palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-020-0394-7 OPEN

Writing impact case studies: a comparative study of high-scoring and low-scoring case studies from REF2014

Bella Reichard [1], Mark S Reed[1]✉, Jenn Chubb[2], Ged Hall [3], Lucy Jowett [4], Alisha Peart[4] & Andrea Whittle[1]

ABSTRACT This paper reports on two studies that used qualitative thematic and quantitative linguistic analysis, respectively, to assess the content and language of the largest ever sample of graded research impact case studies, from the UK Research Excellence Framework 2014 (REF). The paper provides the first empirical evidence across disciplinary main panels of statistically significant linguistic differences between high- versus low-scoring case studies, suggesting that implicit rules linked to written style may have contributed to scores alongside the published criteria on the significance, reach and attribution of impact. High-scoring case studies were more likely to provide specific and high-magnitude articulations of significance and reach than low-scoring cases. High-scoring case studies contained attributional phrases which were more likely to attribute research and/or pathways to impact, and they were written more coherently (containing more explicit causal connections between ideas and more logical connectives) than low-scoring cases. High-scoring case studies appear to have conformed to a distinctive new genre of writing, which was clear and direct, and often simplified in its representation of causality between research and impact, and less likely to contain expressions of uncertainty than typically associated with academic writing. High-scoring case studies in two Main Panels were significantly easier to read than low-scoring cases on the Flesch Reading Ease measure, although both high-scoring and low-scoring cases tended to be of "graduate" reading difficulty. The findings of our work enable impact case study authors to better understand the genre and make content and language choices that communicate their impact as effectively as possible. While directly relevant to the assessment of impact in the UK's Research Excellence Framework, the work also provides insights of relevance to institutions internationally who are designing evaluation frameworks for research impact.

[1] Newcastle University, Newcastle, UK. [2] University of York, York, UK. [3] University of Leeds, Leeds, UK. [4] Northumbria University, Newcastle, UK.
✉email: mark.reed@ncl.ac.uk

Introduction

Academics are under increasing pressure to engage with non-academic actors to generate "usable" knowledge that benefits society and addresses global challenges (Clark et al., 2016; Lemos, 2015; Rau et al., 2018). This is largely driven by funders and governments that seek to justify the societal value of public funding for research (Reed et al., 2020; Smith et al., 2011) often characterised as 'impact'. While this has sometimes been defined narrowly as reflective of the need to demonstrate a return on public investment in research (Mårtensson et al., 2016; Tsey et al., 2016; Warry, 2006), there is also a growing interest in the evaluation of "broader impacts" from research (cf. Bozeman and Youtie, 2017; National Science Foundation, 2014), including less tangible but arguably equally relevant benefits for society and culture. This shift is exemplified by the assessment of impact in the UK's Research Excellence Framework (REF) in 2014 and 2021, the system for assessing the quality of research in UK higher education institutions, and in the rise of similar policies and evaluation systems in Australia, Hong Kong, the United States, Horizon Europe, The Netherlands, Sweden, Italy, Spain and elsewhere (Reed et al., 2020).

The evaluation of research impact in the UK has been criticised by scholars largely for its association with a 'market logic' (Olssen and Peters, 2005; Rhoads and Torres, 2005). Critics argue that a focus of academic performativity can be seen to "destabilise" professional identities (Chubb and Watermeyer, 2017), which in the context of research impact evaluation can further "dehumanise and deprofessionalise" academic performance (Watermeyer, 2019), whilst leading to negative unintended consequences (which Derrick et al., 2018, called "grimpact"). MacDonald (2017), Chubb and Reed (2018) and Weinstein et al. (2019) reported concerns from researchers that the impact agenda may be distorting research priorities, "encourag[ing] less discovery-led research" (Weinstein et al., 2019, p. 94), though these concerns were questioned by University managers in the same study who were reported to "not have enough evidence to support that REF was driving specific research agendas in either direction" (p. 94), and further questioned by Hill (2016).

Responses to this critique have been varied. Some have called for civil disobedience (Watermeyer, 2019) and organised resistance (Back, 2015; MacDonald, 2017) against the impact agenda. In a review of Watermeyer (2019), Reed (2019) suggested that attitudes towards the neoliberal political roots of the impact agenda may vary according to the (political) values and beliefs of researchers, leading them to pursue impacts that either support or oppose neoliberal political and corporate interests. Some have defended the benefits of research impact evaluation. For example, Weinstein et al. (2019) found that "a focus on changing the culture outside of academia is broadly valued" by academics and managers. The impact agenda might enhance stakeholder engagement (Hill, 2016) and give "new currency" to applied research (Chubb, 2017; Watermeyer, 2019). Others have highlighted the long-term benefits for society of incentivising research impact, including increased public support and funding for a more accountable, outward-facing research system (Chubb and Reed, 2017; Hill, 2016; Nesta, 2018; Oancea, 2010, 2014; Wilsdon et al., 2015).

In the UK REF, research outputs and impact are peer reviewed at disciplinary level in 'Units of Assessment' (36 in 2014, 34 in 2021), grouped into four 'Main Panels'. Impact is assessed through case studies that describe the effects of academic research and are given a score between 1∗ ("recognised but modest") and 4∗ ("outstanding"). The case studies follow a set structure of five sections: 1—Summary of the impact; 2—Underpinning research; 3—References to the research; 4—Details of the impact; 5—Sources to corroborate the impact (HEFCE, 2011). The publication of over 6000 impact case studies in 2014[1] by Research England (formerly Higher Education Funding Council for England, HEFCE) was unique in terms of its size, and unlike the recent selective publication of high-scoring case studies from Australia's 2018 Engagement and Impact Assessment, both high-scoring and low-scoring case studies were published. This provides a unique opportunity to evaluate the construction of case studies that were perceived by evaluation panels to have successfully demonstrated impact, as evidenced by a 4∗ rating, and to compare these to case studies that were judged as less successful.

The analysis of case studies included in this research is based on the definition of impact used in REF2014, as "an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" (HEFCE, 2011, p. 26). According to REF2014 guidance, the primary functions of an impact case study were to articulate and evidence the *significance* and *reach* of impacts arising from research beyond academia, clearly demonstrating the contribution that research from a given institution contributed to those impacts (HEFCE, 2011).

In addition to these explicit criteria driving the evaluation of impact in REF2014, a number of analyses have emphasised the role of implicit criteria and subjectivity in shaping the evaluation of impact. For example, Pidd and Broadbent (2015) emphasised the implicit role a "strong narrative" plays in high-scoring case studies (p. 575). This was echoed by the fears of one REF2014 panellist interviewed by Watermeyer and Chubb (2018) who said, "I think with impact it is literally so many words of persuasive narrative" as opposed to "giving any kind of substance" (p. 9). Similarly, Watermeyer and Hedgecoe (2016), reporting on an internal exercise at Cardiff University to evaluate case studies prior to submission, emphasised that "style and structure" were essential to "sell impact", and that "case studies that best sold impact were those rewarded with the highest evaluative scores" (p. 651).

Recent research based on interviews with REF2014 panellists has also emphasised the subjectivity of the peer-review process used to evaluate impact. Derrick's (2018) research findings based on panellist interviews and participant observation of REF2014 sub-panels argued that scores were strongly influenced by who the evaluators were and how the group assessed impact together. Indeed, a panellist interviewed by Watermeyer and Chubb (2018) concurred that "the panel had quite an influence on the criteria" (p. 7), including an admission that some types of (more intangible) evidence were more likely to be overlooked than other (more concrete) forms of evidence, "privileg[ing] certain kinds of impact". Other panellists interviewed spoke of their emotional and intellectual vulnerability in making judgements about an impact criterion that they had little prior experience of assessing (Watermeyer and Chubb, 2018). Derrick (2018) argued that this led many evaluators to base their assessments on more familiar proxies for excellence linked to scientific excellence, which led to biased interpretations and shortcuts that mimicked "groupthink" (p. 193).

This paper will for the first time empirically assess the content and language of the largest possible sample of research impact case studies that received high versus low scores from assessment panels in REF2014. Combining qualitative thematic and quantitative linguistic analysis, we ask:

1. How do high-scoring versus low-scoring case studies articulate and evidence impacts linked to underpinning research?
2. Do high-scoring and low-scoring case studies have differences in their linguistic features or styles?

3. Do high-scoring and low-scoring case studies have lexical differences (words and phrases that are statistically more likely to occur in high- or low-scoring cases) or text-level differences (including reading ease, narrative clarity, use of cohesive devices)?

By answering these questions, our goal is to provide evidence for impact case study authors and their institutions to reflect on in order to optimally balance the content and to use language that communicates their impact as effectively as possible. While directly relevant to the assessment of impact in the UK's REF, the work also provides insights of relevance to institutions internationally who are designing evaluation frameworks for research impact.

Methods

Research design and sample. The datasets were generated by using published institutional REF2014 impact scores to deduce the scores of some impact case studies themselves. Although scores for individual case studies were not made public, we were able to identify case studies that received the top mark of 4∗ based on the distribution of scores received by some institutions, where the whole submission by an institution in a given Unit of Assessment was awarded the same score. In those 20 Units of Assessment (henceforth UoA) where high-scoring case studies could be identified in this way, we also accessed all case studies known to have scored either 1∗ or 2∗ in order to compare the features of high-scoring case studies to those of low-scoring case studies.

We approached our research questions with two separate studies, using quantitative linguistic and qualitative thematic analysis respectively. The thematic analysis, explained in more detail in the section "Qualitative thematic analysis" below, allowed us to find answers to research question 1 (see above). The quantitative linguistic analysis was used to extract and compare typical word combinations for high-scoring and low-scoring case studies, as well as assessing their readability. It mainly addressed research questions 2 and 3.

The quantitative linguistic analysis was based on a sample of all identifiable high-scoring case studies in any UoA ($n = 124$) and all identifiable low-scoring impact case studies in those UoAs where high-scoring case studies could be identified ($n = 93$). As the linguistic analysis focused on identifying characteristic language choices in running text, only those sections designed to contain predominantly text were included (1—Summary of the impact; 2—Underpinning research; 4—Details of the impact). Figure 1 shows the distribution of case studies across Main Panels in the quantitative analysis. Table 1 summarises the number of words included in the analysis.

