

1. How did you use the impact template (REF3a) and how it could have been improved, if at all?
2. How well did the format and guidance work?
3. What was the relationship between the impact template (REF3a) and case studies (REF3b)?
4. How was the assessment of the impact template (REF3a) reviewed in relation to the impact case studies (REF3b) and the environment template (REF5)?

Putting definitions into practice

1. How did you 'measure' (or 'assess') impact?
2. What sorts of metrics were deemed to be suitable for inclusion in case studies, and what sorts of metrics were considered 'invalid'?
3. How did the panel agree the final score for each case study?
4. Can you explain to me the relative strength of quantitative and qualitative evidence for impact
5. How did the panels interpret 'outstanding' reach and significance as opposed to 'very considerable' and 'considerable'?
6. How did you balance 'significance' and 'reach'?
7. How did the panels benchmark between case studies within and across sub-panels? Did they have any particular difficulties or concerns?
8. How did panels compare the more nebulous (and legitimate in terms of the REF criteria) impacts such as impacting public discourse, awareness raising, influence on public or policy debate with more 'concrete' impacts such as visitor numbers, sales figures, patents, etc.?
9. Did the panels feel that they had sufficient time to make robust judgements on each of the impact case studies and then to cross-refer judgements across panels to ensure parity?
10. Did you have a good understanding of how beneficiaries actually used the research?

Corroborating evidence

1. Did you feel that the panels had sufficient information/evidence to form a reliable judgement?
2. What additions to the evidence set would you like to see next time, if any?
3. Who normally led the assessment of impact when looking at individual case studies?
4. Was it straightforward to confirm that the underpinning research was of 2 star and above?
5. How did the panels decide which sources of evidence to follow up, on what basis, and what new information did it bring to light?

Conclusion

1. Do you think this was a fair, reliable and robust process for assessing case studies and templates?
2. Why do you think this?
3. Can you provide examples that we haven't discussed so far?
4. What could have been done differently?

Appendix C NVivo code book for coding interviewee responses from the focus groups[65]

- 01. Panel management
  - 1.1. Communications
    - 1.1.1. Clusters
    - 1.1.2. Sub-panel level
    - 1.1.3. Main Panel level
    - 1.1.4. Supra-panel level
  - 1.2. Calibration
  - 1.3. Moderation & Validation
  - 1.4. Auditing
    - 1.4.1. Underpinning research
    - 1.4.2. Corroborating evidence
  - 1.5. Burden
  - 1.6. Other
- 02. Dealing with difference
  - 2.1. Type of impact
  - 2.2. Type of evidence
  - 2.3. Inter and multidisciplinary research
  - 2.4. Multi-institutional research
  - 2.5. Confidential impact case studies
  - 2.6. Redacted impact case studies
  - 2.7. Size of submission
  - 2.8. Other
- 03. Criteria and Rules
  - 3.1. Reach
  - 3.2. Significance
  - 3.3. Levels
  - 3.4. Contribution of research
  - 3.5. Eligibility
    - 3.5.1. Impact timeframe
    - 3.5.2. Research timeframe
    - 3.5.3. Quality of underpinning research
  - 3.6. Other
- 04. Holistic approach to review
  - 4.1. Allocation
  - 4.2. Format of impact template (REF3a)
  - 4.3. Format of impact case study template (REF3b)
  - 4.4. Relationship between impact case studies and template
  - 4.5. Membership

[65] This approach follows an analytical process such as that outlined in Bazely & Jackson (2013).