In order to detect patterns of content in high-scoring and low-scoring case studies across all four Main Panels, a sub-sample of case studies was selected for a qualitative thematic analysis. This included 60% of high-scoring case studies and 97% of low-scoring case studies from the quantitative analysis, such that only UoAs were included where both high-scoring and low-scoring case studies are available (as opposed to the quantitative sample, which includes all available high-scoring case studies). Further selection criteria were then designed to create a greater balance in the number of high-scoring and low-scoring case studies across Main Panels. Main Panels A (high) and C (low) were particularly over-represented, so a lower proportion of those case studies were selected and 10 additional high-scoring case studies were considered in Panel B, including institutions where at least 85% of the case studies scored 4∗ and the remaining scores were 3∗. As this added a further UoA, we could also include 14 more low-scoring case studies in Main Panel B. This resulted in a total of 85 high-scoring and 90 low-scoring case studies. Figure 2 shows the distribution of case studies across Main Panels in the thematic analysis, illustrating the greater balance compared to the sample used in the quantitative analysis. The majority (75%) of the case studies analysed are included in both samples (Table 2).

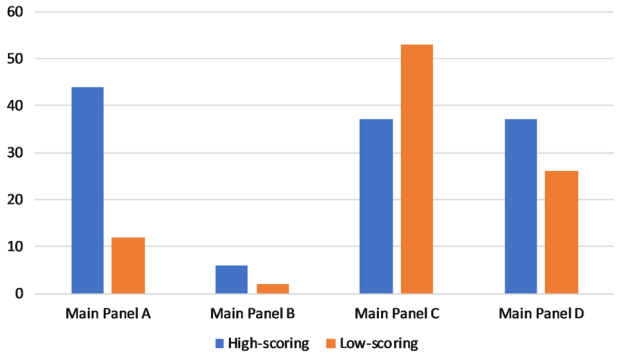

Fig. 1 Distribution of case studies across Main Panels used for the linguistic analysis sample.

Table 1 Number of words in case studies used for the linguistic analysis sample, by Main Panel (MP).

Main panel	High-scoring cases–number of words	Low-scoring cases–number of words	Total–number of words
MP A	69,267	16,262	85,529
MP B	11,021	3,291	14,312
MP C	69,836	73,771	143,607
MP D	70,607	37,958	108,565
Total	**220,731**	**131,282**	**352,013**

Quantitative linguistic analysis. Quantitative linguistic analysis can be used to make recurring patterns in language use visible and to assess their significance. We treated the dataset of impact case studies as a text collection (the 'corpus') divided into two sections, namely high-scoring and low-scoring case studies (the two 'sub-corpora'), in order to explore the lexical profile and the readability of the case studies.

One way to explore the lexical profile of groups of texts is to generate frequency-based word lists and compare these to word lists from a reference corpus to determine which words are characteristic of the corpus of interest ("keywords", cf. Scott, 1997). Another way is to extract word combinations that are particularly frequent. Such word combinations, called "lexical bundles", are "extended collocations" (Hyland, 2008, p. 41) that appear across a set range of texts (Esfandiari and Barbary, 2017). We merged these two approaches in order to uncover meanings that could not be made visible through the analysis of single-word frequencies, comparing lexical bundles from each sub-corpus to the other. Lexical bundles of 2–4 words were extracted with AntConc (specialist software developed by Anthony, 2014) firstly from the corpus of all high-scoring case studies and then separately from the sub-corpora of high-scoring case studies in Main Panel A, C and D.[2] The corresponding lists were extracted from low-scoring case studies overall and separated by panel. The lists of lexical bundles for each of the high-scoring corpus parts were then compared to the corresponding low-scoring parts (High-Overall vs. Low-Overall, High-Main Panel A vs. Low-Main Panel A, etc.) to detect statistically significant over-use and under-use in one set of texts relative to another.

Two statistical measures were used in the analysis of lexical bundles. Log Likelihood was used as a measure of the statistical significance of frequency differences (Rayson and Garside, 2000), with a value of >3.84 corresponding to $p < 0.05$. This measure had the advantage, compared to the more frequently used chi-square test, of not assuming a normal distribution of data (McEnery et al., 2006). The Log Ratio (Hardie, 2014) was used as a measure of effect size, which quantifies the scale, rather than the statistical significance, of frequency differences between two datasets. The Log Ratio is technically the binary log of the relative risk, and a value of >0.5 or <−0.5 is considered meaningful in corpus linguistics (Hardie, 2014), with values further removed from 0 reflecting a bigger difference in the relative frequencies found in each corpus. There is currently no agreed standard effect size measure for keywords (Brezina, 2018, p. 85) and the Log Ratio was chosen because it is straightforward to interpret. Each lexical bundle that met the 'keyness' threshold (Log Likelihood > 3.84 in the case of expected values > 12, with higher significance levels needed for expected values < 13—see Rayson et al., 2004, p. 8) was then assigned a code according to its predominant meaning in the texts, as reflected in the contexts captured in the concordance lines extracted from the corpus.

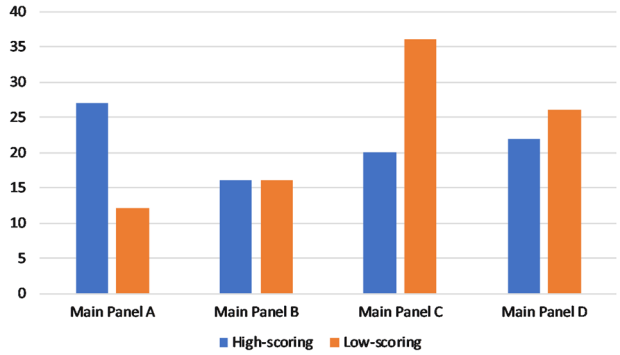

Fig. 2 Distribution of case studies across Main Panels used for the thematic analysis sample.

In the thematic analysis, it appeared that high-scoring case studies were easier to read. In order to quantify the readability of the texts, we therefore analysed them using the Coh-Metrix online tool (www.cohmetrix.com, v3.0) developed by McNamara et al. (2014). This tool provides 106 descriptive indices of language features, including 8 principal component scores developed from combinations of the other indices (Graesser et al., 2011). We selected these principal component scores as comprehensive measures of "reading ease" because they assess multiple characteristics of the text, up to whole-text discourse level (McNamara et al., 2014, p. 78). This was supplemented by the traditional and more wide-spread Flesch Reading Ease score of readability measuring the lengths of words and sentences, which are highly correlated with reading speed (Haberlandt and Graesser, 1985). The selected measures were compared across corpus sections using *t*-tests to evaluate significance. The effect size was measured using Cohen's *D*, following Brezina (2018, p. 190), where $D > 0.3$ indicates a small, $D > 0.5$ a medium, and $D > 0.8$ a high effect size. As with the analysis of lexical bundles,

Table 2 Overview of units of assessment and ratings included in each sample.

Main Panel	Unit of Assessment Name	UoA Number	Quantitative linguistic sample			Qualitative thematic sample		
			4* per UoA	1*/2* per UoA	Total	4* per UoA	1*/2* per UoA	Total
A	Clinical medicine	1	15	0	15	0	0	0
A	Public health, health services and primary care	2	8	0	8	0	0	0
A	Allied Health Professions, Dentistry, Nursing and Pharmacy	3	6	2	8	12	2	14
A	Psychology, Psychiatry and Neuroscience	4	10	6	16	10	6	**16**
A	Agriculture, Veterinary and Food Science	6	5	4	9	5	4	**9**
Total			**44**	**12**	**56**	**27**	**12**	**39**
B	Computer science	11	0	0	0	10	14	24
B	Electrical and Electronic Engineering, Metallurgy and Materials	13	4	2	6	6	2	8
B	Civil and Construction engineering	14	2	0	2	0	0	0
Total			**6**	**2**	**8**	**16**	**16**	**32**
C	Economics and econometrics	18	3	0	3	0	0	0
C	Law	20	3	4	7	3	4	7
C	Social work and social policy	22	16	6	22	6	6	12
C	Sociology	23	3	5	8	3	5	8
C	Anthropology and development studies	24	2	0	2	0	0	0
C	Education	25	8	22	30	6	11	17
C	Sport and Exercise Sciences, Leisure and Tourism	26	2	16	18	2	10	12
Total			**37**	**53**	**90**	**20**	**36**	**56**
D	Area studies	27	5	0	5	0	0	0
D	Modern Languages and Linguistics	28	2	2	4	2	2	4
D	English Language and Literature	29	12	8	20	6	8	14
D	History	30	2	4	6	2	4	6
D	Music, Drama, Dance and Performing Arts	35	10	6	16	6	6	12
D	Communication, Cultural and Media Studies, Library and Information Management	36	6	6	12	6	6	12
Total			**37**	**26**	**63**	**22**	**26**	**48**
		Total	**124**	**93**	**217**	**85**	**90**	**175**

comparisons were made between high- and low-scoring case studies in each of Main Panels A, C and D, as well as between all high-scoring and all low-scoring case studies across Main Panels.

Qualitative thematic analysis. While a quantitative analysis as described above can make differences in the use of certain words visible, it does not capture the narrative or content of the texts under investigation. In order to identify common features of high-scoring and low-scoring case studies, thematic analysis was chosen to complement the quantitative analysis by identifying patterns and inferring meaning from qualitative data (Auerbach and Silverstein, 2003; Braun and Clarke, 2006; Saldana, 2009). To familiarise themselves with the data and for inter-coder reliability, two research team members read a selection of REF2014 impact case studies from different Main Panels, before generating initial codes for each of the five sections of the impact case study template. These were discussed with the full research team, comprising three academic and three professional services staff who had all read multiple case studies themselves. They were piloted prior to defining a final set of themes and questions against which the data was coded (based on the six-step process outlined by Braun and Clarke, 2006) (Table 3). An additional category was used to code stylistic features, to triangulate elements of the quantitative analysis (e.g. readability) and to include additional stylistic features difficult to assess in quantitative terms (e.g. effective use of testimonials). In addition to this, 10 different types of impact were coded for, based on Reed's (2018) typology: capacity and preparedness, awareness and understanding, policy, attitudinal change, behaviour change and other forms of decision-making, other social, economic, environmental, health and well-being, and cultural impacts. There was room for coders to include additional insights arising in each section of the case study that had not been captured in the coding system; and there was room to summarise other key factors they thought might account for high or low scores.

Coders summarised case study content pertaining to each code, for example by listing examples of effective or poor use of structure and formatting as they arose in each case study. Coders also quoted the original material next to their summaries so that their interpretation could be assessed during subsequent analysis. This initial coding of case study text was conducted by six coders,

Table 3 Themes and questions that guided the qualitative analysis of case studies, structured around the main sections of the REF2014 impact case study template (titles in bold).

Theme or question
Case study title
Unit of Assessment
Impact type: understanding and awareness
Impact type: attitudinal
Impact type: economic
Impact type: environmental
Impact type: health and wellbeing
Impact type: policy
Impact type: other forms of decision-making and behavior change
Impact type: cultural
Impact type: other social
Impact type: capacity or preparedness
Additional types of impact not currently included in typology
Overall, what features might account for this being a high or low scoring case study?
Underpinning research and references to the research
Do the titles of publications/journals fit to the UoA? If no, quote example publications that suggest poor fit
Are there indications that the research is likely to be >2*? Provide examples of indications that research may or may not reach threshold
Are the research findings described concisely and clearly? Quote example text
Is the underpinning research adequately linked to the claimed impacts?
Other examples of good/poor practice in underpinning research and references to the research that may account for scores?
Summary and details of the impact
Is the framing of reach justified (and how)?
How does the pathway to impact contribute towards high/low score?
Evidence of ineligible content? Quote examples
Are the claims for impact credible? Are there any doubts or concerns that would lead you to distrust the claims? Quote examples
Is pedagogy a major component of this impact case study?
Is public engagement a major component of this impact case study?
To what extent does the case study argue effectively the case that impacts ultimately arose, or does it focus only/mainly on the pathway/engagement?
How clearly articulated and evidenced are the benefits? Quote examples
How clearly are beneficiaries identified? Quote examples
Other examples of good/poor practice in the summary and details of the impact that may account for scores?
Corroborating evidence
Examples of high or low quality corroborating evidence with justification for why high/low
Does the impact stand alone without reading the corroborating evidence?
Other examples of good/poor practice in corroborating evidence that may account for scores?
Structure and style
Examples of effective or poor use of structure and formatting
Easy or hard to read (e.g. academic jargon, acronyms) by non-specialist? Examples of effective/poor language
Are adjectives used appropriately e.g. backed up with evidence to justify their use or used as unsubstantiated claims? Quote examples
Examples of effective/poor use of testimonials?
Other examples of good/poor practice in structure and style that may account for scores?

with intercoder reliability (based on 10% of the sample) assessed at over 90%. Subsequent thematic analysis *within* the codes was conducted by two of the co-authors. This involved categorising coded material into themes as a way of assigning meaning to features that occurred across multiple case studies (e.g. categorising types of corroborating evidence typically used in high-scoring versus low-scoring case studies).

Results and discussion

In this section, we integrate findings from the quantitative linguistic study and the qualitative analysis of low-scoring versus high-scoring case studies. The results are discussed under four headings based on the key findings that emerged from both analyses. Taken together, these findings provide the most comprehensive evidence to date of the characteristics of a top-rated (4*) impact case study in REF2014.

Highly-rated case studies provided specific, high-magnitude and well-evidenced articulations of significance and reach. One finding from our qualitative thematic analysis was that 84% of high-scoring cases articulated benefits to specific groups and provided evidence of their significance and reach, compared to 32% of low-scoring cases which typically focused instead on the *pathway* to impact, for example describing dissemination of research findings and engagement with stakeholders and publics without citing the benefits arising from dissemination or engagement. One way of conceptualising this difference is using the content/process distinction: whereas low-scoring cases tended to focus on the *process* through which impact was sought (i.e. the pathway used), the high-scoring cases tended to focus on the *content* of the impact itself (i.e. what change or improvement occurred as a result of the research).

Examples of global reach were evidenced across high-scoring case studies from all panels (including Panel D for Arts and Humanities research), but were less often claimed or evidenced in low-scoring case studies. Where reach was more limited geographically, many high-scoring case studies used context to create robust arguments that their reach was impressive in that context, describing reach for example in social or cultural terms or arguing for the importance of reaching a narrow but hard-to-reach or otherwise important target group.

Table 4 provides examples of evidence from high-scoring cases and low-scoring cases that were used to show significance and reach of impacts in REF2014.

Findings from the quantitative linguistic analysis in Table 5 show how high-scoring impact case studies contained more phrases that *specified* reach (e.g. "in England and", "in the US"), compared to low-scoring case studies that used the more generic term "international", leaving the reader in doubt about the actual reach. They also include more phrases that implicitly specified the significance of the impact (e.g. "the government's" or "to the House of Commons"), compared to low-scoring cases which provided more generic phrases, such as "policy and practice", rather than detailing specific policies or practices that had been changed.

The quantitative linguistics analysis also identified a number of words and phrases pertaining to engagement and pathways, which were intended to deliver impact but did not actually specify impact (Table 6). A number of phrases contained the word "dissemination", and there were several words and phrases specifying types of engagement that could be considered more one-way dissemination than consultative or co-productive (cf. Reed et al.'s (2018) engagement typology), e.g. "the book" and "the event". The focus on dissemination supports the finding from the qualitative thematic analysis that low-scoring case tended to focus more on pathways or routes than on impact. Although it is not possible to infer this

Table 4 Examples of evidence used to show significance and reach of impacts from research in high and low-scoring impact case studies from REF2014.

Examples from high-scoring case studies	Examples from low-scoring case studies
Significance	
• Evidence of benefits for specific beneficiary groups that have happened during the eligibility period (rather than anticipated future impacts) • Evidence is shown to come from credible sources and is used to substantiate specific claims, e.g. official data showing 430% increase in approvals of biopesticides, or peer-reviewed analysis showing that the BBC changed its coverage based on recommendations from research • Evidence that a new policy or practice works and has delivered benefits (e.g. via an internal or external independent review, primary or secondary data collection or testimonials) or limiting the claim to changes in policy or practice (where it is too early to assess their effect) • Use of robust research or evaluation designs to evidence impact, with robustness demonstrated through triangulation for qualitative and mixed methods evaluations, or through statistical significance and treatment-control designs (e.g. randomised control trials)	• Research leads to an activity or other pathway, but with no evidence that these pathways led to *actual* impacts (in some cases the claim is for *potential* future impacts) • Evidence is used vaguely, e.g. "evaluative data indicate the majority of users have…changed the way they work" without describing the number of users or the nature of the change • The impact of future policy implementation is claimed (or implied), but the evidence only relates to policy formation • Poorly designed evaluation undermines credibility of evidence, e.g. no baseline, before/after or comparison group to demonstrate changes were a result of the research • Testimonials describe impacts of their organisation rather than the research, or describe engagement with researchers but no impacts • Over-reliance on estimates (e.g. in testimonials) without more concrete evidence
Reach	
• Addressing a challenge that was uniquely felt by a particular group on a sub-national scale • Successfully helping hard-to-reach groups that others have previously not been able to reach • Reaching significantly more than previous initiatives, e.g. poetry events that attracted "twice the national average for such events" • Evidence of strong pathways to impact from well-respected international organisations or groups with strong influence at other relevant scales, for example via funding for research or dissemination of research via policy documents or new working practices	• Reach is claimed internationally or across multiple groups (sometimes implicitly), but convincing evidence is only presented for national (or sub-national) benefits or for a small proportion of the groups who are said to have benefited • Claims of reach based on the global reach of an organisation or initiative using the output of research without specifying the impact the research activity or output has had on this organisation

Table 5 Examples of lexical bundles that were common in the high-scoring case studies and largely absent from the low-scoring case studies–Significance and reach.

	Search term	Appears in Section	Example
Significance	the government's	4	"cited in *the Government's* consultation document" *"The government's* anti-poverty strategy" *"The Government's* education policy"
	the department for	4	*All occurrences refer to government departments, mostly UK but some US*
	the department	2 and 4	*As above, but:* "work done by *the Department* around the topic" (*inside a testimonial quote, referring to university department*)
	produced by	4	"has informed the revised media guidelines *produced by* the Samaritans" "the impact of the Unit's work has been visible in key documents *produced by* the UK Government and the EU"
	(to) the house of	4	Lords: 13x Commons: 15x
Reach	in the US	4	*"In the US* the stimulator is sold by…" "legislation on immigration reform *in the US"* "impact on policy makers *in the US"*
	the UK's	4 but also 2	*"the UK's* investment strategy" *"the UK's* regulatory body" *"the UK's* poorest children"
	of the UK	1 and 4	*"of the UK* general population" "six different sectors *of the UK* economy" "the socio-cultural context *of the UK"*
	in England	1 and 4	"SIDS deaths *in England and* Wales" "has been adopted widely *in England and* Wales" "the largest … study … to be conducted *in England and* one of only three worldwide"
	the UK and	2 but also 1 and some 4	"three key sites across *the UK and* Europe" "supports major health policy change in *the UK and* informed…" "facilitate … in *the UK and* therefore"

Table 6 Examples of Lexical bundles that were common in the low-scoring case studies and largely absent from the high-scoring case studies–Describing pathway.

Search term	Appears in Section	Problem	Example
involved in	2 and 4	If applied to the researcher, this phrase does not convey agency; involved in what ways?	"Stakeholders were *involved in* the work" "[the institution's] invitation to be *involved in* the EU Framework 7" "[the researcher] was *involved in* the project"
has been disseminated	Mainly 4	"disseminated" is one-directional – what happened then?	"The research *has been disseminated* through a report" "The work *has been disseminated* through publication in international journals" "it *has been disseminated* to larger coaching groups"
the event	4	What event? Could be more specific.	"impact was demonstrated through *the event*" "those attending *the event* feel strongly" "*The event* was advertised in the [local newspaper]"

directly from the data, it is possible that this may represent a deeper epistemological position underpinning some case studies, where impact generation was seen as one-way knowledge or technology transfer, and research findings were perceived as something that could be given unchanged to publics and stakeholders through dissemination activities, with the assumption that this would be understood as intended and lead to impact.