- 4.6. Other
- A. Strengths of the process
  - A.1. Suggested good practice
- B. Weaknesses of the process
  - B.1. Suggested improvements
- C. Consequences
- D. Good quotations
- E. Main Panel A
  - E.01. UOA 1
  - E.02. UOA 2
  - E.03. UOA 3
  - E.04. UOA 4
  - E.05. UOA 5
  - E.06. UOA 6
- F. Main Panel B
  - F.01. UOA 7
  - F.02. UOA 8
  - F.03. UOA 9
  - F.04. UOA 10
  - F.05. UOA 11
  - F.06. UOA 12
  - F.07. UOA 13
  - F.08. UOA 14
  - F.09. UOA 15
- G. Main Panel C
  - G.01. UOA 16
  - G.02. UOA 17
  - G.03. UOA 18
  - G.04. UOA 19
  - G.05. UOA 20
  - G.06. UOA 21
  - G.07. UOA 22
  - G.08. UOA 23
  - G.09. UOA 24
  - G.10. UOA 25
  - G.11. UOA 26
- H. Main Panel D
  - H.01. UOA 27
  - H.02. UOA 28
  - H.03. UOA 29
  - H.04. UOA 30
  - H.05. UOA 31
  - H.06. UOA 32
  - H.07. UOA 33
  - H.08. UOA 34
  - H.09. UOA 35
  - H.10. UOA 36
- I. Cross Main Panels
- J. Academic panel member
- K. User member and impact assessor

Appendix D Survey protocols for those involved in the assessment of the impact element of REF 2014

D.1. Academic sub-panel survey

Profile

What was your knowledge of research impact prior to joining the sub-panel?

1. Very limited 2 3 4. Very good

Were you involved with the REF impact process at your own HEI?

Yes No

If you answered yes to the previous question, what was your REF impact role? Please select all that apply:

- Wrote own case study
- Coordinated submissions for school/department
- Wrote impact template
- Other (please specify)

Rules and guidance

To what extent did the rules and guidance you received allow you to:

Assess the impact of the case studies fairly and reliably (where fair means fair to both the institution and all other institutions in the UOA)?

1. Not at all 2 3 4. Very much so

Assess the impact templates fairly and reliably?

1. Not at all 2 3 4. Very much so

Compare different types of impact (i.e. impacts on the economy, society, culture, public policy or services, health, the environment and quality of life) fairly and reliably?

1. Not at all 2 3 4. Very much so

Assess confidential case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the rules and guidance you received allow you to determine the following eligibility of case studies fairly and reliably:

Research time frame

1. Not at all 2 3 4. Very much so

Impact time frame

1. Not at all 2 3 4. Very much so

Quality of underpinning research

1. Not at all 2 3 4. Very much so

Contribution of research to impact

1. Not at all 2 3 4. Very much so

Where underpinning research was conducted

1. Not at all 2 3 4. Very much so

To what extent did the definition of impact allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the criterion of reach allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the criterion of significance allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Was the case study format useful in enabling you to assess case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Was the impact template format useful in enabling you to assess case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Did you receive sufficient training to assess case studies and templates?

1. Not at all 2 3 4. Very much so

Please provide comments on your answers above.

Process of assessment

Were you allocated case studies that you felt you had the expertise to assess? [Options: Yes/No]

Yes No

Where you reviewed both the case studies and templates for one HEI, was the relationship between the case studies and the template clear?

1. Not at all clear 2 3 4. Very Clear

How satisfied were you with the process of moderation and the management of disagreements in relation to the following areas:

All members listened to equally

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Transparency and clarity of the processes

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Equal application of criteria and guidance in every case

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Efficiency of the process

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Did you feel there was effective communication to ensure consistency of assessment across the sub-panels?`

1. Very ineffective 2 3 4. Very effective

How easy did you find it to assess the different types of impact?

Impacts on the economy

1. Very difficult 2 3 4. Very easy

Impacts on society

1. Very difficult 2 3 4. Very easy

Impacts on culture

1. Very difficult 2 3 4. Very easy

Impacts on public policy or services

1. Very difficult 2 3 4. Very easy

Impacts on health

1. Very difficult 2 3 4. Very easy

Impacts on the environment

1. Very difficult 2 3 4. Very easy

Impacts on quality of life

1. Very difficult 2 3 4. Very easy

Please provide comments on your answers above.

Burden of involvement

In answering these questions, please only refer to the additional time taken to act as a sub-panel reviewer, rather than the time you spent on the REF impact exercise as part of your academic role at your HEI, etc. Please provide an *estimate* of time spent, as opposed to referring to records you may have.