It is worth noting that none of the four UK countries appear significantly more often in either high-scoring or low-scoring case studies (outside of the phrase "in England and"). Wales ($n = 50$), Scotland ($n = 71$) and Northern Ireland ($n = 32$) appear slightly more often in high-scoring case studies, but the difference is not significant (England: $n = 162$). An additional factor to take into account is that our dataset includes only submissions that are either high-scoring or low-scoring, and the geographical spread of the submitting institutions was not a factor in selecting texts. There was a balanced number of high-scoring and low-scoring case studies in the sample from English, Scottish and Welsh universities, but no guaranteed low-scoring submissions from Northern Irish institutions. The REF2014 guidance made it clear that impacts in each UK country would be evaluated equally in comparison to each other, the UK and other countries. While the

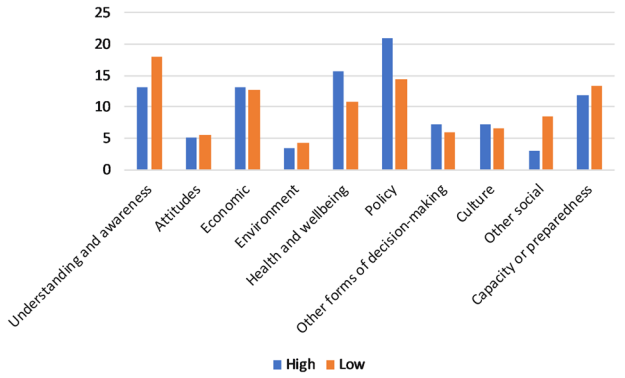

Fig. 3 Number of impacts claimed in high- versus low-scoring case studies by impact type.

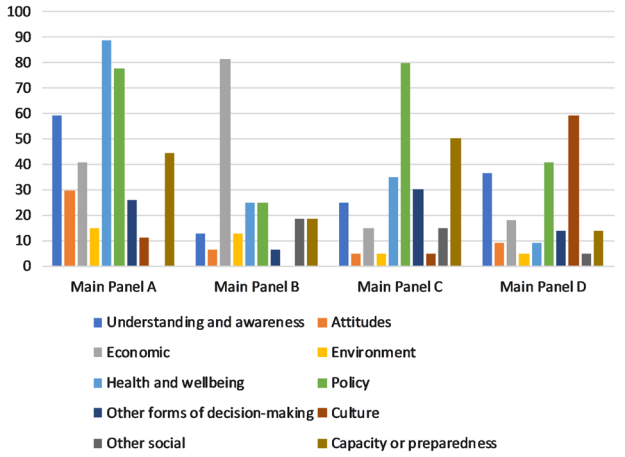

Fig. 4 Percentage of high-scoring case studies that claimed different types of impact.

quantitative analysis of case studies from our sample only found a statistically significant difference for the phrase "in England and", this, combined with the slightly higher number of phrases containing the other countries of the UK in high-scoring case studies, might indicate that this panel guidance was implemented as instructed.

Figures 3–5 shows which types of impact could be identified in high-scoring or low-scoring case studies, respectively, in the qualitative thematic analysis (based on Reed's (2018) typology of impacts). Note that percentages do not add up to 100% because it was possible for each case study to claim more than one type of impact (high-scoring impact case studies described on average 2.8 impacts, compared to an average of 1.8 impacts described by low-scoring case studies)[3]. Figure 3 shows the number of impacts per type as a percentage of the total number of impacts claimed in high-scoring versus low-scoring case studies. This shows that high-scoring case studies were more likely to claim health/wellbeing and policy impacts, whereas low-scoring case studies were more likely to claim understanding/awareness impacts. Looking at this by Main Panel, over 50% of high-scoring case studies in Main Panel A claimed health/wellbeing, policy and understanding/awareness impacts (Fig. 4), whereas over 50% of low-scoring case studies in Main Panel A claimed capacity building impacts (Fig. 5). There were relatively high numbers of economic and policy claimed in both high-scoring and low-scoring case studies in Main Panels B and C, respectively, with no impact type dominating strongly in Main Panel D (Figs. 4 and 5).

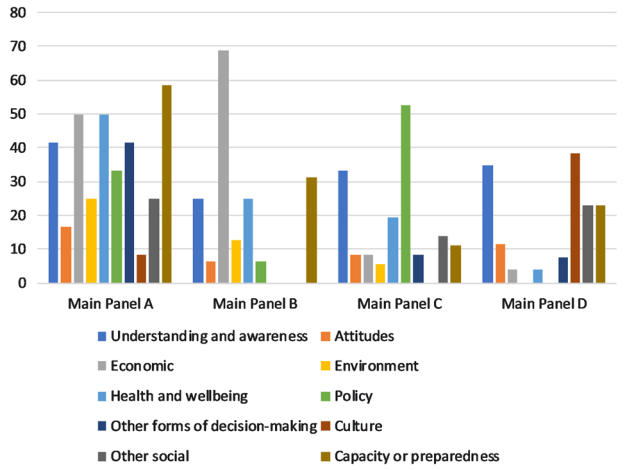

Fig. 5 Percentage of low-scoring case studies that claimed different types of impact.

Highly-rated case studies used distinct features to establish links between research (cause) and impact (effect). Findings from the quantitative linguistic analysis show that high-scoring case studies were significantly more likely to include attributional phrases like "cited in", "used to" and "resulting in", compared to low-scoring case studies (Table 7 provides examples for some of the 12 phrases more frequent in high-scoring case studies). However, there were some attributional phrases that were more likely to be found in low-scoring case studies (e.g. "from the", "of the research" and "this work has"—total of 9 different phrases).

To investigate this further, all 564 and 601 instances[4] of attributional phrases in high-scoring and low-scoring case studies, respectively, were analysed to categorise the context in which they were used, to establish the extent to which these phrases in each corpus were being used to establish attribution to impacts. The first word or phrase preceding or succeeding the attributional content was coded. For example, if the attributional content was "used the", followed by "research to generate impact", the first word succeeding the attributional content (in this case "research") was coded rather than the phrase it subsequently led to ("generate impact"). According to a Pearson Chi Square test, high-scoring case studies were significantly more likely to establish attribution to impact than low-scoring cases ($p < 0.0001$, but with a small effect size based on Cramer's $V = 0.22$; bold in Table 8). 18% ($n = 106$) of phrases in the low-scoring corpus established attribution to impact, compared to 37% ($n = 210$) in the high-scoring corpus, for example, stating that research, pathway or something else *led to* impact. Instead, low-scoring case studies were more likely to establish attribution to research (40%; $n = 241$) compared to high-scoring cases (28%; $n = 156$; $p < 0.0001$, but with a small effect size based on Cramer's $V = 0.135$). Both high- and low-scoring case studies were similarly likely to establish attribution to pathways (low: 32%; $n = 194$; high: 31% $n = 176$).

Moreover, low-scoring case studies were more likely to include ambiguous or uncertain phrases. For example, the phrase "a number of" can be read to imply that it is not known how many instances there were. This occurred in all sections of the impact case studies, for example in the underpinning research section as "The research explores *a number of* themes" or in the summary or details of the impact section as "The work has also resulted in *a number of* other national and international impacts", or "has influenced approaches and practices of *a number of* partner organisations". Similarly, "an impact on" could give the impression that the nature of the impact is not known. This

Table 7 Examples of lexical bundles that were common in the high-scoring case studies and largely absent from the low-scoring case studies–Attribution.

Search term	Appears in Section	Example
led by Professor	Start of 2	*usually followed by name but sometimes by specialism and name; often preceded by* "team"/"group" or "studies"/"research"
cited in	4	"*cited in* the guideline on Organ donation" "*cited in* the Mental Health Strategy for Scotland" "*cited in* the Financial Times"
used to	4, 3x in 1	"*used to* inform and target a range of strategies" "our survey methods and evaluation measures are *used to* assess [...] quality" "has been *used to* inform Government policy"
improve the	throughout	"to *improve the* nation's public health" "to *improve the* availability of data" "to *improve the* quality of teaching and learning"
resulting in	throughout	"*resulting in* a funded study" "Based on this research, [company] updated its [...] guidelines [...] *resulting in* cheaper [...] costs"

Table 8 Contexts in which attributional phrases in high and low-scoring case studies were used.

	High		Low	
	Percentage	**Number (total: 564)**	**Percentage**	**Number (total: 601)**
To the left of the attributional phrase				
Research	57	319	34	206
Pathway[1]	23	130	21	125
Impact	8	45	10	59
Other[2]	12	70	20	118
None (sentence starts with attributional phrase)	0	0	15	93
To the right of the attributional phrase				
Research	28	**156**	40	**241**
Pathway[1]	31	176	32	195
Impact	37	**210**	18	**104**
Other[2]	4	21	8	51
None (sentence ends with attributional phrase)	<1	1	1	8

[a]Material coded as "pathway" included activities that could lead to impact (e.g. training), but did not articulate a clear immediate benefit yet (e.g. change in understanding or benefits from using skills learned during training). The names of beneficiaries, e.g. in the excerpts "The UK Pesticides Safety Directorate used the research...", "UK Pesticides Safety Directorate", were coded as "pathway".
[b]"Other" (text in bold) included evidence, problem statements and other contextual material, and fragments such as "**First**, **on** the basis of our findings" or "**As well as being** *used to* inform...". Where fragments clearly referred to a previous sentence, e.g. "**This** was *used to*" or "**they** *used the*", the subject of the previous sentence was coded (e.g. research if "this" referred to a paper, or pathway if "they" referred to a beneficiary) (attributional phrases in italics).

phrase occurred only in summary and details of the impact sections, for example, "These activities have had *an impact on* the professional development", "the research has had *an impact on* the legal arguments", or "there has also been *an impact on* the work of regional agency".

In the qualitative thematic analysis, we found that only 50% of low-scoring case studies clearly linked the underpinning research to claimed impacts (compared to 97% of high-scoring cases). This gave the impression of over-claimed impacts in some low-scoring submissions. For example, one case study claimed "significant impacts on [a country's] society" based on enhancing the security of a new IT system in the department responsible for publishing and archiving legislation. Another claimed "economic impact on a worldwide scale" based on billions of pounds of benefits, calculated using an undisclosed method by an undisclosed evaluator in an unpublished final report by the research team. One case study claimed attribution for impact based on similarities between a prototype developed by the researchers and a product subsequently launched by a major corporation, without any evidence that the product as launched was based on the prototype. Similar assumptions were made in a number of other case studies that appeared to conflate correlation with causation in their attempts to infer attribution between research and impact. Table 9 provides examples of different ways in which links between research and impact were evidenced in the details of the research section.

Table 10 shows how corroborating sources were used to support these claims. 82% of high-scoring case studies compared to 7% of low-scoring cases were identified in the qualitative thematic analysis as having generally high-quality corroborating evidence. In contrast, 11% of high-scoring case studies, compared to 71% of low-scoring cases, were identified as having corroborating evidence that was vague and/or poorly linked to claimed impacts. Looking at only case studies that claim policy impact, 11 out of 26 high-scoring case studies in the sample described both policy and implementation (42%), compared to just 5 out of 29 low-scoring case studies that included both policy and implementation (17%; the remainder described policy impacts only with no evidence of benefits arising from implementation). High- scoring case studies were more likely to cite evidence of impacts rather than just citing evidence pertaining to the pathway (which was more common in low-scoring cases). High-scoring policy case studies also provided evidence pertaining to the pathway, but because they typically also included evidence of policy change, this evidence helped attribute policy impacts to research.

Highly-rated case studies were easy to understand and well written. In preparation for the REF, many universities invested heavily in writing assistance (Coleman, 2019) to ensure that impact case studies were "easy to understand and evaluation-friendly" (Watermeyer and Chubb, 2018) for the assessment panels, which comprised academics and experts from other sectors (HEFCE, 2011, p. 6). With this in mind, we investigated readability and style, both in the quantitative linguistic and in the qualitative thematic analysis.

High-scoring impact case studies scored more highly on the Flesch Reading Ease score, a readability measure based on the

Table 9 Examples of different ways in which links between research and impact were evidenced in REF2014 case studies.

Examples of how links between research and impact were evidenced	Examples of problems establishing links between research and impact
Complete causal chain	
• Description of pathways to impact demonstrates causal chain from impact all the way back to research, with each link in the chain evidenced clearly • All claimed impacts clearly arise from the research	• Research leads to an activity or other pathway, but with no evidence that these pathways led to impacts • Claims that research was used without explaining how or to what effect • Cause and effect implied but not stated or evidenced explicitly • Link to research only established for some (not all) impacts claimed • Important missing links in causal chains from research to impact • The nature of the claim means it would be impossible to attribute impact to the research (this was acknowledged explicitly in some cases)
Policy	
• Citation of the research in policy documents, often supported by testimonials detailing the contribution that the research made	• Policy change that co-incidentally matches research recommendations without citation or testimony to demonstrate the change was linked to research
Spin-out companies	
• Spin-out companies that commercialise specific research findings	• Spin-out companies that work in a similar area to the research with no explicit link between products/services and specific research findings, or whose main activities are not linked to the research
Link to research	
• Clear distinction between research, pathways to impact and impact, showing how excellent research led to impact • Impacts (in section 4, "details of the impact") mapped against research findings (in Section 2, "underpinning research")	• Descriptions of underpinning research that describes the pathway to impact more than (or instead of) the originality, significance and rigour of the research, making it difficult to identify the research findings that impacts have arisen from • No explicit reference back to underpinning research in the description of impact
Convincing pathway	
• Research was commissioned by organisation that implemented findings • Other evidence of close collaboration and buy-in from early in research process e.g. via researchers in organisational roles or placements, researchers as practitioners, or evidence of embeddedness of researchers with community or culture	• Limited information about pathway to impact means causal links between research and impact are implicit only, rather than explicitly described and credible

length of words and sentences. The scores in Table 11 are reported out of 100, with a higher score indicating that a text is easier to read. While the scores reveal a significant difference between 4* and 1*/2* impact case studies, they also indicate that impact case studies are generally on the verge of "graduate" difficulty (Hartley, 2016, p. 1524). As such our analysis should not be understood as suggesting that these technical documents should be adjusted to the readability of a newspaper article, but they should be maintained at interested and educated non-specialist level.

Interestingly, there were differences between the main panels.[5] In Social Science and Humanities case studies (Main Panels C and D), high-scoring impact case studies scored significantly higher on reading ease than low-scoring ones. There was no significant difference in Main Panel A between 4* and 1*/2* cases. However, all Main Panel A case studies showed, on average, lower reading ease scores than the low-scoring cases in Main Panels C and D. This means that their authors used longer words and sentences, which may be explained in part by more and longer technical terms needed in Main Panel A disciplines; the difference between high- and low-scoring case studies in Main Panels C and D may be explained by the use of more technical jargon (confirmed in the qualitative analysis).

The Flesch Reading Ease measure assesses the sentence- and word-level, rather than capturing higher-level text-processing difficulty. While this is recognised as a reliable indicator of comparative reading ease, and the underlying measures of sentence-length and word-length are highly correlated with reading speed (Haberlandt and Graesser, 1985), Hartley (2016) is right in his criticism that the tool takes neither the meaning of the words nor the wider text into account. The Coh-Metrix tool (McNamara et al., 2014) provides further measures for reading ease based on textual cohesion in these texts compared to a set of general English texts. Of the eight principal component scores computed by the tool, most did not reveal a significant difference between high- and low-scoring case studies or between different Main Panels. Moreover, in most measures, impact case studies overall were fairly homogenous compared to the baseline of general English texts. However, there were significant differences between high- and low-scoring impact case studies in two of the measures: "deep cohesion" and "connectivity" (Table 12).

"Deep cohesion" shows whether a text makes causal connections between ideas explicit (e.g. "because", "so") or leaves them for the reader to infer. High-scoring case studies had a higher level of deep cohesion compared to general English texts (Graesser et al., 2011), while low-scoring case studies tended to sit below the general English average. In addition, Main Panel A case studies (Life Sciences), which received the lowest scores in Flesch Reading Ease, on average scored higher on deep cohesion than case studies in more discursive disciplines (Main Panel C—Social Sciences and Main Panel D—Arts and Humanities). "Connectivity" measures the level of explicit logical connectives (e.g. "and", "or" and "but") to show relations in the text. Impact case studies were low in connectivity compared to general English texts, but within each of the Main Panels, high-scoring case studies had more explicit connectivity than low-scoring case studies. This means that Main Panel A case studies, while using on average longer words and sentences as indicated by the Flesch Reading Ease scores, compensated for this by making causal and logical relationships more explicit in the texts. In Main Panels C and D, which on average scored lower on these measures, there was a clearer difference between high- and low-scoring case studies than in Main Panel A, with high-scoring case studies being easier to read.

Table 10 Examples of corroborating evidence identified from qualitative analysis of high versus low-scoring REF2014 case studies.

Examples of corroborating evidence from high-scoring case studies	Examples of corroborating evidence from low-scoring case studies
Credibility of sources	
• Testimonials from high-level stakeholders in highly relevant organisations, e.g. NHS and WHO • Independent evidence from other research teams; highly credible organisations, e.g. WHO report or secondary data sources (e.g. Government statistics) • Peer-reviewed evidence of impact from impact case study authors e.g. showing impact on computing speed or RCTs, quote from journal article by a museum's Head of Research showing impact of research on curatorial practice	Potential conflicts of interest undermine credibility of source, for example: • A case study corroborated by testimonials from those who commissioned the research • A publisher commenting on the success of the book they published • Statements on spin-out company websites • Unpublished or non-peer-reviewed reports by the team responsible for the impact • Testimonial from staff at submitting unit
Evidence of pathways versus impacts	
• Evidence of claimed impacts, e.g. links to NICE guidelines or new industry standard explaining how and where research is cited, evidence of audience or visitor numbers • Link to Government press release showing a policy was based on research by the submitting unit • Testimonials about the impact of the research contained in media reports • Evidence of policy engagement to attribute impact to research in cases where policy impacts were achieved • Evidence of impacts arising from evidence-based policy, rather than just evidence of policy change	• Download figures and other statistics relating to pathway rather than reach of impact • A funding proposal (e.g. original Knowledge Transfer Partnership application) • Collaboration agreements • Links to project websites and Facebook pages • Lists of media coverage without explaining what impact they evidence • Link to training materials rather than evidence that training had benefits • Links to conference and other presentations • Evidence of policy engagement with no evidence of policy impacts • Evidence of policy change in contexts where there are doubts over likelihood of implementation or enforcement • Evidence of policy change without explaining which aspects were linked to the research
Eligibility of impacts evidenced	
• Only eligible impacts are evidenced	• Evidence of potential future interest, rather than retrospective impact claims • Evidence that research was cited by other researchers • Indicators of esteem such as keynote presentations, invitation to contribute articles to *The Guardian* newspaper
Specificity and link to impacts	
• Narrative explaining what each source corroborates with references to page numbers where relevant • Corroborating evidence is provided for all claimed impacts	• Lists of names (with or without positions and affiliations) that do not state what the person is able to corroborate (and are not cross-referenced to a quote from a testimonial in the case study) • Lists of hyperlinks, reports or other forms of evidence that are not cited in the "Details of the impact" section and do not explain what claims they evidence • Generic customer service email address to corroborate impact • Lists of research outputs without explaining how they corroborate impacts • No evidence provided to support key claims, e.g. missing economic data or testimonials to corroborate economic impact • Missing evidence for claimed impacts, e.g. a single piece of corroborating evidence from one individual beneficiary saying they were using an endangered language in a new way • Claim for causality based on similarity of two devices is not supported as an image/ specification is only given for one of the devices

Linked to this, low-scoring case studies across panels were more likely than high-scoring case studies to contain phrases linked to the research process (suggesting an over-emphasis on the research rather than the impact, and a focus on process over findings or quality; Table 18) and filler-phrases (Table 13).

High-scoring case studies were more likely to clearly identify individual impacts via subheadings and paragraph headings ($p < 0.0001$, with effect size measure Log Ratio 0.54). The difference is especially pronounced in Main Panel D (Log Ratio 1.53), with a small difference in Main Panel C and no significant difference in Main Panel A. In Units of Assessment combined in Main Panel D, a more discursive academic writing style is prevalent (see e.g. Hyland, 2002) using fewer visual/typographical distinctions such as headings. The difference in the number of headings used in case studies from those disciplines suggests that high-scoring case studies showed greater divergence from disciplinary norms than low-scoring case studies. This may have allowed them to adapt the presentation of their research impact to the audience of panel members to a greater extent than low-scoring case studies.