How long, in hours, did you spend at meetings discussing impact?

How long, in hours, did you spend reading and assessing case studies and templates independently?

How long, in hours, did you spend travelling to meetings to discuss impact (if additional to time stated above)?

How long, in hours, did you spend on any other preparation or impact assessment activities? Please specify what these additional activities were:

Reflections on impact for the next REF

What were the (up to) three main benefits in participating in the sub-panel assessment for you and your work?

What were the (up to) three main challenges in participating in the sub-panel assessment for you and your work?

What (up to) three changes would you make to improve the assessment process for subsequent REF impact exercises?

Please provide any additional comments you would like to add to the impact assessment process in the box below.

D.2. Impact assessor sub-panel survey

Profile

What was your knowledge of research impact prior to joining the sub-panel?

1. Very limited 2 3 4. Very good

What is your academic background? Please select all that apply:

- PhD
- Masters
- Undergraduate

What is your current engagement with academia?

- Regular engagement
- Ad-hoc engagement
- No engagement

What was your reason for joining the sub-panel?

Which of the following best describes the type of organisation in which you work:

- Private sector (SME/multinational)
- Policymaking body
- Other public sector body
- NGO/charity
- Practice organisation
- Other (please specify)

Rules and guidance

To what extent did the rules and guidance you received allow you to:

Assess the impact of the case studies fairly and reliably (where fair means fair to both the institution and all other institutions in the UOA)?

1. Not at all 2 3 4. Very much so

Assess the impact templates fairly and reliably?

1. Not at all 2 3 4. Very much so

Compare different types of impact fairly and reliably (i.e. impacts on the economy, society, culture, public policy or services, health, the environment and quality of life)?

1. Not at all 2 3 4. Very much so

Assess confidential case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the rules and guidance you received allow you to determine the following eligibility of case studies fairly and reliably:

Research time frame

1. Not at all 2 3 4. Very much so

Impact time frame

1. Not at all 2 3 4. Very much so

Quality of underpinning research

1. Not at all 2 3 4. Very much so

Contribution of research to impact

1. Not at all 2 3 4. Very much so

Where underpinning research was conducted

1. Not at all 2 3 4. Very much so

To what extent did the definition of impact allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the criterion of reach allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent did the criterion of significance allow you to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Was the case study format useful in enabling you to assess case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Was the impact template format useful in enabling you to assess case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Did you receive sufficient training to assess case studies and templates?

1. Not at all 2 3 4. Very much so

Please provide comments on your answers above.

Process of assessment

Were you allocated case studies that you felt you had the expertise to assess?

Yes No

Where you reviewed both the case studies and the templates for one HEI, was the relationship between the case studies and the template clear?

1. Not at all clear 2 3 4. Very Clear

How satisfied were you with the process of moderation and the management of disagreements in relation to the following areas:

All members listened to equally

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Transparency and clarity of the processes

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Equal application of criteria and guidance in every case

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Efficiency of the process

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Did you feel there was effective communication to ensure consistency of assessment across the sub-panels?

1. Very ineffective 2 3 4. Very effective

How easy did you find it to assess the different types of impact?

Impacts on the economy

1. Very difficult 2 3 4. Very easy

Impacts on society

1. Very difficult 2 3 4. Very easy

Impacts on culture

1. Very difficult 2 3 4. Very easy

Impacts on public policy or services

1. Very difficult 2 3 4. Very easy

Impacts on health

1. Very difficult 2 3 4. Very easy

Impacts on the environment

1. Very difficult 2 3 4. Very easy

Impacts on quality of life

1. Very difficult 2 3 4. Very easy

Please provide comments on your answers above.

Burden of involvement

In answering these questions, please only refer to the additional time taken to act as a sub-panel reviewer, rather than the time you spent on the REF impact exercise as part of your academic role at your HEI, etc. Please provide an *estimate* of time spent, as opposed to referring to records you may have.

How long, in hours, did you spend at meetings discussing impact?

How long, in hours, did you spend reading and assessing case studies and templates independently?

How long, in hours, did you spend travelling to meetings to discuss impact (if additional to time stated above)?

How long, in hours, did you spend on any other preparation or impact assessment activities?

Please specify what these additional activities were:

Reflections on impact for the next REF

What were the (up to) three main benefits in participating in the sub-panel assessment for you and your work?

What were the (up to) three main challenges in participating in the sub-panel assessment for you and your work?

What (up to) three changes would you make to improve the assessment process for subsequent REF impact exercises?

Please provide any additional comments you would like to add to the impact assessment process in the box below.

D.3. Adviser/secretariat sub-panel survey

Profile

What was your knowledge of research impact prior to joining the sub-panel?

1. Very limited 2 3 4. Very good

Rules and guidance

To what extent do you believe the rules and guidance allowed sub-panel members to:

Assess the impact of the case studies fairly and reliably (where fair means fair to both the institution and all other institutions in the UOA)?

1. Not at all 2 3 4. Very much so

Assess the impact templates fairly and reliably?

1. Not at all 2 3 4. Very much so

Compare different types of impact (i.e. impacts on the economy, society, culture, public policy or services, health, the environment and quality of life) fairly and reliably?

1. Not at all 2 3 4. Very much so

Assess confidential case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent do you believe the rules and guidance the sub-panel members received allow them to determine the following eligibility of case studies fairly and reliably:

Research time frame

1. Not at all 2 3 4. Very much so

Impact time frame

1. Not at all 2 3 4. Very much so

Quality of underpinning research

1. Not at all 2 3 4. Very much so

Contribution of research to impact

1. Not at all 2 3 4. Very much so

Where underpinning research was conducted

1. Not at all 2 3 4. Very much so

To what extent do you believe the definition of impact allowed sub-panel members to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent do you believe the criterion of reach allowed sub-panel members to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

To what extent do you believe the criterion of significance allowed sub-panel members to assess impact case studies fairly and reliably?

1. Not at all 2 3 4. Very much so

Do you think the briefing sub-panel members and impact assessors received provided sufficient training to allow them to assess case studies and templates?

1. Not at all 2 3 4. Very much so

Please provide comments on your answers above.

Process of assessment

On average, how long did the auditing process for one case study take in hours?

Did you think the process of assessment was fair and robust?

Questions for advisors

Approximately how many case studies were you asked to provide advice on during the process of assessment?

Approximately how many templates were you asked to provide advice on during the process of assessment?

Based on the queries you received, which of these areas do you think the sub-panel found most challenging?

Criterion of reach

1. Not at all challenging	2	3	4. Very challenging

Criterion of significance

1. Not at all challenging	2	3	4. Very challenging

Definition of impact

1. Not at all challenging	2	3	4. Very challenging

Quality of underpinning research

1. Not at all challenging	2	3	4. Very challenging

Research and impact time frames

1. Not at all challenging	2	3	4. Very challenging

Location research carried out

1. Not at all challenging	2	3	4. Very challenging

Contribution of research to impact

1. Not at all challenging	2	3	4. Very challenging

Evidence supporting impact

1. Not at all challenging	2	3	4. Very challenging

Multi-institutional research submissions

1. Not at all challenging | 2 | 3 | 4. Very challenging

Multi- or inter-disciplinary research

1. Not at all challenging | 2 | 3 | 4. Very challenging

Did you feel there were adequate processes in place to ensure the consistency of assessment across the sub-panels?

1. Not at all | 2 | 3 | 4. Very much so

How satisfied were you with the process of moderation and the management of disagreements in relation to the following areas:

All members listened to equally

1. Very dissatisfied | 2 | 3 | 4. Very satisfied | Not assessed | No opinion

Transparency and clarity of the processes

1. Very dissatisfied | 2 | 3 | 4. Very satisfied | Not assessed | No opinion

Equal application of criteria and guidance in every case

1. Very dissatisfied | 2 | 3 | 4. Very satisfied | Not assessed | No opinion

Efficiency of the process

1. Very dissatisfied | 2 | 3 | 4. Very satisfied | Not assessed | No opinion

How effective did you perceive the scoring process to be across sub-panels?