The qualitative thematic analysis of Impact Case Studies indicates that it is not simply the number of subheadings that matters, although this comparison is interesting especially in the context of the larger discrepancy in Main Panel D. Table 14 summarises formatting that was considered helpful and unhelpful from the qualitative analysis.

The observations in Tables 11–13 stem from quantitative linguistic analysis, which, while enabling statistical testing, does not show directly the effect of a text on the reader. When conducting the qualitative thematic analysis, we collected examples of formatting and stylistic features from the writing

Table 11 Flesch reading ease scores for high- and low-scoring impact case studies.

	High-scoring case studies	Low-scoring case studies	p (one-tailed t-test)	Cohen's D (effect size)
Overall	30.9	27.5	<0.01**	>0.4
Main Panel A	28.4	26.2	>0.05	<0.3
Main Panel C	32.3	27.4	<0.001***	>0.5
Main Panel D	32.8	28.3	<0.05*	>0.3

Table 12 Deep cohesion and connectivity - difference between high- and low-scoring impact case studies.

	Deep Cohesion-p (one-tailed t-test)	Deep Cohesion-Cohen's D (effect size)	Connectivity-p (one-tailed t-test)	Connectivity-Cohen's D (effect size)
Overall	<0.001	>0.5	<0.001	>0.5
Main Panel A	-	-	<0.05	>0.5
Main Panel C	<0.01	>0.5	<0.001	>0.8
Main Panel D	<0.01	>0.5	-	-

Table 13 Examples of Lexical bundles that were common in the low-scoring case studies and largely absent from the high-scoring case studies-filler phrases.

Search term	Appears in	Example
in terms of	across sections	"children have benefited *in terms of* enhanced [...] awareness" "research into [...] is demonstrated *in terms of* its reach by citation in" "The impact *in terms of* awareness-raising"
the way(s) in which	across sections	"[researcher's] work was significant in *the way in which* the [...] were devised" "evidence for *the way in which* coaches influence"
in relation to	2 and 4	"*In relation to* (i) participants disclosed that" "This is important *in relation to* two approaches" "[Researcher's] work *in relation to* [research topic] has led to"

Table 14 Examples of formatting identified from qualitative analysis of high and low-scoring REF2014 case studies.

Examples of formatting from high-scoring case studies	Examples of formatting from low-scoring case studies
Headings	
• Meaningful and consistent	• There is a danger of breaking the text up too much at the expense of a coherent narrative
• Correspond to structure that may be signposted in Section 1 (or at start of relevant Section)	• Headings which are titles of research projects or names of researchers can give the impression that these are the focus of the case study, rather than the impact
• One or two levels of subheadings	
Bullet points, lists	
• List of testimonials	• Bullets announce a list that is then not fully elaborated on
• Details of impact by beneficiary	• Points don't link together
• Highlighting the central research questions of projects	• Danger of highlighting irrelevant details and therefore *weakening* the claim for reach and significance
• In Section 2 breaking down research findings	
Bold or italics	
• Bold is used for impacts, beneficiaries, researcher names, dates, references to Section 3/5	• Italics are less effective for impacts/beneficiaries
• Italics for testimonial quotes	• Testimonials as block quotations can give the impression of taking over from the main narrative

and presentation of high and low-scoring case studies that might have affected clarity of the texts (Tables 14 and 15). Specifically, 38% of low-scoring case studies made inappropriate use of adjectives to describe impacts (compared to 20% of high-scoring; Table 16). Inappropriate use of adjectives may have given an impression of over-claiming or created a less factual impression than case studies that used adjectives more sparingly to describe impacts. Some included adjectives to describe impacts in testimonial quotes, giving third-party endorsement to the claims rather than using these adjectives directly in the case study text.

Highly-rated case studies were more likely to describe underpinning research findings, rather than research processes. To be eligible, case studies in REF2014 had to be based on underpinning research that was "recognised internationally in terms of originality, significance and rigour" (denoted by a 2* quality profile, HEFCE, 2011, p. 29). Ineligible case studies were excluded from our sample (i.e. those in the "unclassifiable" quality profile), so all the case studies should have been based on strong research. Once this research quality threshold had been passed, scores were based on the significance and reach of impact, so case studies with higher-rated research should not, in theory, get better scores on

Table 15 Examples of stylistic features identified from qualitative analysis of high and low-scoring REF2014 case studies.

Feature	Stylistic features in high-scoring case studies	Stylistic features in low-scoring case studies
Clarity of writing	• Simple style and vocabulary • Claims are made directly • Avoids long, complex sentences and breaks text into paragraphs, sub-sections and lists where relevant	• Long sentences, unnecessarily complex language • Text not broken up, poor organisation • Hard to follow even if technical vocabulary is not used • Long-winded descriptions, poor explanations
Use of technical jargon and acronyms	• Avoids "isms" and "lenses" • Explains necessary technical terms and context • Spells out (sparingly used) acronyms	• Especially in crucial places e.g. when describing the impact • Too much background knowledge is assumed • Jargon disguises how vague the claims are • Unexplained technical terms and acronyms • Over-use of acronyms makes text difficult to follow
Narrative progression	• Narrative clearly shows progression	• No coherent narrative linking research to pathways and impacts or linking different pathways and impacts together • Spelling mistakes and grammatical errors • Swapping between first and third person

Table 16 Examples of use of adjectives that may have given an impression of over-claiming or may have cast doubts on claims, identified from qualitative analysis of REF2014 impact case studies.

Inappropriate use	Examples
Unsubstantiated use of adjectives giving impression of over-claiming	Adjectives such as "promising", "significant", "invested heavily", "excellent", "fundamental", "expanding rapidly" were over-used across a number of cases and were often unsubstantiated
Vague use of adjectives weakening or casting doubt on claims	• Claims of impact on "many" without a definition of "many" • "Substantial" is used to describe estimate of millions of dollars of benefit, drawing attention to the fact that there is no specific number and it is only an estimate • "Accumulated impact" implies impact was incremental or is only emerging slowly • "Very well received and some very valuable feedback" without being able provide examples casts doubt on the claim

the basis of their underpinning research. However, there is evidence that units whose research outputs scored well in REF2014 also performed well on impact (unpublished Research England analysis cited in Hill, 2016). This observation only shows that high-quality research and impact were co-located, rather than demonstrating a causal relationship between high-quality research and highly rated impacts. However, our qualitative thematic analysis suggests that weaker descriptions of research (underpinning research was not evaluated directly) may have been more likely to be co-located with lower-rated impacts at the level of individual case studies. We know that the majority of underpinning research in the sample was graded 2* or above (because we excluded unclassifiable case studies from the analysis) but individual ratings for outputs in the underpinning research section are not provided in REF2014. Therefore, the qualitative analysis looked for a range of indicators of strong or weak research in four categories: (i) indicators of publication quality; (ii) quality of funding sources; (iii) narrative descriptions of research quality; and (iv) the extent to which the submitting unit (versus collaborators outside the institution) had contributed to the underpinning research. As would be expected (given that all cases had passed the 2* threshold), only a small minority of cases in the sample gave grounds to doubt the quality of the underpinning research. However, both our qualitative and quantitative analyses identified research-related differences between high- and low-scoring impact case studies.

Based on our qualitative thematic analysis of indicators of research quality, a number of low-scoring cases contained indications that underpinning research may have been weak. This was very rare in high-scoring cases. In the most extreme case, one case study was not able to submit any published research to underpin the impact, relying instead on having secured grant funding and having a manuscript under review. Table 17 describes indicators that underpinning research may have been weaker (presumably closer to the 2* quality threshold for eligibility). It also describes the indications of higher quality research (which were likely to have exceeded the 2* threshold) that were found in the rest of the sample. High-scoring case studies demonstrated the quality of the research using a range of direct and indirect approaches. Direct approaches included the construction of arguments that articulated the originality, significance and rigour of the research in the "underpinning research" section of the case study (sometimes with reference to outputs that were being assessed elsewhere in the exercise to provide a quick and robust check on quality ratings). In addition to this, a wide range of indirect proxies were used to infer quality, including publication venue, funding sources, reviews and awards.

These indicators are of particular interest given the stipulation in REF2021 that case studies must provide evidence of research quality, with the only official guidance suggesting that this is done via the use of indicators. The indicators identified in Table 17 overlap significantly with example indicators proposed by panels in the REF2021 guidance. However, there are also a number of additional indicators, which may be of use for demonstrating the quality of research in REF2021 case studies. In common with proposed REF2021 research quality indicators, many of the indicators in Table 17 are highly context dependent, based on subjective disciplinary norms that are used as short-cuts to assessments of quality by peers within a given context. Funding sources, publication venues and reviews that are considered prestigious in one disciplinary context are often perceived very differently in other disciplinary contexts. While REF2021 does not allow the use of certain indicators (e.g. journal impact factors), no comment is given on the appropriateness of the suggested indicators. While this may be problematic, given that an indicator by definition sign-posts, suggests or indicates by

Table 17 Indications that underpinning research was stronger (likely to have exceeded this threshold) or weaker (and likely to have been closer to the 2* quality threshold) in REF2014 case studies.

Type of indicator	Indications of stronger research likely to have exceeded 2* threshold	Indications of weaker underpinning research, closer to 2* quality threshold
Publication quality	• Peer-reviewed in journals that are well-regarded within the discipline, even if journals are not highly ranked • Monographs published by respected academic publishers • Reviews in broadsheet newspapers, specialist magazines and awards (or nominations), coupled with translation into multiple languages (Main Panel D) • Research met the inclusion criteria for a systematic review (Main Panels A-C)	• Absence of peer-reviewed work or "in press" with no DOI (Main Panels A-C) • Publications only in magazines targeted at practitioners e.g. trade magazines • Underpinning research consists only of narrative literature review or other pieces with no original research • Publication in apparently "predatory" journals with limited, poor or no peer review • Reliance on conference papers or lectures in disciplines where this is not widely respected • Books published by non-academic publishers
Funding	• Peer-reviewed funding from sources considered prestigious in the Unit of Assessment the case study was submitted to (even if small total amounts)	• Less prestigious, non-peer-reviewed funding sources • Non peer-reviewed reports submitted as underpinning research
Narrative description of underpinning research quality	• Strong narrative justification of originality, significance and rigour • Awards for research and/or researchers showing academic recognition	• Work described in ways that suggests no original or significant knowledge was generated, e.g. the University's role in a project was to generate impact from research done by other partners
Contribution to underpinning research		• Researcher is far down authorship lists with no explanation of the significance of their role in the work • Research included that is unrelated to the impact in the case study • Work based on a funded network where it is not clear if the research emerged from network members or the submitting institution

Table 18 Examples of Lexical bundles that were common in the low-scoring case studies and largely absent from the high-scoring case studies–Research.