1. Very ineffective | 2 | 3 | 4. Very effective

Please provide comments on your answers above.

Reflections on impact for the next REF

What were the (up to) three main benefits in participating in the sub-panel assessment for you and your work?

What were the (up to) three main challenges in participating in the sub-panel assessment for you and your work?

What (up to) three changes would you make to improve the assessment process for subsequent REF impact exercises?

Please provide any additional comments you would like to add to the impact assessment process in the box below.

D.4. Main panel member survey

Profile

What was your knowledge of research impact prior to joining the panel?

1. Very limited 2 3 4. Very good

Were you involved with the REF impact process at your own HEI?

Yes No

If you answered yes to the previous question, what was your REF impact role? Please select all that apply:

- Wrote own case study
- Coordinated submissions for school/department
- Wrote impact template
- Other (please specify)

Rules and guidance

How far do you believe the guidance was interpreted consistently across sub-panels in your main panel?

1. Not at all 2 3 4. Very much so

Do you think the definition of impact allowed impact case studies to be assessed fairly and reliably (where fair means fair to both the institution and all other institutions in the UOA)?

1. Not at all 2 3 4. Very much so

Do you think the criterion of reach allowed impact case studies to be assessed fairly and reliably?

1. Not at all 2 3 4. Very much so

Do you think the criterion of significance allowed impact case studies to be assessed fairly and reliably?

1. Not at all 2 3 4. Very much so

How far do you believe the definitions of reach and significance were interpreted consistently across sub-panels in your main panel?

1. Not at all 2 3 4. Very much so

Do you think the case study format was useful in enabling case studies to be assessed fairly and reliably?

1. Not at all 2 3 4. Very much so

Do you think the impact template format was useful in enabling case studies to be assessed fairly and reliably?

1. Not at all 2 3 4. Very much so

Process of assessment

How satisfied were you with the process of moderation and the management of disagreements in relation to the following areas:

All members listened to equally

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Transparency and clarity of the processes

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Equal application of criteria and guidance in every case

1. Very dissatisfied 2 3 4. Very satisfied Not assessed No opinion

Efficiency of the process

1. Very dissatisfied | 2 | 3 | 4. Very satisfied | Not assessed | No opinion

Did you feel there was effective communication across the sub-panels to ensure consistency of assessment?

1. Very ineffective | 2 | 3 | 4. Very effective

Did you feel there was effective communication between the sub-panels and the main panel to ensure consistency of assessment?

1. Very ineffective | 2 | 3 | 4. Very effective

How fair and consistent do you think the process of assessment was between the sub-panels falling under your main panel?

1. Very inconsistent | 2 | 3 | 4. Very consistent

How fair and consistent do you think the results of the assessment were between the sub-panels falling under your main panel?

1. Very inconsistent | 2 | 3 | 4. Very consistent

How fair and consistent do you think the process of assessment was between the different main panels?

1. Very inconsistent | 2 | 3 | 4. Very consistent

How fair and consistent do you think the results of the assessment were between the different main panels?

1. Very inconsistent | 2 | 3 | 4. Very consistent

Approximately how many times did you calibrate scores between sub-panels?

Approximately how many times did you calibrate scores between main panels?

Please provide comments on your answers above.

Burden of involvement

In answering these questions, please only refer to the additional time taken to act as a panel reviewer, rather than the time you spent on the REF impact exercise as part of your academic role at your HEI, etc. Please provide an *estimate* of time spent, as opposed to referring to records you may have.

How long, in hours, did you spend at meetings discussing impact?

How long, in hours, did you spend travelling to meetings to discuss impact (if additional to time stated above)?