Search term	Appears in Section	Problem	Examples
the paper; journal of; peer-reviewed; et al.	mainly 2	Names of research output and justification of quality should be placed in Section 3 and referenced with a number in Section 2 (and 4).	"to be published in the International *Journal of*" "This research featured in two *peer-reviewed* publications" "the paper by [name] *et al.* (date)"
relationship between	2 and 4	Often framing a research question – but appears vague if used to describe impact.	"the *relationship between* learning disability and sport" "the *relationship between* national security and the protection of fundamental rights" "the project enhanced the *relationship between*"

proxy rather than representing the outcome of any rigorous assessment, we make no comment on whether it is appropriate to judge research quality via such proxies. Instead, Table 17 presents a subjective, qualitative identification of indicators of high or low research quality, which were as far as possible considered within the context of disciplinary norms in the Units of Assessments to which the case studies belonged.

The quantitative linguistic analysis also found differences between the high-scoring and low-scoring case studies relating to underpinning research. There were significantly more words and phrases in low-scoring case studies compared to high-scoring cases relating to research outputs (e.g. "the paper", "peer-reviewed", "journal of", "et al"), the research process (e.g. "research project", "the research", "his work", "research team") and descriptions of research ("relationship between", "research into", "the research") (Table 18). The word "research" itself appears frequently in both (high: 91× per 10,000 words; low: 110× per 10,000 words), which is nevertheless a small but significant over-use in the low-scoring case studies (effect size measure log ratio = 0.27, $p < 0.0001$).

There are two alternative ways to interpret these findings. First, the qualitative research appears to suggest a link between higher-quality underpinning research and higher impact scores. However, the causal mechanism is not clear. An independent review of REF2014 commissioned by the UK Government (Stern, 2016) proposed that underpinning research should only have to meet the 2* threshold for rigour, as the academic significance and novelty of the research is not in theory a necessary precursor to significant and far-reaching impact. However, a number of the indications of weaker research in Table 17 relate to academic significance and originality, and many of the indicators that suggested research exceeded the 2* threshold imply academic

significance and originality (e.g. more prestigious publication venues often demand stronger evidence of academic significance and originality in addition to rigour). As such, it may be possible to posit two potential causal mechanisms related to the originality and/or significance of research. First, it may be argued that major new academic breakthroughs may be more likely to lead to impacts, whether directly in the case of applied research that addresses societal challenges in new and important ways leading to breakthrough impacts, or indirectly in the case of major new methodological or theoretical breakthroughs that make new work possible that addresses previously intractable challenges. Second, the highest quality research may have sub-consciously biased reviewers to view associated impacts more favourably. Further research would be necessary to test either mechanism.

However, these mechanisms do not explain the higher frequency of words and phrases relating to research outputs and process in low-scoring case studies. Both high-scoring and low-scoring cases described the underpinning research, and none of the phrases that emerged from the analysis imply higher or lower quality of research. We hypothesised that this may be explained by low-scoring case studies devoting more space to underpinning research at the expense of other sections that may have been more likely to contribute towards scores. Word limits were "indicative", and the real limit of "four pages" in REF2014 (extended to five pages in REF2021) was operationalised in various way. However, a *t*-test found no significant difference between the underpinning research word counts (mean of 579 and 537 words in high and low-scoring case studies, respectively; $p = 0.11$). Instead, we note that words and phrases relating to research in the low-scoring case studies focused more on descriptions of research outputs and processes rather than descriptions of research findings or the quality of research, as requested in REF2014 guidelines. Given that eligibility evidenced in this section is based on whether the research findings underpin the impacts and the quality of the research (HEFCE, 2011), we hypothesise that the focus of low-scoring case studies on research outputs and processes was unnecessary (at best) or replaced or obscured research findings (at worst). This could be conceptualised as another instance of the content/process distinction, whereby high-scoring case studies focused on what the research found and low-scoring case studies focused on the process through which the research was conducted and disseminated. It could be concluded that this tendency may have contributed towards lower scores if unnecessary descriptions of research outputs and process, which would not have contributed towards scores, used up space that could otherwise have been used for material that may have contributed towards scores.

Limitations

These findings may be useful in guiding the construction and writing of case studies for REF2021 but it is important to recognise that our analyses are retrospective, showing examples of what was judged to be 'good' and 'poor' practice in the authorship of case studies for REF2014. Importantly, the findings of this study should not be used to infer a causal relationship between the linguistic features we have identified and the judgements of the REF evaluation panel. Our quantitative analysis has identified similarities and differences in their linguistic features, but there are undoubtedly a range of considerations taken into account by evaluation panels. It is also not possible to anticipate how REF2021 panels will interpret guidance and evaluate case studies, and there is already evidence that practice is changing significantly across the sector. This shift in expectations regarding impact is especially likely to be the case in research concerned with public policy, which are increasingly including policy implementation as well as design in their requirements, and research involving public engagement, which is increasingly being expected to provide longitudinal evidence of benefits and provide evidence of cause and effect. We are unable to say anything conclusive from our sample about case studies that focused primarily on public engagement and pedagogy because neither of these types of impact were common enough in either the high-scoring or low-scoring sample to infer reliable findings. While this is the largest sample of known high-scoring versus low-scoring case studies ever analysed, it is important to note that this represents <3% of the total case studies submitted to REF2014. Although the number of case studies was fairly evenly balanced between Main Panels in the thematic analysis, the sample only included a selection of Units of Assessment from each Main Panel, where sufficient numbers of high and low-scoring cases could be identified (14 and 20 out of 36 Units of Assessment in the qualitative and quantitative studies, respectively). As such, caution should be taken when generalising from these findings.

Conclusion

This paper provides empirical insights into the linguistic differences in high-scoring and low-scoring impact case studies in REF2014. Higher-scoring case studies were more likely to have articulated evidence of significant and far-reaching impacts (rather than just presenting the activities used to reach intended future impacts), and they articulated clear evidence of causal links between the underpinning research and claimed impacts. While a cause and effect relationship between linguistic features, styles and the panel's evaluation cannot be claimed, we have provided a granularity of analysis that shows how high-scoring versus low-scoring case studies attempted to meet REF criteria. Knowledge of these features may provide useful lessons for future case study authors, submitting institutions and others developing impact assessments internationally. Specifically, we show that high-scoring case studies were more likely to provide specific and high-magnitude articulations of significance and reach, compared to low-scoring cases, which were more likely to provide less specific and lower-magnitude articulations of significance and reach. Lower-scoring case studies were more likely to focus on pathways to impact rather than articulating clear impact claims, with a particular focus on one-way modes of knowledge transfer. High-scoring case studies were more likely to provide clear links between underpinning research and impacts, supported by high-quality corroborating evidence, compared to low-scoring cases that often had missing links between research and impact and were more likely to be underpinned by corroborating evidence that was vague and/or not clearly linked to impact claims. Linked to this, high-scoring case studies were more likely to contain attributional phrases, and these phrases were more likely to attribute research and/or pathways to impact, compared to low-scoring cases, which contained fewer attributional phrases, which were more likely to provide attribution to pathways rather than impact. Furthermore, there is evidence that high-scoring case studies had more explicit causal connections between ideas and more logical connective words (and, or, but) than low-scoring cases.

However, in addition to the explicit REF2014 rules, which appear to have been enacted effectively by sub-panels, there is evidence that implicit rules, particularly linked to written style, may also have played a role. High-scoring case studies appear to have conformed to a distinctive new genre of writing, which was clear and direct, often simplified in its representation of causality between research and impact, and less likely to contain expressions of uncertainty than might be normally expected in academic writing (cf. e.g. Vold, 2006; Yang et al., 2015). Low-scoring case

studies were more likely to contain filler phrases that could be described as "academese" (Biber and Gray, 2019, p. 1), more likely to use unsubstantiated or vague adjectives to describe impacts, and were less likely to signpost readers to key points using sub-headings and paragraph headings. High-scoring case studies in two Main Panels (out of the three that could be analysed in this way) were significantly easier to read, although both high- and low-scoring case studies tended to be of "graduate" (Hartley, 2016) difficulty.

These findings suggest that aspects of written style may have contributed towards or compromised the scores of some case studies in REF2014, in line with previous research emphasising the role of implicit and subjective factors in determining the outcomes of impact evaluation (Derrick, 2018; Watermeyer and Chubb, 2018). If this were the case, it may raise questions about whether case studies are an appropriate way to evaluate impact. However, metric-based approaches have many other limitations and are widely regarded as inappropriate for evaluating societal impact (Bornmann et al., 2018; Pollitt et al., 2016; Ravenscroft et al., 2017; Wilsdon et al., 2015). Comparing research output evaluation systems across different countries, Sivertsen (2017) presents the peer-review-based UK REF as "best practice" compared to the metrics-based systems elsewhere. Comparing the evaluation of impact in the UK to impact evaluations in USA, the Netherlands, Italy and Finland, Derrick (2019) describes REF2014 and REF2021 as "the world's most developed agenda for evaluating the wider benefits of research and its success has influenced the way many other countries define and approach the assessment of impact".

We cannot be certain about the extent to which linguistic features or style shaped the judgement of REF evaluators, nor can such influences easily be identified or even consciously recognised when they are at work (cf. research on sub-conscious bias and tacit knowledge; the idea that "we know more than we can say"—Polanyi, 1958 cited in Goodman, 2003, p. 142). Nonetheless, we hope that the granularity of our findings proves useful in informing decisions about presenting case studies, both for case study authors (in REF2021 and other research impact evaluations around the world) and those designing such evaluation processes. In publishing this evidence, we hope to create a more "level playing field" between institutions with and without significant resources available to hire dedicated staff or consultants to help write their impact case studies.