How long, in hours, did you spend on any other preparation or impact related activities? Please specify what these additional activities were:

Reflections on impact for the next REF

What were the (up to) three main benefits in participating in the sub-panel assessment for you and your work?

What were the (up to) three main challenges in participating in the sub-panel assessment for you and your work?

What (up to) three changes would you make to improve the assessment process for subsequent REF impact exercises?

Please provide any additional comments you would like to add to the impact assessment process in the box below.

D.5. Additional main panel questions (for survey respondents who are already answering another section of the survey)

Rules and guidance

How far do you believe the guidance was interpreted consistently across sub-panels in your main panel?

1. Not at all 2 3 4. Very much so

How far do you believe the definitions of reach and significance were interpreted consistently across sub-panels in your main panel?

1. Not at all 2 3 4. Very much so

Process of assessment

How consistent do you think the process of assessment was between the sub-panels falling under your main panel?

1. Very inconsistent 2 3 4. Very consistent

How fair and consistent do you think the results of the assessment were between the sub-panels falling under your main panel?

1. Very inconsistent 2 3 4. Very consistent

How consistent do you think the process of assessment was between the different main panels?

1. Very inconsistent 2 3 4. Very consistent

How fair and consistent do you think the results of the assessment were between the different main panels?

1. Very inconsistent 2 3 4. Very consistent

Approximately how many times did you calibrate scores between sub-panels?

Approximately how many times did you calibrate scores between main panels?

Please provide comments on your answers above.

Appendix E Protocol for individual interviews with panellists involved in the assessment of impact

1. Could you start by telling me how you became involved in the assessment process of REF 2014?

Prompt: How much did you know about impact prior to your engagement in REF 2014? How were you approached?

2. How were impact case studies and templates allocated in your sub-panel?

Prompt: How many were you allocated? How were allocations divided between panel members and impact assessors? How did this compare to other sub-panels you were aware of? Did you feel you had the expertise to assess the case studies and templates you were assigned?

3. What training and guidance did you receive in preparing for assessing impact case studies and impact templates?

Prompt: How useful was the training and guidance? Did you feel fully prepared to conduct the assessment? What more or less would you like to have received? What input (if any) did you have at the initial briefing stage on the way the assessment would be conducted for your sub-panel? At what point(s) did calibration take place?

4. How did you go about conducting your individual assessment of the impact case studies and impact templates you were assigned?

Prompt: how did you evaluate different types of impact? How did you use the evidence provided? In what circumstances did you raise an audit query? How did you apply the rules and guidance? What was the most difficult thing in judging impact scores for impact case studies? Did you, or anyone in your sub-panels develop a formula to help structure your individual scoring?

5. How did your sub-panel come to a decision on scores for impact case studies and impact templates?

Prompt: What was the process of moderation and validation? Did the Panel agree in advance the parameters in which the various scores for impact could be allocated (i.e. 4 star as opposed to 3 star), and the types of things that would, and would not, be 'counted' as impact – for example, the place of 'public engagement' in the assessment? How formulaically was this applied – did it feel more like a process of discussion and deliberation or the application of a formula? What level of consensus was there around scores from different individuals? What role did different members of the panel play in the decisionmaking process? Do you feel all views were heard and valued appropriately? What level of disagreement was there between the panel, and how was consensus reached around disagreements?

6. For future exercises what would you change and what elements of the impact assessment process would you retain?

Prompt: please tell us about things that worked/should be improved for both you as an individual member and for the whole process.

7. Was the workload what you expected?

Prompt: Can you estimate the time (in hours) you spent on impact during the assessment process. We have categorised this as: meetings discussing impact, reading and assessing case studies and templates, travelling to meetings to discuss impact (if additional to time stated above)? Please give an appropriate proportion of the travel time for meetings where impact formed only part of the discussion, any other impact assessment related activities? [note these questions were asked in the survey and therefore they may already have answered them]. Can you explain why it is difficult to estimate the time taken?

8. Is there anything else you would like the opportunity to say, or that you think we should be covering in our evaluation of the impact element of the assessment that we haven't discussed?