Data availability

The dataset analysed during the current study corresponds to the publicly available impact case studies defined through the method explained in Section "Research design and sample" and Table 2. A full list of case studies included can be obtained from the corresponding author upon request.

Received: 10 July 2019; Accepted: 9 January 2020;
Published online: 25 February 2020

Notes

1 https://impact.ref.ac.uk/casestudies/search1.aspx
2 For Main Panel B, only six high-scoring and two low-scoring case studies are clearly identifiable and available to the public (cf. Fig. 1). The Main Panel B dataset is therefore too small for separate statistical analysis, and no generalisations should be made on the basis of only one high-scoring and one low-scoring submission.
3 However, in the qualitative analysis, there were a similar number of high-scoring case studies that were considered to have reached this score due to a clear focus on one single, highly impressive impact, compared to those that were singled out for their impressive range of different impacts.
4 Note that there were more instances of the smaller number of attributional phrases in the low-scoring corpus.
5 For Main Panel B, only six high-scoring and two low-scoring case studies are clearly identifiable and available to the public. The Main Panel B dataset is therefore too small for separate statistical analysis, and no generalisations should be made on the basis of only one high-scoring and one low-scoring submission.

References

Anthony L (2014) AntConc, 3.4.4 edn. Waseda University, Tokyo
Auerbach CF, Silverstein LB (2003) Qualitative data: an introduction to coding and analyzing data in qualitative research. New York University Press, New York, NY
Back L (2015) On the side of the powerful: the 'impact agenda' and sociology in public. https://www.thesociologicalreview.com/on-the-side-of-the-powerful-the-impact-agenda-sociology-in-public/. Last Accessed 24 Jan 2020
Biber D, Gray B (2019) Grammatical complexity in academic English: linguistic change in writing. Cambridge University Press, Cambridge
Bornmann L, Haunschild R, Adams J (2018) Do altmetrics assess societal impact in the same way as case studies? An empirical analysis testing the convergent validity of altmetrics based on data from the UK Research Excellence Framework (REF). J Informetr 13(1):325–340
Bozeman B, Youtie J (2017) Socio-economic impacts and public value of government-funded research: lessons from four US National Science Foundation initiatives. Res Policy 46(8):1387–1398
Braun V, Clarke V (2006) Using thematic analysis in psychology. Quale Res Psychol 3(2):77–101
Brezina V (2018) Statistics in corpus linguistics: a practical guide. Cambridge University Press, Cambridge
Chubb J (2017) Instrumentalism and epistemic responsibility: researchers and the impact agenda in the UK and Australia. University of York
Chubb J, Watermeyer R (2017) Artifice or integrity in the marketization of research impact? Investigating the moral economy of (pathways to) impact statements within research funding proposals in the UK and Australia. Stud High Educ 42(2):2360–2372
Chubb J, Reed MS (2017) Epistemic responsibility as an edifying force in academic research: investigating the moral challenges and opportunities of an impact agenda in the UK and Australia. Palgrave Commun 3:20
Chubb J, Reed MS (2018) The politics of research impact: academic perceptions of the implications for research funding, motivation and quality. Br Politics 13 (3):295–311
Clark WC et al. (2016) Crafting usable knowledge for sustainable development. Proc Natl Acad Sci USA 113(17):4570–4578
Coleman I (2019) The evolution of impact support in UK universities. Cactus Communications Pvt. Ltd
Derrick G (2018) The evaluators' eye: impact assessment and academic peer review. Palgrave Macmillan
Derrick G (2019) Cultural impact of the impact agenda: implications for social sciences and humanities (SSH) research. In: Bueno D et al. (eds.), Higher education in the world, vol. 7. Humanities and higher education: synergies between science, technology and humanities. Global University Network for Innovation (GUNi)
Derrick G et al. (2018) Towards characterising negative impact: introducing Grimpact. In: Proceedings of the 23rd international conference on Science and Technology Indicators (STI 2018). Centre for Science and Technology Studies (CWTS), Leiden, The Netherlands
Esfandiari R, Barbary F (2017) A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. J Engl Academic Purp 29:21–42
Goodman CP (2003) The tacit dimension. Polanyiana 2(1):133–157
Graesser AC, McNamara DS, Kulikowich J (2011) Coh-Metrix: providing multilevel analyses of text characteristics. Educ Res 40:223–234
Haberlandt KF, Graesser AC (1985) Component processes in text comprehension and some of their interactions. J Exp Psychol: Gen 114(3):357–374
Hardie A (2014) Statistical identification of keywords, lockwords and collocations as a two-step procedure. ICAME 35, Nottingham
Hartley J (2016) Is time up for the Flesch measure of reading ease? Scientometrics 107(3):1523–1526
HEFCE (2011) Assessment framework and guidance on submissions. Ref. 02.2011
Hill S (2016) Assessing (for) impact: future assessment of the societal impact of research. Palgrave Commun 2:16073
Hyland K (2002) Directives: argument and engagement in academic writing. Appl Linguist 23(2):215–238
Hyland K (2008) As can be seen: lexical bundles and disciplinary variation. Engl Specif Purp 27(1):4–21

Lemos MC (2015) Usable climate knowledge for adaptive and co-managed water governance. Curr Opin Environ Sustain 12:48–52
MacDonald R (2017) "Impact", research and slaying Zombies: the pressures and possibilities of the REF. Int J Sociol Soc Policy 37(11–12):696–710
Mårtensson P et al. (2016) Evaluating research: a multidisciplinary approach to assessing research practice and quality. Res Policy 45(3):593–603
McEnery T, Xiao R, Tono Y (2006) Corpus-based language studies: an advanced resource book. Routledge, Abingdon
McNamara DS et al. (2014) Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press, New York, NY
National Science Foundation (2014) Perspectives on broader impacts
Nesta (2018) Seven principles for public engagement in research and innovation policymaking. https://www.nesta.org.uk/documents/955/Seven_principles_HlLwdow.pdf. Last Accessed 12 Dec 2019
Oancea A (2010) The BERA/UCET review of the impacts of RAE 2008 on education research in UK higher education institutions. ERA/UCET, Macclesfield
Oancea (2014) Research assessment as governance technology in the United Kingdom: findings from a survey of RAE 2008 impacts. Z Erziehungswis 17 (S6):83–110
Olssen M, Peters MA (2005) Neoliberalism, higher education and the knowledge economy: from the free market to knowledge capitalism. J Educ Policy 20 (3):313–345
Pidd M, Broadbent J (2015) Business and management studies in the 2014 Research Excellence Framework. Br J Manag 26:569–581
Pollitt A et al. (2016) Understanding the relative valuation of research impact: a best–worst scaling experiment of the general public and biomedical and health researchers. BMJ Open 6(8):e010916
Rau H, Goggins G, Fahy F (2018) From invisibility to impact: recognising the scientific and societal relevance of interdisciplinary sustainability research. Res Policy 47(1):266–276
Ravenscroft J et al. (2017) Measuring scientific impact beyond academia: an assessment of existing impact metrics and proposed improvements. PLoS ONE 12(3):e0173152
Rayson P, Garside R (2000) Comparing corpora using frequency profiling, Workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000), Hong Kong, pp. 1–6
Rayson P, Berridge D, Francis B (2004) Extending the Cochran rule for the comparison of word frequencies between corpora. In: Purnelle G, Fairon C, Dister A (eds.), Le poids des mots: Proceedings of the 7th international conference on statistical analysis of textual data (JADT 2004) (II). Presses universitaires de Louvain, Louvain-la-Neuve, Belgium, pp. 926–936
Reed MS (2018) The research impact handbook, 2nd edn. Fast Track Impact, Huntly, Aberdeenshire
Reed MS (2019) Book review: new book calls for civil disobedience to fight "dehumanising" impact agenda. Fast Track Impact
Reed MS et al. (under review) Evaluating research impact: a methodological framework. Res Policy
Rhoads R, Torres CA (2005) The University, State, and Market: The Political Economy of Globalization in the Americas. Stanford University Press, Stanford
Saldana J (2009) The Coding Manual for Qualitative Researchers. Sage, Thousand Oaks
Scott M (1997) PC analysis of key words—and key key words. System 25(2):233–245
Sivertsen G (2017) Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. Palgrave Commun 3:17078
Smith S, Ward V, House A (2011) 'Impact' in the proposals for the UK's Research Excellence Framework: shifting the boundaries of academic autonomy. Res Policy 40(10):1369–1379
Stern LN (2016) Building on success and learning from experience: an independent review of the Research Excellence Framework
Tsey K et al. (2016) Evaluating research impact: the development of a research for impact tool. Front Public Health 4:160
Vold ET (2006) Epistemic modality markers in research articles: a cross-linguistic and cross-disciplinary study. Int J Appl Linguist 16(1):61–87
Warry P (2006) Increasing the economic impact of the Research Councils (the Warry report). Research Council UK, Swindon
Watermeyer R (2019) Competitive accountability in academic life: the struggle for social impact and public legitimacy. Edward Elgar, Cheltenham
Watermeyer R, Hedgecoe A (2016) 'Selling 'impact': peer reviewer projections of what is needed and what counts in REF impact case studies. A retrospective analysis. J Educ Policy 31:651–665
Watermeyer R, Chubb J (2018) Evaluating 'impact' in the UK's Research Excellence Framework (REF): liminality, looseness and new modalities of scholarly distinction. Stud Higher Educ 44(9):1–13
Weinstein N et al. (2019) The real-time REF review: a pilot study to examine the feasibility of a longitudinal evaluation of perceptions and attitudes towards REF 2021
Wilsdon J et al. (2015) Metric tide: report of the independent review of the role of metrics in research assessment and management
Yang A, Zheng S, Ge G (2015) Epistemic modality in English-medium medical research articles: a systemic functional perspective. Engl Specif Purp 38:1–10

Acknowledgements

Thanks to Dr. Adam Mearns, School of English Literature, Language & Linguistics at Newcastle University for help with statistics and wider input to research design as a co-supervisor on the Ph.D. research upon which this article is based.

Competing interests

MR is CEO of Fast Track Impact Ltd, providing impact training to researchers internationally. JC worked with Research England as part of the Real-Time REF Review in parallel with the writing of this article. BR offers consultancy services reviewing REF impact case studies.

Additional information

Correspondence and requests for materials should be addressed to M.S.R.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.